HAL
open science

Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs

Violaine Rebouillat

To cite this version:

Violaine Rebouillat. Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs. Sciences de l'information et de la communication. Conservatoire national des arts et metiers - CNAM, 2019. Français. NNT : 2019CNAM1254 . tel-02447653

HAL Id: tel-02447653

https://theses.hal.science/tel-02447653

Submitted on 21 Jan 2020

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

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

le cnam

École doctorale Abbé Grégoire

Dispositifs d'Information et de Communication

à l'Ère Numérique – Paris Île de France (DICEN-IdF)

THÈSE

Présentée par : **Violaine REBOUILLAT**

Soutenue le : **3 décembre 2019**

Pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline : Sciences de l'information et de la communication

Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs

Thèse dirigée par :

Mme CHARTRON Ghislaine	Professeur, Cnam

Et co-encadrée par :

M. SCHÖPFEL Joachim	Maître de conférences, Université de Lille

RAPPORTEURS :

M. IHADJADENE Madjid	Professeur, Université Paris 8
M. LIQUÈTE Vincent	Professeur, Université de Bordeaux

JURY :

Mme PAGANELLI Céline	Professeur, Université Paul Valéry (présidente du jury)
Mme CHARTRON Ghislaine	Professeur, Cnam (directrice)
M. SCHÖPFEL Joachim	Maître de conférences, Université de Lille (co-encadrant)
M. IHADJADENE Madjid	Professeur, Université Paris 8 (rapporteur)
M. LIQUÈTE Vincent	Professeur, Université de Bordeaux (rapporteur)
M. DACOS Marin	Conseiller pour la science ouverte du Directeur général de la recherche et de l'innovation, Ministère de l'enseignement supérieur, de la recherche et de l'innovation (membre invité)

Remerciements

Mes premiers remerciements vont à Ghislaine Chartron et Joachim Schöpfel, mes directeurs de thèse. Je les remercie pour leur disponibilité, malgré la difficulté que nous avions d'habiter chacun dans une ville différente. Je les remercie également pour leurs conseils et leurs encouragements, qui ont été déterminants tout au long de ce travail de recherche.

Mes remerciements vont ensuite à Madjid Ihadjadene et Vincent Liquète qui ont accepté d'être rapporteurs de ce travail. Je remercie également Céline Paganelli et Marin Dacos pour avoir accepté de prendre part au jury de thèse. Merci à eux quatre pour le temps octroyé à l'évaluation de ce travail.

Je remercie Bernard Jacquemin et Joumana Boustany, qui ont accepté de faire partie du comité de suivi de thèse. Leur évaluation m'a permis d'ajuster mes recherches.

Merci à Francis André et Paul-Antoine Hervieux pour m'avoir permis d'utiliser dans la thèse les résultats de la cartographie. Merci à Ourida Aberkane, Anne Ciolek-Figiel et Marie-Christine Jacquemot-Perbal de l'Inist-CNRS, qui ont conçu Cat OPIDoR et qui contribuent encore aujourd'hui à son amélioration. Merci à mes responsables, Adeline Rege et Anne Pelletier, qui ont su tenir compte de mon statut de doctorante et m'ont toujours laissé une grande souplesse dans l'organisation de mon emploi du temps. Une pensée particulière pour Adeline, qui était là au tout début, lorsque la thèse n'était encore qu'un projet.

Je pense à tous les doctorants, chercheurs et enseignants-chercheurs que j'ai eu l'occasion de rencontrer lors de colloques, séminaires ou cafés doctorants. Je garde un très bon souvenir de nos échanges, au cours desquels j'ai pu partager mes hésitations et dont j'ai beaucoup appris.

Je tiens à exprimer ma gratitude envers les chercheurs et responsables de service, qui ont accepté de participer à mes enquêtes. Je les remercie pour leur disponibilité et leur accueil.

Je suis reconnaissante envers tous ceux qui m'ont entourée au cours de ces quatre années. Je remercie mes parents pour leur soutien. Ils ont toujours fait en sorte que je puisse concrétiser mes projets d'études. Je voudrais également remercier chaleureusement Josette Le Fur pour son écoute attentive et ses conseils. Un grand merci à Valériane qui a relu et commenté la thèse en un temps record. Enfin, je remercie mon compagnon, Jori, qui a su me redonner de l'énergie, quand j'en manquais. La mise en page de cette thèse lui doit beaucoup.

Résumé

Cette thèse s'intéresse aux données de la recherche, dans un contexte d'incitation croissante à leur ouverture. Les données de la recherche sont des informations collectées par les scientifiques dans la perspective d'être utilisées comme preuves d'une théorie scientifique. Il s'agit d'une notion complexe à définir, car contextuelle. Depuis les années 2000, le libre accès aux données occupe une place de plus en plus stratégique dans les politiques de recherche. Ces enjeux ont été relayés par des professions intermédiaires, qui ont développé des services dédiés, destinés à accompagner les chercheurs dans l'application des recommandations de gestion et d'ouverture. La thèse interroge le lien entre philosophie de l'ouverture et pratiques de recherche. Quelles formes de gestion et de partage des données existent dans les communautés de recherche et par quoi sont-elles motivées ? Quelle place les chercheurs accordent-ils à l'offre de services issue des politiques de gestion et d'ouverture des données ? Pour tenter d'y répondre, 57 entretiens ont été réalisés avec des chercheurs de l'Université de Strasbourg dans différentes disciplines. L'enquête révèle une très grande variété de pratiques de gestion et de partage de données. Un des points mis en évidence est que, dans la logique scientifique, le partage des données répond un besoin. Il fait partie intégrante de la stratégie du chercheur, dont l'objectif est avant tout de préserver ses intérêts professionnels. Les données s'inscrivent donc dans un cycle de crédibilité, qui leur confère à la fois une valeur d'usage (pour la production de nouvelles publications) et une valeur d'échange (en tant que monnaie d'échange dans le cadre de collaborations avec des partenaires). L'enquête montre également que les services développés dans un contexte d'ouverture des données correspondent pour une faible partie à ceux qu'utilisent les chercheurs. L'une des hypothèses émises est que l'offre de services arrive trop tôt pour rencontrer les besoins des chercheurs. L'évaluation et la reconnaissance des activités scientifiques étant principalement fondées sur la publication d'articles et d'ouvrages, la gestion et l'ouverture des données ne sont pas considérées comme prioritaires par les chercheurs. La seconde hypothèse avancée est que les services d'ouverture des données sont proposés par des acteurs relativement éloignés des communautés de recherche. Les chercheurs sont davantage influencés par des réseaux spécifiques à leurs champs de recherche (revues, infrastructures…). Ces résultats invitent donc à reconsidérer la question de la médiation dans l'ouverture des données scientifiques.

Abstract

The thesis investigates research data, as there is a growing demand for opening them. Research data are information that is collected by scientists in order to be used as evidence for theories. It is a complex, contextual notion. Since the 2000s, open access to scientific data has become a strategic axis of research policies. These policies has been relayed by third actors, who developed services dedicated to support researchers with data management and sharing. The thesis questions the relationship between the ideology of openness and the research practices. Which kinds of data management and sharing practices already exist in research communities? What drives them? Do scientists rely on research data services? Fifty-seven interviews were conducted with researchers from the University of Strasbourg in many disciplines. The survey identifies a myriad of different data management and sharing practices. It appears that data sharing is embedded in the researcher's strategy: his main goal is to protect his professional interests. Thus, research data are part of a credibility cycle, in which they get both use value (for new publications) and exchange value (as they are traded for other valuable resources). The survey also shows that researchers rarely use the services developed in a context of openness. Two explanations can be put forward. (1) The service offer comes too early to reach researchers' needs. Currently, data management and sharing are not within researchers' priorities. The priority is publishing, which is defined as source of reward and recognition of the scientific activities. (2) Data management services are offered by actors outside the research communities. But scientists seem to be more influenced by internal networks, close to their research topics (like journals, infrastructures…). These results prompt us to reconsider the mediation between scientific communities and open research data policies.

Sommaire

Remerciements.....3

Résumé.....4

Abstract.....5

Sommaire.....6

Liste des tableaux.....8

Liste des figures.....9

Introduction.....11

Première partie - Qu'est-ce qu'une donnée de la recherche ?.....23

1. Des tentatives de définition.....25
2. Vers une non définition.....34
3. Conclusion.....45

Deuxième partie - Les politiques publiques de gestion et d'ouverture des données de la recherche.....47

1. Mouvements à l'origine des politiques de gestion et d'ouverture des données de la recherche.....50
2. Influence de l'Open Data.....56
3. Politiques et initiatives de l'Union européenne.....65
4. Politiques et initiatives de l'État français.....88
5. Conclusion.....96

Troisième partie - Les services d'appui à la gestion et au partage des données de recherche.....101

1. Terrain et méthodologie.....105

2. Le paysage national des services de gestion et d'ouverture des données : constats 114

3. Conclusion.....151

Quatrième partie - Les données dans les pratiques de recherche.....153

1. Terrain et méthodologie.....156

2. Résultats et discussion.....162

3. Conclusion.....210

Cinquième partie - Adéquation entre les services de données et les pratiques des chercheurs.....213

1. Enquête sur l'utilisation de services de données par les chercheurs.....216

2. Utilisation par les chercheurs des services nés sous l'influence du mouvement d'ouverture des données.....224

3. Utilisation des services disciplinaires par les chercheurs.....228

Conclusion.....249

Bibliographie.....263

Sources.....281

Sitographie & Acronymes.....293

Annexes.....307

Sommaire détaillé.....481

Liste des tableaux

Tableau 1 : Degrés de traitement des données produites par le système d'observation de la Terre de la NASA.....32

Tableau 2 : Volumétrie des entrepôts de données en 2016.....128

Tableau 3 : Volumétrie des annuaires de données en 2016.....129

Tableau 4 : Profil des professionnels gérant les services de données.....134

Tableau 5 : Services gérés par des professionnels de l'IST.....135

Tableau 6 : Nature de l'offre de service des 10 structures IST répertoriées.....143

Tableau 7 : Taux d'évolution des annuaires de données entre 2016 et 2019.....144

Tableau 8 : Taux d'évolution des entrepôts de données entre 2016 et 2019.....145

Tableau 9 : Composition de l'échantillon initial.....159

Tableau 10 : Composition du second échantillon.....160

Tableau 11 : Liste des chercheurs interrogés.....163

Tableau 12 : Définitions de « donnée de recherche » proposées par les chercheurs interrogés 169

Tableau 13 : Liste des chercheurs interrogés sur l'utilisation de services de données.....217

Tableau 14 : Proportion de chercheurs par laboratoire à avoir déjà partagé des données sous-jacentes à une publication.....226

Liste des figures

Figure 1 : Réponses des chercheurs interrogés par F. Cabrera à la question « Dans le cadre de votre pratique, qu'est-ce qu'une donnée de la recherche ? »....29

Figure 2 : Variables influençant le type de données collectées....40

Figure 3 : Définition quadridimensionnelle des données de recherche....44

Figure 4 : Chronologie des politiques publiques d'ouverture des données....97

Figure 5 : Cycle de vie des données de la recherche....107

Figure 6 : Nombre de services analysés par type....117

Figure 7 : Répartition des services par domaine scientifique....121

Figure 8 : Types de structures à l'origine d'un service....131

Figure 9 : Evolution du nombre d'annuaires répertoriés dans Cat OPIDoR par grand domaine scientifique....147

Figure 10 : Evolution du nombre d'entrepôts répertoriés dans Cat OPIDoR par grand domaine scientifique....147

Figure 11 : Cycle de crédibilité selon B. Latour....171

Figure 12 : Cycles de crédibilité dans la recherche appliquée....199

Introduction

Le 4 juillet 2018, la Ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation, Frédérique Vidal, dévoilait un plan national pour la science ouverte[1]. Doté d'un budget de 5,4 millions d'euros pour l'année 2018-2019, ce plan entend généraliser le libre accès aux publications et aux données de la recherche financée sur fonds publics. Le Ministère instaure donc une politique en matière de données de recherche, avec pour ambition de systématiser leur structuration, leur préservation et leur ouverture.

Dans son discours, Frédérique Vidal présente l'ouverture des données comme un moyen de renouveler le lien entre science et société.

> *« La science est un bien commun, que nous devons partager le plus largement possible. Le rôle des pouvoirs publics est de rétablir la fonction initiale de la science, comme facteur d'enrichissement collectif. Car la diffusion des connaissances scientifiques a un impact direct en termes de développement économique, sanitaire, social. Je m'inspire de la stratégie qui a été mise en place dans le cadre du Human Genome Project, qui a coûté 3,8 milliards de dollars mais dont il a été décidé que les résultats seraient publics et considérés comme patrimoine de l'humanité.*
>
> *Cette décision a permis l'exploitation scientifique et médicale formidable que nous connaissons tous. D'ailleurs, on estime qu'en 2012 ce projet avait eu un impact économique de 796 milliards. Les données astronomiques sont traditionnellement diffusées après une année d'embargo. Deux tiers de la littérature scientifique en astronomie s'appuient sur des données ouvertes ! Ce potentiel énorme, exploité par les astronomes, ne l'est pas encore dans toutes les disciplines. »*[2]

La Ministre justifie le plan national pour la science ouverte en s'appuyant sur l'exemple de deux communautés de recherche : la génomique et l'astronomie. Dans ces disciplines, l'ouverture des données est le fruit d'une initiative des communautés. En assignant à l'État le

1 MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION (2018). *Plan national pour la science ouverte*. http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf (consulté le 18 septembre 2019)

2 VIDAL, F. (2018). 'Plan national pour la science ouverte'. Discours prononcé lors du *Congrès LIBER*, 4 juillet 2018, Villeneuve d'Ascq. http://www.enseignementsup-recherche.gouv.fr/cid132531/plan-national-pour-la-science-ouverte-discours-de-frederique-vidal.html (consulté le 18 septembre 2019)

rôle de généraliser l'ouverture des données à toutes les disciplines, Frédérique Vidal déplace cette prérogative des communautés de recherche à un niveau politique.

Le plan national pour la science ouverte n'est qu'un exemple parmi d'autres d'intervention du politique dans l'auto-détermination des communautés à diffuser leurs résultats de recherche. Depuis une quinzaine d'années, les incitations politiques en faveur de l'ouverture des données semblent s'être multipliées. Aux États-Unis, les agences de financement de la recherche ont très tôt instauré des politiques de données : les *National Institutes of Health* (NIH) en 2003 ; la *National Science Foundation* (NSF) en 2010 (Gaillard 2014, p.29-30). Sur le même modèle, la Commission européenne imposera à partir de 2021 l'ouverture des données issues de l'ensemble des projets de recherche qu'elle finance[3]. Ce mandat d'ouverture se veut compatible avec d'éventuelles perspectives de valorisation des résultats de recherche. Aussi le principe retenu est-il celui d'un accès *« aussi ouvert que possible, aussi fermé que nécessaire »*[4]. Ces incitations ont été relayées par des corps intermédiaires, comme les professionnels de la documentation (bibliothécaires, archivistes…), qui ont développé des services dédiés, destinés à accompagner les chercheurs dans l'application des recommandations de gestion et d'ouverture.

Ces différentes initiatives traduisent une logique descendante dans la volonté d'orienter les pratiques de gestion et de diffusion des données – logique dont on est en droit de questionner la pertinence. Comme le souligne Ghislaine Chartron (2018), *« des corps intermédiaires prônent les orientations et prennent les décisions, à la place même des principaux acteurs concernés. Ainsi, dans l'évolution du mouvement Open Access/Open Science, une grande différence est évidente entre la liberté d'ouvrir ses travaux des années 1990 et l'injonction politique actuelle à tout ouvrir, heurtant profondément l'autonomie de décision du chercheur ».*

3 COMMISSION EUROPÉENNE (2018a). *Budget de l'Union : La Commission propose le programme de recherche et d'innovation le plus ambitieux à ce jour*. Communiqué de presse, 7 juin 2018, Bruxelles. https://europa.eu/rapid/press-release_IP-18-4041_fr.htm (consulté le 18 septembre 2019)

4 COMMISSION EUROPÉENNE (2018c). *Recommandation de la Commission du 25.4.2018 relative à l'accès aux informations scientifiques et à leur conservation*. http://data.europa.eu/eli/reco/2018/790/oj (consulté le 8 octobre 2019).

Pour ma part, j'ai commencé à m'intéresser à la question des données de recherche dans le cadre de mon mémoire de master. À cette occasion, j'ai rencontré plusieurs chercheurs de l'Université de Strasbourg, pour les interroger sur leurs besoins en termes de valorisation de données. Constatant un décalage entre leurs discours et celui des acteurs politiques prônant l'ouverture, je me suis interrogée sur la pertinence de ce mouvement pour la recherche.

À cette première expérience s'est ajoutée en 2015 une mission de chargée d'étude pour la Bibliothèque Scientifique Numérique. L'étude a consisté à recenser les différents services existant en France autour de la gestion et de l'ouverture des données de la recherche. Elle a renforcé mes interrogations sur le lien entre philosophie de l'ouverture et pratiques de recherche et m'a conduite à étudier de plus près l'utilisation des services de données par les chercheurs.

De ces deux expériences est venu le questionnement de la thèse. L'instauration de politiques d'ouverture et le développement de services dédiés suffisent-ils à modifier les pratiques ? Quelle est la perspective du chercheur ? L'ouverture des données contribue-t-elle à faire avancer sa recherche ou bien, au contraire, le ralentit-elle ?

A l'aune d'injonctions politiques nouvelles de gestion et d'ouverture, l'objectif de ce travail de thèse a donc été d'identifier la place occupée par les données dans les pratiques de recherche.

Questions de recherche

La question de la gestion des données n'est pas nouvelle au sein de la communauté scientifique. Ce sont des chercheurs en astronomie (Borgman et al. 2016), en cristallographie (Bruno et al. 2017) et en génomique (Lander et al. 2001) qui, les premiers, ont instauré des moyens pour le partage de leurs données[5]. C'est pourquoi j'ai choisi d'explorer le point de vue des communautés scientifiques. A l'heure où les politiques publiques tendent à imposer une structuration et une ouverture généralisées des données scientifiques, il semble important

5 D'autres communautés ont elles aussi entrepris de développer des moyens pour gérer et réutiliser les données de façon partagée. On peut citer le domaine de l'écologie et celui, plus transversal, des données géospatiales. La définition de standards d'échange s'opère au sein de consortiums comme la Global Spatial Data Infrastructure Association (GDSI), aujourd'hui dissoute, ou au sein d'organisations comme la Research Data Alliance (RDA).

de cerner le contexte dans lequel elles s'insèreront, d'en identifier les possibles pierres d'achoppement et de mettre en lumière les points d'accroche qui rendront leur adoption plus aisée par les équipes de recherche. Il paraît également important d'anticiper la façon dont les services de données peuvent être pensés pour s'adapter le mieux possible aux pratiques de recherche. Pour ce faire, je me suis posé les deux questions suivantes :

(1) Comment, dans un projet de recherche, sont déterminés les modes de gestion et de partage des données ?

(2) Quelle place les chercheurs accordent-ils à l'offre de services issue des politiques de gestion et d'ouverture des données ?

Pour tenter de répondre à la première question de recherche, trois hypothèses ont été émises.

Première hypothèse : la gestion et le partage des données sont déterminés par le cadre épistémique dans lequel s'ancre le projet de recherche.

La notion de « cadre épistémique » est à entendre dans le sens que lui donne Karin Knorr-Cetina (1981), lorsqu'elle développe le concept de « cultures épistémiques ». La sociologue désigne à travers ce concept la diversité des cultures scientifiques, la façon de faire la science variant en fonction des outils, des méthodes et des raisonnements employés dans une discipline. La première hypothèse stipule donc que les spécificités d'une thématique scientifique ont une influence sur les pratiques de gestion et de partage des données. Dans un domaine comme l'astronomie, la collecte des données passe par de grands instruments (les télescopes). Étant coûteux et générateurs d'importantes quantités de données, ces instruments sont aujourd'hui mutualisés à l'échelle internationale. On constate que la communauté des astronomes et astrophysiciens se structure autour de ces grands équipements : les chercheurs collaborent entre eux pour acquérir et analyser les données. En termes de gestion et de partage, des processus standardisés ont été mis en place. Les données sont structurées et archivées au niveau des télescopes. Elles sont rendues librement accessibles à la communauté scientifique, afin que celle-ci puisse les analyser et les interpréter (Borgman et al. 2016).

La gestion standardisée et le partage systématique des données, comme en astronomie, sont-ils caractéristiques de communautés collectant les données à partir de grands instruments ? Ou bien sont-ils un mode d'organisation spécifique à l'astronomie ? Dans ce cas, l'utilisation de grands instruments générerait des modes d'organisation différents selon les communautés, sans être systématiquement fondée sur le partage des données.

Deuxième hypothèse : la gestion et le partage des données sont influencés par le cadre institutionnel qui environne les chercheurs impliqués dans un projet.

La notion de cadre institutionnel est entendue ici au sens large : elle englobe à la fois le laboratoire dans lequel le chercheur travaille, l'établissement de recherche auquel il est affilié, les éditeurs qui vont publier ses articles et ouvrages, les agences de financement auprès desquelles il dépose des demandes de subventions, ainsi que les autres partenaires, publics ou privés, avec lesquels il établit des contrats de recherche. Cette hypothèse propose de considérer la dimension normative du cadre institutionnel. Celui-ci produit toutes sortes de recommandations, de règles, de protocoles, de contrats ou de chartes, plus ou moins prescriptifs, qui environnent l'activité scientifique. Dans un article de la revue *COSSI*, Joachim Schöpfel (2018a) développe le concept de « norme », reprenant la définition donnée par Cacaly (1997) : une norme est un *« document établi par consensus et approuvé par un organisme reconnu, qui fournit, pour des usages communs et répétés, des règles, des lignes directrices ou des caractéristiques, pour des activités ou leurs résultats, garantissant un niveau d'ordre optimal dans un contexte donné ».* Selon J. Schöpfel, les normes qui s'appliquent, directement ou indirectement, aux données de recherche sont multiples. Elles peuvent avoir un caractère légal (la loi « Informatique et Libertés » pour les données personnelles, par exemple), un caractère éthique (l'avis des Comités de Protection des Personnes, chargés de valider les protocoles de recherche impliquant la personne humaine) ou un caractère industriel (la norme ISO 9001 pour le management de la qualité, fondée sur un principe de réplicabilité de la production). L'objectif est d'étudier l'incidence de ces normes sur la gestion et le partage des données de recherche. Dans quelle mesure les chercheurs observent-ils ces normes ? En quoi cela influence-t-il la gestion et le partage des données ?

Troisième hypothèse : la gestion et le partage des données dépendent du cadre social dans lequel s'insère un projet de recherche.

Cette hypothèse envisage le milieu scientifique dans sa dimension sociale. La culture scientifique semble fondée sur un principe d'échanges privés et réciproques, allant à contre-courant de la notion d'ouverture prônée par l'Open Science. Wallis et al. (2013) évoquent une « culture du don » (*gift culture*), en référence au modèle de don/contre-don mis en évidence par l'anthropologue Marcel Mauss dans les sociétés archaïques (Mauss 2007). Dans la sphère scientifique, ce principe consisterait pour un chercheur à ne partager ses données qu'avec des pairs de sa connaissance et à la condition implicite d'obtenir une rétribution symbolique en échange (être co-auteur d'une publication par exemple). De cette manière, le chercheur parviendrait à préserver ses intérêts professionnels. Ce mécanisme de don/contre-don avait déjà été repris par Warren O. Hagstrom (1965), qui l'appliquait au principe de publication scientifique.

Quant à la seconde question de recherche, l'hypothèse suivante a été émise.

Quatrième hypothèse : Les services de gestion et d'ouverture des données sont peu connus des chercheurs.

Depuis la Déclaration de Berlin en 2003[6] et dans un contexte d'engouement croissant pour les « *data* » (Cukier 2010) et l'ouverture de la science (Leonelli 2013b), de nombreux services ont vu le jour, dans la perspective de proposer aux chercheurs un appui pour gérer et partager les données scientifiques. « Service » est entendu ici comme la fourniture de ressources humaines et/ou techniques pour gérer les données à une ou plusieurs étapes d'un projet de recherche. Je me suis donc également interrogée sur la réception des services de données par les chercheurs. Plusieurs enquêtes (Fecher et al. 2015 ; Tenopir et al. 2011) ont montré que l'ouverture des données était un sujet connu des chercheurs, mais qu'elle était devancée dans leur quotidien par des préoccupations de publication. La mise à disposition peu fréquente des données aurait donc des répercussions sur l'utilisation des services de gestion et d'ouverture, les chercheurs ne trouvant pas forcément l'occasion d'y avoir recours.

6 MAX PLANCK GESELLSCHAFT (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. http://openaccess.mpg.de/Berlin-Declaration (consulté le 18 septembre 2019).
La Déclaration de Berlin sera présentée dans la partie 2 de la thèse (voir 1.1, p.50).

Méthodologie et cadre théorique

La thèse s'inscrit dans le champ des sciences de l'information et de la communication. Elle s'appuie, par ailleurs, sur certains aspects conceptuels et méthodologiques de la sociologie.

Pour répondre aux questions de recherche énoncées ci-dessus, j'ai réalisé une enquête sous forme d'entretiens semi-directifs auprès de 57 chercheurs de l'Université de Strasbourg. L'approche choisie est une approche qualitative. L'objectif n'était pas de quantifier les pratiques de recherche, mais de les rendre intelligibles dans leur variété et leur complexité (Kaufmann 1996; Lejeune 2014).

Sur le plan théorique, les sciences de l'information et de la communication (SIC) m'ont fourni une première approche critique du domaine de la communication scientifique. Le mouvement d'ouverture de la science y est étudié dans sa dynamique d'acteurs. Les travaux de Ghislaine Chartron (Chartron 2010 ; Chartron 2016 ; Chartron et Schöpfel 2017) montrent les tensions politiques, économiques et sociales qui le traversent et le font évoluer. La question des données de la recherche a également été abordée par les sciences de l'information et de la communication. Des enquêtes comme celles réalisées par Tenopir et al. (2015a), Chowdhury et al. (2017), Serres et al. (2017) et Schöpfel (2018b), s'attachent à comprendre comment et avec qui les chercheurs gèrent et partagent leurs données. Elles ont contribué à mieux connaître les motivations et réticences des chercheurs.

Pour comprendre de l'intérieur quels enjeux recouvre l'ouverture des données, je me suis tournée vers la sociologie des sciences. Je me suis notamment appuyée sur les premières ethnographies de laboratoires (Knorr-Cetina 1981 ; Latour et Woolgar 1979). Ces travaux s'efforcent de restituer la façon dont se fait la recherche, en décrivant les mécanismes par lesquels des résultats deviennent connaissances.

La sociologie des sciences donne à voir l'univers de la recherche comme un système social à part entière, avec ses modes d'échanges propres. Au travers du concept de « crédit scientifique », Pierre Bourdieu (1975) et Bruno Latour (2001) montrent que les connaissances scientifiques sont des biens symboliques que le chercheur troque en échange de crédit. Ce crédit sera lui-même réinvesti dans de nouvelles recherches, permettant de produire de nouvelles connaissances et ainsi d'acquérir encore davantage de crédit. Ce système repose

donc sur la plus-value informationnelle, dont font notamment partie les données de recherche. Choisir d'ouvrir les données a, par conséquent, un impact inévitable sur cet équilibre entre connaissances et crédit.

Une **première partie** sera consacrée à éclairer le terme de « donnée de recherche » (tout au long de la thèse, seront utilisés indifféremment les termes de « données de la recherche », « données de recherche » et de « données scientifiques »). Nous verrons que les premières définitions du terme sont nées pour accompagner le mouvement d'ouverture des résultats scientifiques, mais qu'il n'existe à ce jour aucune définition faisant consensus. Il s'agira alors d'expliquer pourquoi il est si difficile de définir ce qu'est une donnée de recherche. Le concept de *« Information as thing »* (Buckland 1991) sera convoqué pour montrer le caractère relatif de la donnée. On retrouvera cette dimension dans des travaux récents sur les données de la recherche, en philosophie (S. Leonelli) et en sciences de l'information et de la communication (C. L. Borgman ; J. Schöpfel). Les données n'y sont pas définies selon leurs propriétés intrinsèques mais selon leur fonction au sein de processus de recherche particuliers. Cela revient à dire qu'on ne peut répondre à la question « qu'est-ce qu'une donnée ? » qu'en faisant référence à des situations de recherche concrètes.

Une **deuxième partie** révèlera la dimension politique que recouvre l'ouverture des données de recherche. Elle présentera d'abord les courants de pensée qui sont à l'origine du principe d'ouverture (le libre accès aux résultats scientifiques et la transparence des données du secteur public, notamment). Elle s'attachera ensuite à décrire les initiatives politiques de l'Union européenne et de l'État français en matière de gestion et d'ouverture des données scientifiques, depuis la création d'un Espace Européen de la Recherche (2000) jusqu'au déploiement d'un plan national pour la science ouverte (2018). Cette partie vise à identifier les enjeux et manifestations politiques du mouvement d'ouverture. L'enjeu est d'évaluer l'adéquation de ces derniers avec les dynamiques de la recherche, qui seront décrites dans la quatrième partie de la thèse. Nous verrons que ces initiatives politiques sous-tendent une vision libérale de la science. Dans les discours de la Commission européenne notamment, le libre accès aux données de la recherche est toujours associé au champ de l'innovation et de la croissance économique. Il est utilisé comme vecteur pour rapprocher science et économie.

Nous montrerons que l'adhésion au principe d'ouverture va toutefois au-delà de cette vision socio-économique. L'ouverture des données est une expression qui a le vent en poupe. Son sens est suffisamment large et mélioratif pour parvenir à fédérer des acteurs aux perspectives diverses. Ce phénomène permet d'expliquer le ralliement de différentes organisations d'appui à la recherche autour des politiques publiques d'ouverture et l'engagement de ces dernières dans le développement de services pour la gestion et l'ouverture des données de la recherche.

La **troisième partie** se concentrera sur les services de gestion et de partage des données. L'objectif est d'analyser l'offre de services permettant de répondre à la demande d'ouverture des données formulée par les instances politiques et d'en identifier les acteurs. Cette partie constitue un préambule à la dernière partie de la thèse, dans la mesure où elle introduit une réflexion sur l'adéquation des services de données avec les pratiques des chercheurs (seconde question de recherche). Elle s'appuiera sur le travail de cartographie réalisé entre 2015 et 2017 pour la Bibliothèque Scientifique Numérique. L'analyse sera donc circonscrite au paysage des services présents en France à cette période (soit 44 services). Seront d'abord présentés les aspects méthodologiques du recensement et l'aboutissement de ce dernier, avec la création du catalogue en ligne Cat OPIDoR[7]. Les services seront ensuite étudiés dans leur diversité, à travers leurs fonctions, leurs fournisseurs et leurs modes de financement. Nous aboutirons à la conclusion d'un paysage hétérogène, dépourvu de coordination nationale à la période étudiée et composé de services en partie conçus par des intermédiaires.

Une **quatrième partie** portera sur les pratiques de recherche. Elle s'appuiera sur les résultats d'une enquête qualitative menée auprès de 57 chercheurs de l'Université de Strasbourg. Ces résultats permettront d'étudier la place occupée par les données dans l'activité scientifique des chercheurs. Nous verrons que les données s'inscrivent dans un cycle de crédibilité (Latour 2001), qui leur confère à la fois une valeur d'usage (pour la production de nouvelles publications) et une valeur d'échange (en tant que monnaie d'échange dans le cadre de collaborations avec des partenaires). L'enquête révèle également une diversité des modes de gestion des données, dont on étudiera deux paramètres d'influence : les ressources humaines et matérielles à disposition du chercheur ; les normes s'appliquant aux données personnelles

7 Catalogue pour une Optimisation du Partage et de l'Interopérabilité des Données de la Recherche (https://cat.opidor.fr)

et médicales, ainsi qu'aux données de la recherche appliquée. Dans cette partie seront vérifiées les première, deuxième et troisième hypothèses de recherche.

Enfin, la **cinquième partie** s'intéressera de plus près à l'utilisation des services de gestion et d'ouverture des données. Une partie des chercheurs participant à l'enquête ont été interrogés sur les services de données qu'ils utilisaient, avec pour point de départ de la discussion la présentation du catalogue Cat OPIDoR. L'enquête montre que les services répertoriés dans le catalogue ne sont pas ou peu connus des chercheurs, confirmant la quatrième hypothèse de recherche. Elle révèle l'existence d'une catégorie de services dédiés à l'acquisition et à la réutilisation de données, davantage utilisés par les chercheurs que les services nés dans un contexte d'ouverture de la science. Nous verrons qu'il s'agit de services très spécifiques, peu visibles en dehors de la communauté qui les utilise. L'intérêt sera d'en montrer les caractéristiques, afin que celles-ci puissent être prises en compte dans la réflexion sur l'offre de services de gestion et d'ouverture des données.

Première partie

-

Qu’est-ce qu’une donnée de la recherche ?

1. Des tentatives de définition

1.1. Quand a-t-on commencé à définir les données de recherche ?

Le terme de « donnée de la recherche » (*research data*) a été explicité pour la première fois en 1999 par le gouvernement fédéral américain. Dans une circulaire destinée à délimiter le cadre des projets financés par l'administration, le Bureau de la Gestion et du Budget définit la donnée de recherche comme *« l'enregistrement factuel couramment considéré dans la communauté scientifique comme nécessaire à la validation des résultats de la recherche »*[8].

Cette définition a été reprise en 2007 par l'Organisation de Coopération et de Développement Économiques (OCDE) :

> *« Dans le cadre de ces Principes et Lignes directrices [pour l'accès aux données de la recherche financée sur fonds publics], les « données de la recherche » sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche.*
>
> *Ce terme ne s'applique pas aux éléments suivants :* ***carnets de laboratoire****,* ***analyses préliminaires*** *et* ***projets de documents scientifiques****,* ***programmes de travaux futurs****,* ***examens par les pairs****,* ***communications personnelles avec des collègues*** *et* ***objets matériels*** *(par exemple, les* ***échantillons de laboratoire****, les* ***souches bactériennes*** *et les* ***animaux de laboratoire*** *tels que les souris). L'accès à tous ces produits ou résultats de la recherche est régi par d'autres considérations que celles abordées ici. »*[9]

8 « Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings », dans OFFICE OF MANAGEMENT AND BUDGET (1993). *Circulaire A-110*, amendée le 30 septembre 1999. https://www.govinfo.gov/app/details/CFR-2012-title2-vol1/CFR-2012-title2-vol1-part215 (consulté le 19 septembre 2019).
Traduction de Rémi Gaillard (2014, p.16)

C'est donc dans un objectif d'ouverture, et plus particulièrement à des fins économiques, qu'a été défini le terme de « donnée de la recherche ». A partir des années 2000, l'OCDE développe en effet une stratégie fondée sur l'innovation, pensée comme *« moteur du développement et de la croissance »*[10]. Il s'agit de développer un cadre favorable à la multiplication d'initiatives innovantes. Selon l'OCDE, l'innovation ne doit plus seulement être le fruit de pôles de recherche dédiés, mais doit essaimer à tous les niveaux de la société. Une perspective intersectorielle de l'innovation est prônée. Par conséquent, *« les politiques destinées à la soutenir doivent procurer à des acteurs très divers les moyens de participer à l'innovation et de bénéficier de ses résultats »*[11]. La science étant considérée comme source d'innovation, l'une des incitations de l'OCDE consiste à élargir l'accès aux produits de la recherche – notamment aux données scientifiques. L'objectif est de permettre à la société civile de réutiliser les données pour créer de la valeur, donc de l'innovation et de la croissance économique. Comme le formule Sabina Leonelli (2013b), *« les données tendent à être conceptualisées de plus en plus comme des produits de la recherche détenant une valeur en eux-mêmes, et de moins en moins comme des composants du processus de recherche n'ayant pas de valeur inhérente »*[12].

1.2. Définitions par l'énumération

Dans ce contexte de mise en valeur des données issues de la recherche ont émergé diverses tentatives pour définir leur périmètre.

La Commission européenne, qui a instauré une politique d'ouverture des données dans le cadre de son 8ème programme de financement à la recherche (« Horizon 2020 »), définit les données de la manière suivante :

---

9 ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2007). *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*. Paris: Éditions OCDE. http://www.oecd.org/fr/science/sci-tech/38500823.pdf (consulté le 19 septembre 2019). Page 18

10 ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2010). *La stratégie de l'OCDE pour l'innovation : pour prendre une longueur d'avance*. Paris : Éditions OCDE. https://dx.doi.org/10.1787/9789264084759-fr (consulté le 19 septembre 2019). Page 3

11 Ibid., p.3

12 Traduction de : « Data are increasingly conceptualized as inherently valuable products of scientific research, rather than as components of the research process that have no value in themselves » (Leonelli 2013b).

> *« Les données numériques de la recherche désignent les informations de forme numérique (en particulier les faits ou les chiffres), collectées pour être analysées et utilisées afin d'alimenter des raisonnements, des discussions ou des calculs. Il peut s'agir de* ***statistiques****, de* ***résultats d'expériences****, de* ***mesures****, d'****observations sur le terrain****, de* ***résultats d'enquêtes****, d'****enregistrements d'entretiens*** *ou d'****images****. »*[13]

A l'inverse de l'OCDE, qui livre une définition en creux, énumérant ce que ne recouvrent pas les données de la recherche (cahiers de laboratoire, correspondance entre chercheurs, échantillons biologiques…)[14], la Commission européenne propose une définition axée sur une liste d'exemples de ce que peuvent être des données de recherche. Ces deux propositions ont pour point commun d'être des définitions par l'exemple. Elles présentent l'inconvénient de ne pas établir de frontière claire entre ce qui relève de la notion de donnée de recherche et ce qui n'en relève pas.

Une autre manière de définir les données est proposée par Francisca Cabrera (2014). Sa démarche est intéressante, dans la mesure où elle est allée à la rencontre des chercheurs pour tenter de comprendre ce que recouvre le terme de « donnée de recherche ». En tant qu'instigateurs des processus de recherche, les chercheurs semblent en effet les mieux placés pour définir ce que sont leurs données. Dans son étude, Francisca Cabrera s'est essentiellement intéressée au domaine des sciences humaines et sociales (SHS). Elle a mené des entretiens semi-directifs dans 19 disciplines du domaine, auprès de 53 chercheurs de la sphère académique française. A la question « dans le cadre de votre pratique, qu'est-ce qu'une donnée de la recherche ? », les chercheurs interrogés ont donné des réponses très différentes, en fonction de leur discipline, de leur objet de recherche et de la méthodologie employée.

13 « 'Digital research data' is information in digital form (in particular facts or numbers), collected to be examined and used as a basis for reasoning, discussion or calculation; this includes statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images », dans COMMISSION EUROPÉENNE (2019). *H2020 Programme: Annotated Model Grant Agreement.* Version 5.2. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf (consulté le 18 septembre 2019)

Traduction s'appuyant sur celle délivrée par l'Inist-CNRS (https://www.enssib.fr/bibliotheque-numerique/notices/66433-lignes-directrices-pour-le-libre-acces-aux-publications-scientifiques-et-aux-donnees-de-la-recherche-dans-horizon-2020).

14 Voir supra (p.26)

Francisca Cabrera synthétise cette diversité de réponses dans une carte heuristique (figure 1), qui témoigne bien de l'hétérogénéité des données de la recherche et de la difficulté à embrasser ce concept en une simple définition de quelques lignes. Définir les données par le recensement exhaustif de ce qu'elles recouvrent aboutirait donc probablement à un résultat peu lisible.

1.3. Définitions sous forme de typologies

Une autre approche permettant d'aboutir à un panorama global des données est d'en établir une typologie. Il existe différentes manières de classer les données de la recherche, selon l'angle de vue sous lequel on les étudie.

1.3.1. Typologie selon l'approche méthodologique

Francisca Cabrera a elle-même proposé deux typologies possibles. La première consiste à classer les données selon l'approche méthodologique employée. Deux approches distinctes sont identifiées pour les sciences humaines et sociales (Cabrera 2014, p.54-57) :

- L'approche herméneutique et textuelle, qui concerne en particulier des disciplines comme le droit, l'histoire et la philosophie. Ces disciplines manipulent essentiellement des textes (juridiques, historiques ou philosophiques), qu'elles analysent et interprètent. Ces textes ne sont d'ailleurs pas appelés « données » par les chercheurs, mais plutôt « matériaux ».
- L'approche expérimentale et de terrain, caractéristique de disciplines telles que l'archéologie, l'économie, la géographie, la linguistique, la psychologie ou la sociologie. Ces disciplines sont souvent productrices de données, qu'elles collectent sur le terrain ou à partir d'enquêtes ou de tests.

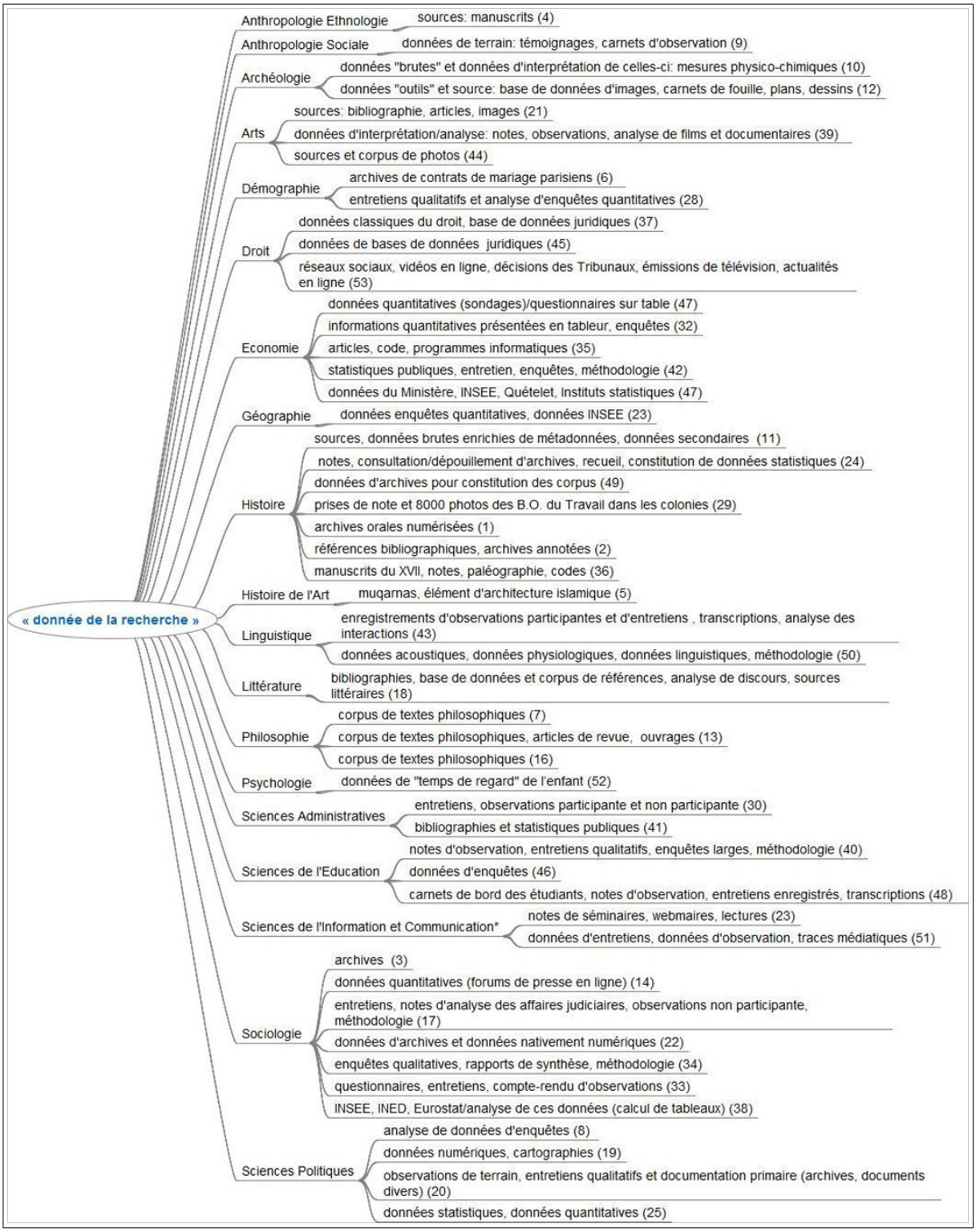


Figure 1 : Réponses des chercheurs interrogés par F. Cabrera à la question « Dans le cadre de votre pratique, qu'est-ce qu'une donnée de la recherche ? »[15]

15 Carte heuristique réalisée par F. Cabrera (2014, p.60)

Cette typologie est à rapprocher de celle de la National Science Foundation[16], qui s'appuie également sur la manière dont les données ont été collectées. Trois catégories de données la composent :

- Première catégorie – les données observationnelles (*observational data*) : ces données résultent de l'enregistrement d'un fait en temps réel. En sciences et technologies, il peut s'agir d'une mesure sismique ou de l'image d'une étoile en fin de vie. En sciences humaines et sociales, ce sont par exemple des notes d'observations ethnographiques ou les résultats d'un sondage de population. Ces données ont pour caractéristique d'être liées à un moment et à un lieu spécifiques. Elles ne sont donc pas reproductibles.
- Deuxième catégorie – les données computationnelles (*computational data*) : ce sont les données résultant de l'exécution d'une simulation ou d'un modèle numérique. Les économistes, par exemple, utilisent des modèles pour simuler des phénomènes comme la croissance économique. En physique, les chercheurs élaborent plutôt des modèles relatifs à des phénomènes naturels par exemple.
- Troisième catégorie – les données d'expérimentation (*experimental data*) : il s'agit de données produites dans un environnement contrôlé. En chimie, une donnée d'expérimentation peut être le résultat d'une réaction effectuée en laboratoire. En psychologie expérimentale, ce sont par exemple des observations relatives au comportement d'individus en situation de test.
- Christine L. Borgman (2015, p.24) ajoute à cette typologie une quatrième catégorie de données, qu'elle intitule *records*. Cette catégorie regroupe tous les documents témoignant d'un phénomène ou d'une activité humaine. Ce sont des documents qu'un chercheur pourra être amené à collecter et à utiliser comme données dans ses recherches. Il peut s'agir de textes de lois, d'ouvrages littéraires, de documents d'archives ou de tout autre document textuel, audio ou vidéo d'origine publique ou privée. Dans ce contexte, « *record* » peut être traduit en français par le terme d'« enregistrement » ou de « source ».

16 NATIONAL SCIENCE BOARD (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Foundation. https://www.nsf.gov/pubs/2005/nsb0540/ (consulté le 15 septembre 2019). Page 19

1.3.2. Typologie selon le niveau de traitement des données

Autre typologie possible : celle qui traduit le niveau de traitement de la donnée. Les processus de recherche conduisent généralement à une transformation des données. Les données initiales, telles qu'elles ont pu être décrites dans la typologie ci-dessus, sont en effet souvent l'objet d'un travail de curation, de combinaison et d'analyse, avant d'être interprétées. On distingue alors données « brutes » (en entrée) et données « dérivées » (en sortie). Les étapes de traitement sont parfois multiples, aboutissant à une succession de versions de fichiers.

Le système d'observation de la Terre de la NASA a mis en place une typologie en fonction du degré de traitement des données. Cette typologie est composée de six niveaux (tableau 1). Le niveau 0 correspond aux données « brutes », c'est-à-dire aux signaux délivrés par l'instrument de télédétection dont elles sont issues. Les données de niveaux 1A et 1B sont ces mêmes signaux, auxquels ont été ajoutées des métadonnées de référencement temporel et géographique et de paramétrage de l'instrument. Les niveaux 2, 3 et 4 font référence à des données traitées encore plus avant, destinées à être comparées ou intégrées dans des modèles[17].

1.3.3. Typologie selon l'origine des données

Enfin, une autre manière de classer les données de la recherche consiste à différencier les données utilisées par les chercheurs, des données produites par les chercheurs. En effet, la recherche ne s'appuie pas seulement sur les données qu'elle produit (en laboratoire ou sur le terrain). Dans certaines disciplines, comme la démographie ou la géographie, les chercheurs utilisent des données préexistantes, parfois acquises hors du cercle de la recherche à des fins tout autres que scientifiques. Les démographes, par exemple, vont être amenés à utiliser les données produites par des instituts de statistique publique comme l'INSEE. Les géographes, quant à eux, utilisent les données générées par de grandes infrastructures nationales ou européennes, disposant d'importants moyens de télédétection (avions, satellites). Il est fréquent qu'un projet de recherche mobilise à la fois des données préexistantes et des données nouvelles produites spécifiquement pour le projet.

17 Cet exemple de classification par degré de traitement est décrit dans l'ouvrage *Big data, little data, no data* de C. L. Borgman (2015, p.21-23).

Data Level	Description
Level 0	Reconstructed, unprocessed instrument and payload data at full resolution, with any and all communications artifacts (e.g., synchronization frames, communications headers, duplicate data) removed. (In most cases, NASA's EOS Data and Operations System [EDOS] provides these data to NASA's Distributed Active Archive Centers [DAACs] as production data sets for processing by NASA's Science Data Processing Segment [SDPS] or by one of NASA's Science Investigator-led Processing System [SIPS] to produce higher-level products.)
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters (e.g., platform ephemeris) computed and appended but not applied to Level 0 data.
Level 1B	Level 1A data that have been processed to sensor units (not all instruments have Level 1B source data).
Level 2	Derived geophysical variables at the same resolution and location as Level 1 source data.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
Level 4	Model output or results from analyses of lower-level data (e.g., variables derived from multiple measurements).

Tableau 1 : Degrés de traitement des données produites par le système d'observation de la Terre de la NASA[18]

18 Source : https://earthdata.nasa.gov/earth-science-data-systems-program/policies/data-information-policy/data-levels

Francisca Cabrera (2014, p.60-61) propose ainsi une seconde typologie distinguant données utilisées et données produites en sciences humaines et sociales.

Les données utilisées regroupent :

- Des documents relevant des archives ;
- Des documents issus de la production scientifique, publiés ou non (articles de revue, ouvrages, littérature grise[19]) ;
- Des sources matérielles ou primaires (objets, tweets, corpus de textes) ;
- Des données de référencement (de type bibliographies) ;
- Des données statistiques issues de l'administration publique.

Les données produites rassemblent :

- Les données générées par un travail de terrain (des données d'entretiens par exemple ou des données issues de démarches expérimentales, comme en psychologie) ;
- Les données nées d'un travail de transformation sur des données brutes de départ (cela peut consister à numériser des archives, structurer des données en une base de données, modéliser des données du web sous forme de graphes…).

Ces définitions typologiques tentent de saisir les données par leur plus petit dénominateur commun, dans une perspective souvent opérationnelle de traitement des données (comme c'est le cas pour le système d'observation de la Terre de la NASA). Toutefois, selon Christine L. Borgman (2012, p.1062), ces définitions tendent à *« obscurcir la diversité des données »*, car l'attribution d'une catégorie n'est jamais que le résultat d'une décision arbitraire – aussi judicieuse soit-elle. Le terme « donnée de recherche » semble donc être une notion complexe, que ni les producteurs de données, ni les institutions qui en demandent l'ouverture ne parviennent à définir sous une forme élémentaire et unanimement reconnue. La raison de cette complexité est peut-être à chercher dans le concept fondamental de « données ».

19 Selon la définition dite « de Luxembourg », approuvée lors de la troisième conférence internationale sur la littérature grise en 1997, la littérature grise est « ce qui est produit par toutes les instances du gouvernement, de l'enseignement et la recherche publique, du commerce et de l'industrie, sous un format papier ou numérique, et qui n'est pas contrôlé par l'édition commerciale » (Schöpfel 2012).

2. Vers une non définition

2.1. Le concept général de « donnée » : un concept fuyant

2.1.1. Étymologie

En français comme en anglais, le terme « donnée » (ou « *datum* ») vient du verbe latin *dare* (donner). « *Datum* » est la forme neutre du participe passé de ce verbe et signifie littéralement « ce qui a été donné ». « *Data* » est la forme au pluriel. L'anglais a donc repris l'orthographe exacte du terme d'origine. Dans un article fondé sur une analyse de corpus de textes, Daniel Rosenberg (2013) étudie l'introduction du mot « *data* » dans la langue anglaise. Bien que spécifique à l'histoire anglo-saxonne (il n'existe pas, à ce jour, d'étude similaire sur le terme français de « donnée »), l'analyse de Rosenberg n'en demeure pas moins intéressante pour comprendre le sens du mot « donnée » dans son acception actuelle. En anglais, « datum » désigne quelque chose qui est tenu pour acquis (qui est « donné ») dans une démonstration. Autrement dit, il s'agit du postulat de départ sur lequel repose un raisonnement. Rosenberg confronte ce terme à celui de « fait » (*fact*). Le mot « *fact* » a pour origine le participe passé du verbe latin *facere* (faire). Il signifie donc littéralement : « ce qui a été fait », ce qui s'est produit ou ce qui existe. Ainsi, par définition, les faits seraient ontologiques (ils sont réels donc vrais) et les données rhétoriques (car liées au discours). Rosenberg constate à partir de son corpus de textes qu'effectivement l'occurrence du terme « datum » n'est jamais reliée à des considérations de vérité ontologique. *« Quand on prouve qu'un fait est faux, il cesse d'être un fait. Une donnée fausse, en revanche, n'en demeure pas moins une donnée »*[20]. Aujourd'hui encore il n'y a pas de lien obligé entre donnée et vérité. Le terme « donnée » a gardé sa dimension rhétorique de postulat dans une démonstration. Selon Rosenberg, c'est cette dimension rhétorique qui a rendu le terme si prégnant dans notre monde d'aujourd'hui, où la communication occupe une place centrale.

20 Traduction de : « When a fact is proven false, it ceases to be a fact. False data is data nonetheless » (Rosenberg 2013, p.18)

2.1.2. Donnée, Information, Connaissance

Les données sont souvent associées au triptyque « donnée, information, connaissance ». Le philosophe Sven Ove Hansson (2002) décrit l'articulation entre ces trois notions à partir de l'exemple d'un ouvrage de sociologie :

> *« Les données diffèrent de l'information en ce qu'elles n'ont pas à se présenter sous une forme qui se prête à l'assimilation. Si au lieu de l'ouvrage [de sociologie que je suis en train de lire], j'avais sur mon bureau les dix mille questionnaires sur lesquels il repose, j'aurais des données au lieu d'information. En résumé, il faut que des données soient assimilables pour pouvoir constituer de l'information et qu'elles soient assimilées pour pouvoir constituer du savoir. »*

En sciences de l'information, Chaim Zins (2007) distingue « donnée », « information » et « connaissance », en s'appuyant sur le concept de connaissance propositionnelle. Théorisée par la philosophie, la connaissance propositionnelle est la pensée ou l'expression de ce qu'une personne pense qu'elle sait (Bernecker et Dretske 2000). Elle prend habituellement la forme de « il sait que + proposition ». Elle se distingue du « savoir-faire » (capacité à réussir une action) et de la « connaissance directe » (fait de connaître une personne, un lieu ou une chose).

La connaissance propositionnelle peut provenir :

- De la compréhension intuitive d'un phénomène (*non-inferential knowledge*) ;
- D'un raisonnement inductif ou déductif (*inferential knowledge*).

Dans la sphère académique, par exemple, les connaissances publiées dans les ouvrages et les articles scientifiques sont le fruit de raisonnements inductifs ou déductifs.

C'est sur ce type de connaissance propositionnelle, issue du raisonnement, que Chaim Zins fonde sa définition de la donnée. Selon lui, la donnée a deux modes d'existence : elle existe dans la sphère subjective et dans la sphère objective. La « connaissance subjective » renvoie à la connaissance du sujet (c'est-à-dire à la connaissance de l'individu qui sait). Elle se limite au for intérieur de l'individu. La « connaissance objective » (ou « connaissance collective »)

équivaut, quant à elle, à la connaissance en tant qu'objet ou chose. Elle est présente dans le monde extérieur à l'individu : c'est par exemple un article publié dans une revue scientifique.

Zins distingue la donnée des concepts d'information et de connaissance, au sein de ces deux sphères subjective et collective.

Dans la sphère subjective :

- Les données sont des stimuli sensoriels (une perception empirique). Zins prend l'exemple d'une voiture qui démarre : le bruit que l'individu perçoit (celui du moteur) est un stimulus sensoriel, donc une donnée.
- L'information, quant à elle, est une connaissance empirique. Pour reprendre l'exemple précédent, la connaissance qu'une voiture démarre est une information. Zins considère l'information comme un type de connaissance, plutôt que comme un niveau intermédiaire entre la donnée et la connaissance.
- Enfin, la connaissance est une pensée que l'individu considère comme vraie.

Dans la sphère collective, données, informations et connaissances sont des artefacts humains, représentés par des signes empiriques (c'est-à-dire par des signes que chacun peut percevoir par le biais de ses sens). Ces signes peuvent prendre la forme d'inscriptions gravées, de formes peintes, de caractères imprimés, de signaux numériques, de rayons lumineux, d'ondes… Dans la sphère collective :

- Une donnée est un ensemble de signes représentant des stimuli sensoriels.
- Une information est un ensemble de signes représentant une connaissance empirique.
- Une connaissance est un ensemble de signes représentant le contenu d'une pensée que l'individu considère comme vrai.

Selon Zins, les données font donc partie du domaine de la connaissance. La définition qu'il en donne n'en demeure pas moins une proposition parmi d'autres, ayant lui-même recueilli une variété de définitions auprès d'un panel de 44 chercheurs en sciences de l'information.

Plus récemment, Evelyne Broudoux (2018) s'est attachée à observer l'articulation entre les concepts d'information, de donnée et de connaissance, telle que celle-ci est pensée dans le

domaine des sciences de l'information et de la communication. Elle s'appuie notamment sur les travaux de Luciano Floridi et de Marcia J. Bates. Dans sa Définition Générale de l'Information (GDI), Floridi présente la donnée comme une entité symbolique qui différencie l'information (Floridi 2005). Quant à Bates (2005), elle propose une approche par « motifs », dans laquelle :

- L'information 1 est un *« motif d'organisation de matière et d'énergie »* (l'information n'est pas le matériel lui-même mais un motif organisationnel) ;
- L'information 2 est un motif d'organisation de matière et d'énergie *« auquel un être vivant accorde une signification »* ;
- La connaissance est l'information 2 s'intégrant aux connaissances pré-existantes ;
- La donnée 1 est une partie d'un environnement informationnel accessible à un organisme qui est intégrée ou traitée par lui ;
- La donnée 2 est une information sélectionnée ou générée par un être humain pour des objectifs sociétaux.

Ces deux définitions (Floridi 2005 ; Bates 2005) présentent la donnée comme une entité indivisible, n'acquérant d'utilité que par son association dans un contexte de production d'informations. E. Broudoux constate cependant que la définition informatique de la donnée tend aujourd'hui à s'imposer comme étant porteuse de sens, conduisant à réinterroger ses rapports avec l'information.

> *« Données, informations et connaissances sont des concepts que l'on relie habituellement ensemble selon un principe d'intégration et de construction mais les données acquièrent leur territoire propre et sont susceptibles de traitement au même titre que les documents. »* (Broudoux 2018, p.51)

2.1.3. « Information as thing »

Michael K. Buckland (1991) classe les données dans la catégorie des « informations en tant que choses » (*information as thing* – que l'on traduira ici par « entités informationnelles »).

Les entités informationnelles se rapportent à toutes les choses auxquelles on attribue le terme d'information, parce qu'on considère qu'elles sont source d'information.

L'information en tant que chose se différencie des deux autres usages qui peuvent être faits du terme « information », à savoir :

- L'information en tant que processus (*information as process*) : c'est l'action d'informer, de communiquer la nouvelle d'un fait.
- L'information en tant que connaissance (*information as knowledge*) : il s'agit du contenu de ce qui est communiqué par l'information-processus. L'information en tant que connaissance est l'information communiquée sur un sujet, un fait ou un évènement particulier.

L'information en tant que connaissance est intangible (on ne peut pas la toucher). Pour pouvoir la communiquer, il faut la représenter sous une forme physique (un discours, un texte, un signal…). Cette représentation constitue une entité informationnelle (une information en tant que chose). Sa caractéristique fondamentale est d'être informative.

Les données, au même titre que les documents, les objets et les événements, sont des entités informationnelles, car elles sont informatives. M. K. Buckland constate que toute chose peut être informative et que tout est susceptible d'être information.

Buckland assimile les entités informationnelles à des « preuves » (*evidences*). Nos connaissances et nos opinions sont influencées par ce que nous voyons, lisons, entendons ou vivons. Cela ne signifie pas que nous y adhérons. Cela n'implique pas non plus nécessairement que ce que nous avons vu, lu ou perçu d'une quelconque manière est pertinent pour ce que nous recherchons. Un objet ou un événement constitue donc en cela une « preuve » susceptible, dans un contexte spécifique, d'influencer notre jugement sur tel ou tel sujet.

Une entité informationnelle possède, par conséquent, la particularité d'être situationnelle. Sa pertinence dépend des circonstances dans lesquelles elle est perçue. La capacité d'une donnée ou d'un événement à être informatif est fonction de la question que se pose la personne qui

l'observe et de l'expertise de cette dernière. La qualité informative d'une chose est donc affaire de jugement et de contexte.

2.2. Les données de recherche se définissent relativement à un contexte épistémologique

La difficulté à définir ce qu'est une donnée de recherche pourrait venir de cette dimension situationnelle inhérente à toute donnée. La diversité des définitions proposées par la littérature (et présentées en début de partie, p.25-34) montre combien il est complexe d'identifier ce que sont les données de recherche. Il semble impossible de les définir dans l'absolu.

2.2.1. Quand une entité devient-elle une donnée ?

C'est ce qu'a tenté d'illustrer Christine L. Borgman dans ses travaux. Elle montre, à partir d'études de cas réalisées dans diverses disciplines scientifiques (Borgman 2015), que les données de recherche englobent une myriade d'objets informationnels[21]. Leur nature varie en fonction de la discipline scientifique, de l'objectif de la recherche, de la méthodologie et de l'instrumentation utilisées…

Plus précisément, Borgman (2012, p.1062-1064) isole trois variables qui, selon elle, ont une influence sur la nature des données collectées.

- La finalité de la collecte de données (*specificity of purpose*) : la collecte de données a-t-elle lieu dans le cadre d'un projet de recherche particulier, avec une question de recherche particulière, ou bien s'agit-il d'observer un phénomène sur le long terme, en collectant des données de manière systématique ?
- L'étendue de la collecte de données (*scope of data collection*) : la collecte est-elle limitée aux données décrivant un événement ou un phénomène particulier ou bien vise-t-elle à rassembler des données sur un système dans son entier ?

21 Six études de cas sont développées dans le chapitre « *Case Studies in Data Scholarship* » (Borgman 2015, p.81-202)

- Le but de la recherche (*goal of research*) : la recherche vise-t-elle à étudier un terrain particulier (recherche empirique) ou bien essentiellement des lois, des principes, des concepts (recherche théorique) ?

Borgman spécifie que ces variables ne sont pas exhaustives et qu'elles peuvent être complémentaires. Pour mieux figurer les multiples formes que celles-ci peuvent prendre, elle les représente dans un repère à trois dimensions (figure 2). Chaque axe va du plus local et flexible au plus global et standardisé. Pour une recherche exploratoire, par exemple, un chercheur souhaitera plutôt collecter de petits jeux de données qui décrivent des évènements ou des phénomènes particuliers. A l'inverse, si le but de la recherche est de modéliser un système dans son ensemble, le chercheur aura besoin de rassembler une collection homogène d'importants volumes de données.

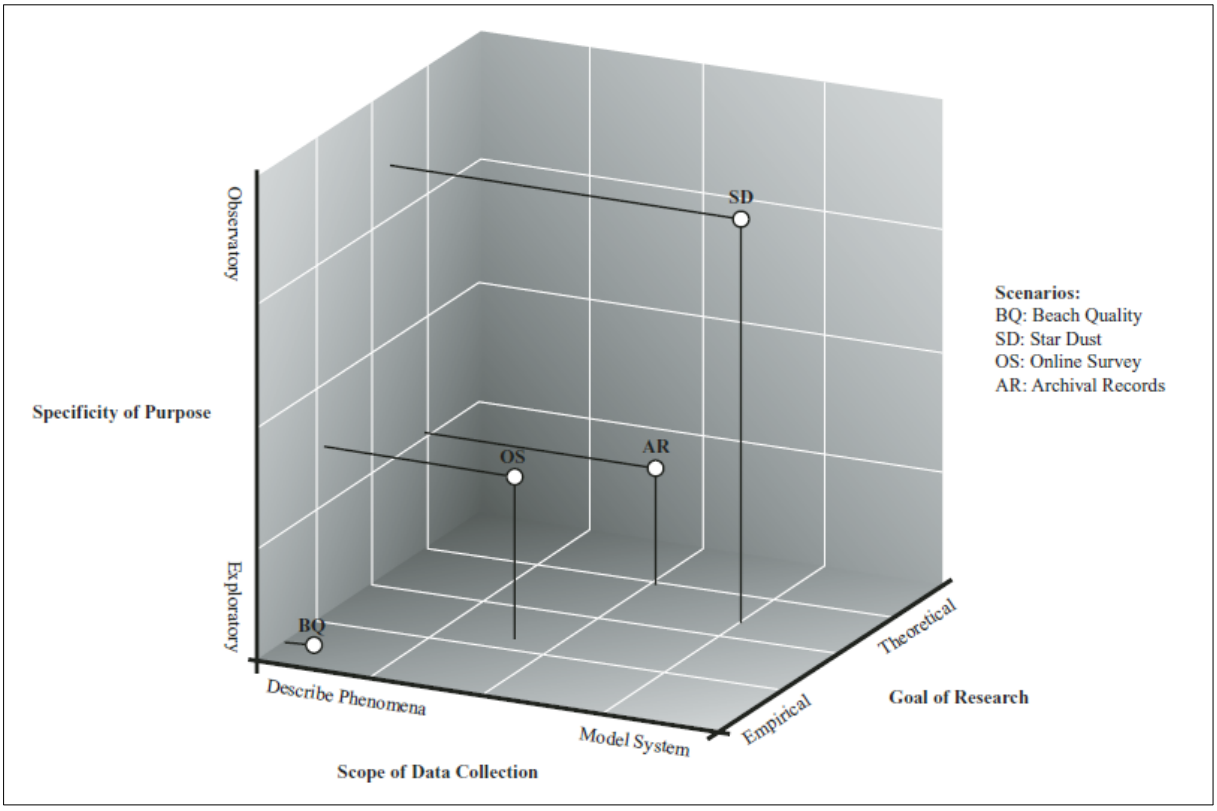


Figure 2 : Variables influençant le type de données collectées[22]

22 Source : Borgman 2012, p.1063

Christine L. Borgman en conclut qu'il est moins pertinent de tenter d'identifier ce qu'est une donnée de recherche que de s'interroger sur les circonstances dans lesquelles une entité devient une donnée. Cela revient à poser la question : à quel moment une entité devient-elle une donnée ? C'est-à-dire par quel processus un objet ou un fait en vient-il à être considéré comme une donnée scientifique ? De manière générale, il est possible de dire que *« la représentation d'une observation, d'un objet ou de toute autre entité devient une donnée de recherche dès lors qu'elle est utilisée comme preuve d'un phénomène, à des fins scientifiques »* (Borgman 2015)[23]. Les données ne sont pas des objets purs ou naturels ayant une essence propre. Elles existent dans un contexte et prennent leur sens en fonction de ce contexte et de la perspective de l'observateur.

2.2.2. Les données comme preuves

Sabina Leonelli (2015) partage cette idée que la donnée est relative à un contexte. Elle se refuse d'ailleurs à produire une définition qui permettrait d'étudier le concept de donnée indépendamment d'un contexte spécifique. Elle s'oppose à la conception selon laquelle les données sont des *« entités représentatives, contenant une information fixe et dépeignant une partie spécifique de la réalité, indépendamment des circonstances dans lesquelles elles sont considérées »*[24].

Selon Leonelli, il n'existe pas de données en elles-mêmes. Ce qu'un scientifique considère comme « donnée » est toujours relatif à une question de recherche spécifique. Les données ne sont pas définies selon leurs propriétés intrinsèques mais selon leur fonction au sein de processus de recherche particuliers. C'est donc dans un cadre circonstanciel que Leonelli pose le concept de donnée. On ne peut répondre à la question « qu'est-ce qu'une donnée » qu'en faisant référence à des situations de recherche concrètes. Ainsi, un même ensemble de données est susceptible d'être utilisé pour expliquer de multiples phénomènes.

23 Traduction de : « Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship. […] Entities become data only when someone uses them as evidence of a phenomenon. » (Borgman 2015, p.28)

24 Traduction de : « Representational entities with a fixed information content, which depict a specific part of reality independently of the circumstances under which they are considered » (Leonelli 2015, p.818).

Leonelli prend l'exemple de photographies d'embryons de poulets prises à différentes étapes de leur développement. Dans le cas de recherches en biologie, ces photographies sont prises pour documenter les caractéristiques d'une entité naturelle (le poulet). Les données sont donc ici des *« traces laissées par le processus de mesure et de manipulation d'échantillons organiques, entrepris dans le cadre d'expériences biologiques »*[25]. Elles ne sont pas la représentation directe d'un phénomène. Elles constituent seulement une manière (parmi d'autres) d'exprimer une information relative à ce phénomène.

Leonelli propose de voir les données scientifiques comme des produits de la recherche collectés, enregistrés et diffusés dans la perspective d'être utilisés ensuite comme preuve d'une théorie scientifique au sujet d'un phénomène particulier. Celui qui collecte une donnée ne sait pas forcément quel phénomène elle décrit, au moment où il la collecte. Ce qui importe, c'est qu'il la collecte dans l'idée que peut-être elle servira de preuve pour l'assertion scientifique qu'il développera. Les données de la recherche se définissent donc par leur utilité potentielle en tant que preuve.

Le chercheur joue un rôle important dans la détermination de ce qui constitue une donnée. C'est lui qui décide de considérer telle entité comme une donnée, c'est-à-dire comme étant la preuve de tel phénomène. Par exemple, il choisira de collecter ou non des photographies d'embryons, pour décrire tel ou tel phénomène biologique. Les données apparaissent donc comme « fabriquées » par les chercheurs. Leonelli considère qu'elles sont le résultat d'interactions entre les chercheurs et le monde qu'ils observent. Ces interactions ont lieu par le biais de techniques d'observations, d'outils d'enregistrement ou d'instruments de mesure. Elles visent à rendre l'objet d'étude « manipulable » et « étudiable ». Les données sont donc des artefacts. Elles sont le fruit de manipulations humaines. C'est le cas des données de terrain comme des données issues d'expérimentations ou de simulations.

Les données de la recherche possèdent par ailleurs une caractéristique qui leur est propre : leur « portabilité ». Leonelli considère en effet les données comme des outils de communication, dont la fonction principale est de permettre les échanges intellectuels et matériels entre individus (Leonelli 2015, p.817-819). Leur portabilité est une condition

25 Traduction de : « They are traces left by the process of measurement and manipulation of organic samples undertaken in a biological experiment » (Leonelli 2015, p.811-813).

nécessaire à leur utilisation comme preuve. Dans le domaine de la recherche, les connaissances scientifiques s'établissent en effet sur la base d'activités sociales. La valeur d'une théorie scientifique n'est établie que si celle-ci a été soumise au jugement des pairs. Ces derniers évaluent notamment la façon dont les données ont été collectées et analysées au cours du projet de recherche, afin de déterminer leur qualité et leur fiabilité en tant que preuve. D'où la notion de portabilité des données. Si les données ne sont pas portables, elles ne pourront être transmises aux pairs chargés de vérifier leur valeur scientifique. Cela implique des étapes de manipulation variées, susceptibles de modifier le format et le support des données. Selon Leonelli (2013a), c'est quelque chose qui est pensé dès l'étape de production des données, orientant le choix des instruments, des méthodes de collecte et d'enregistrement...

La conception de Leonelli fait écho à l'origine étymologique du mot « donnée ». De même que le terme « *datum* » renvoie au postulat d'une démonstration dans un cadre rhétorique, les données de la recherche jouent le rôle de preuve au sein d'un système – la recherche scientifique – fondé sur la communication des résultats.

2.2.3. Une définition multidimensionnelle

Constatant, tout comme Borgman et Leonelli, le caractère relatif des données de la recherche, Schöpfel et al. (2017a) posent les bases d'une définition multidimensionnelle.

Quatre éléments composent cette base de définition (figure 3) :

- **L'enregistrement**
  Une donnée de recherche est avant tout une entité informationnelle qui a été fixée sur un support matériel (physique ou numérique). Elle relève de ce que Chaim Zins appelle la *« sphère objective de la connaissance »*[26].

- **La nature factuelle des données**
  Les données peuvent être de natures extrêmement diverses, allant de la séquence d'un gène à des relevés pluviométriques, en passant par l'enregistrement audio d'un dialecte.

26 Voir supra (2.1.2, p.35)

- **Le lien avec la communauté scientifique**

  Les qualificatifs de données « brutes », « secondaires » ou « dérivées » mettent en évidence leur caractère dynamique. Les données sont des objets évolutifs, car elles sont au cœur même des processus de recherche. Le lien avec la communauté scientifique est donc fort. Le terme « communauté » est entendu par Schöpfel et al. comme un groupe partageant des valeurs, des concepts, des pratiques et des outils communs. Il existerait au sein de chaque communauté scientifique une forme de consensus autour de ce qui constitue une donnée de recherche pour la communauté.

- **La finalité**

  Les données ont une fonction première au sein du processus scientifique : selon l'OCDE (entre autres), elles servent à valider les résultats de recherche ; pour Sabina Leonelli, elles sont une preuve potentielle d'une assertion scientifique. Elles peuvent également avoir des finalités secondaires, comme celles qui leur sont attribuées par le mouvement d'ouverture : à savoir informer les citoyens et les autorités publiques sur leur environnement, ou encore catalyser l'innovation.

Le cadre délimité par Schöpfel et al. imbrique donc concepts, typologies et éléments contextuels.

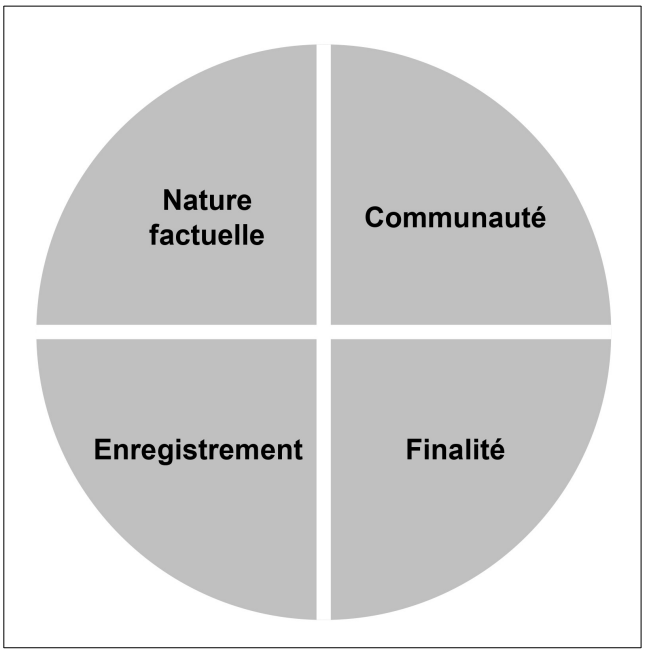


Figure 3 : Définition quadridimensionnelle des données de recherche[27]

27 Cette figure reprend en partie le schéma proposé par Schöpfel et al. (2017a, p.6).

3. Conclusion

Pour conclure, citons la définition adoptée par les organes de la recherche britannique dans le *Concordat on Open Research Data*, synthétisant les différents aspects des données mis en évidence dans cette partie :

> *« Les données de la recherche sont les preuves qui sous-tendent la réponse à une question de recherche. Elles peuvent être utilisées pour valider les résultats, quelle que soit leur forme (imprimée, numérique ou physique). Il peut s'agir d'informations quantitatives ou qualitatives collectées par les chercheurs au cours de leur travail par le biais de l'expérimentation, l'observation, la modélisation, l'enquête ou tout autre méthode. Il peut aussi s'agir d'informations dérivées de preuves existantes. Les données peuvent être : brutes ou primaires (par exemple être le résultat direct d'une mesure ou d'une collecte) ; dérivées de données primaires en vue de leur analyse ou de leur interprétation (par exemple être nettoyées ou extraites d'un jeu de données plus grand) ; dérivées de sources existantes, dont les droits sont détenus par des tiers. Les données peuvent être définies comme des composants « relationnels » ou « fonctionnels » de la recherche, dans la mesure où leur identification et leur valeur dépendent de leur utilisation par les chercheurs en tant que preuves d'une assertion scientifique.*
>
> *Les données sont par exemple des statistiques, des collections d'images numériques, des enregistrements audio, des transcriptions d'entretiens, des données d'enquêtes et des observations de terrain accompagnées d'annotations, une performance, une œuvre d'art, des archives, des objets trouvés, des textes publiés ou un manuscrit.*
>
> *La finalité première des données de recherche est d'apporter des informations permettant de soutenir ou de valider les observations, les résultats ou les publications d'un projet de recherche. »*[28]

28 Traduction de : « Research data are the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Data may be raw or primary (e.g. direct from measurement or collection) or derived from primary data for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger data set), or

La présente recherche ne vise pas à élaborer une nouvelle définition des données de la recherche. Elle portera certes une attention particulière à ce que les acteurs considèrent comme des données, mais elle n'a pas pour objectif de retravailler le concept de données de la recherche.

Elle s'appuiera donc sur les différents aspects mis en évidence dans les travaux précités, à savoir :

- La notion de « donnée de la recherche » est une notion complexe. Cette complexité repose en partie sur la difficulté à définir le terme initial de « donnée ».
- La donnée relève du domaine de la connaissance. Elle est une entité informationnelle, se matérialisant sur un support physique ou numérique.
- Elle possède un lien étymologique avec le discours, en particulier avec la forme de l'argumentation. Une des premières acceptions du terme « donnée » en anglais est « *datum* », désignant le postulat de départ d'une démonstration logique.
- La notion de « donnée de recherche » est une notion contextuelle. La donnée n'est pas un objet ayant une essence propre. Elle prend son sens en fonction du contexte, dans lequel elle existe.
- Les données sont collectées par les chercheurs dans la perspective d'être utilisées comme preuve d'une théorie scientifique au sujet d'un phénomène particulier. Afin d'assurer cette fonction auprès des pairs, elles sont rendues « portables », c'est-à-dire présentées sous une forme adéquate à leur communication.

---

derived from existing sources where the rights may be held by others. Data may be defined as 'relational' or 'functional' components of research, thus signalling that their identification and value lies in whether and how researchers use them as evidence for claims.
They may include, for example, statistics, collections of digital images, sound recordings, transcripts of interviews, survey data and fieldwork observations with appropriate annotations, an interpretation, an artwork, archives, found objects, published texts or a manuscript.
The primary purpose of research data is to provide the information necessary to support or validate a research project's observations, findings or outputs. », dans RESEARCH COUNCILS UK (2016). *Concordat on Open Research Data*. https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/ (consulté le 18 septembre 2019).

Deuxième partie

-

Les politiques publiques de gestion et d’ouverture des données de la recherche

La première partie a montré que les données de recherche avaient pour la première fois été définies dans un contexte d'ouverture. Cette démarche révèle la construction politique dont elles sont l'objet. Cette partie sera consacrée aux politiques publiques de l'Union européenne et de l'État français, qui progressivement contribuent à instaurer l'ouverture des données de la recherche. Décrire ces mesures permettra de nourrir une réflexion sur ce qui est de l'ordre d'un hiatus entre intentions politiques et pratiques de recherche. Il s'agira, dans un premier temps, d'exposer les philosophies à l'origine des politiques d'ouverture (du libre accès aux résultats scientifiques jusqu'à l'ouverture des données publiques, en passant par l'économie fondée sur la connaissance). Seront ensuite décrites les mesures prises par l'Union européenne et par l'État français en la matière.

1. Mouvements à l'origine des politiques de gestion et d'ouverture des données de la recherche

1.1. Déclaration de Berlin (2003)

La Déclaration de Berlin[29] est issue du mouvement du libre accès (*Open Access*), promouvant l'accès à la recherche sans barrière financière, légale ou technique. Initié dès 1991 par des membres de la communauté scientifique[30] dans un but de communication directe, ce mouvement a été porté à l'échelle mondiale à partir de la conférence de Budapest, en 2002[31]. Les données de la recherche ont été incluses dans le débat du libre accès dès 2003, avec la Déclaration de Berlin. Initialement soutenue par 19 institutions de recherche, cette déclaration élargit le concept de libre accès à l'ensemble de la production scientifique :

> *« Les contributions au libre accès se composent de résultats originaux de recherches scientifiques, de données brutes et de métadonnées, de documents sources, de représentations numériques de documents picturaux et graphiques, de documents scientifiques multimédia. »*[32]

Le discours en faveur du libre accès à la production scientifique est aujourd'hui porté par la communauté scientifique, les institutions de recherche, les agences de financement et les ministères. Il défend les valeurs traditionnelles de la science, telles que les a définies Robert

29 MAX PLANCK GESELLSCHAFT (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. http://openaccess.mpg.de/Berlin-Declaration (consulté le 18 septembre 2019).

30 La première archive ouverte, arXiv, a été imaginée par le physicien Paul Ginsparg en 1991. Elle a inspiré la « proposition subversive » de Stevan Harnad, professeur en sciences cognitives, incitant les chercheurs à déposer leurs *preprints* dans des archives ouvertes (Harnad 1994).

31 *Budapest Open Access Initiative* (2002). https://www.budapestopenaccessinitiative.org/read (consulté le 16 septembre 2019).

32 « Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material » (Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities 2003, op. cit.).

Traduction disponible en ligne sur https://openaccess.mpg.de/68042/BerlinDeclaration_wsis_fr.pdf.

K. Merton (1973)[33]. Dans le discours du libre accès, la connaissance scientifique est en effet considérée comme un bien commun.

Le mouvement est né dans un contexte où l'Internet ouvrait de nouvelles possibilités pour diffuser la connaissance scientifique. Celui-ci était perçu comme le moyen de communication idéal pour diffuser rapidement et à moindre coût les résultats de la recherche. L'Initiative de Budapest explicite ainsi le recours à l'Internet comme outil de partage des connaissances scientifiques :

> *« Une tradition ancienne et une technologie nouvelle ont convergé pour rendre possible un bienfait public sans précédent. La tradition ancienne est la volonté des scientifiques et universitaires de publier sans rétribution les fruits de leur recherche dans des revues savantes, pour l'amour de la recherche et de la connaissance. La nouvelle technologie est l'Internet. Le bienfait public qu'elles rendent possible est la diffusion électronique à l'échelle mondiale de la littérature des revues à comité de lecture avec accès complètement gratuit et sans restriction à tous les scientifiques, savants, enseignants, étudiants et autres esprits curieux. Supprimer les obstacles restreignant l'accès à cette littérature va accélérer la recherche, enrichir l'enseignement, partager le savoir des riches avec les pauvres et le savoir des pauvres avec les riches, rendre à cette littérature son potentiel d'utilité, et jeter les fondements de l'unification de l'humanité à travers un dialogue intellectuel, et une quête du savoir communs. »*[34]

33 Quatre normes, selon Merton (1973, p.268-278), forment l'« ethos de la science » :
- l'universalisme : les énoncés soumis à la communauté scientifique doivent être évalués selon des critères impersonnels, sans égard pour les caractéristiques sociales ou institutionnelles de leurs auteurs ;
- le communisme : en tant que produit de la collaboration entre chercheurs, toute découverte scientifique est un bien commun et doit donc systématiquement être publiée ;
- le désintéressement : les scientifiques cherchent de nouvelles connaissances non pas dans leur intérêt personnel, mais pour le bien commun ;
- le scepticisme organisé : tout énoncé nouveau doit être examiné, vérifié et reproduit avant d'être considéré comme valide.

34 « An old tradition and a new technology have converged to make possible an unprecedented public good. The old tradition is the willingness of scientists and scholars to publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge. The new technology is the internet. The public good they make possible is the worldwide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds. Removing access barriers to this literature will accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge. » (Budapest Open Access Initiative 2002, op. cit.)
Traduction en ligne sur https://www.budapestopenaccessinitiative.org/translations/french-translation.

Dans la philosophie du libre accès, l'Internet est un outil au service d'une science désintéressée, dont les résultats ont vocation à être ouverts à tous.

La réaffirmation du caractère commun de la science s'inscrit en réaction à une situation de monopole de l'édition scientifique. La littérature scientifique et, en particulier, les articles de revues sont devenus captifs de quelques grands éditeurs privés qui, ayant le monopole de ce qu'ils publient, en déterminent librement le prix d'accès. Or depuis la transition vers des offres d'édition numérique, la hausse constante des prix d'abonnement, conjuguée aux restrictions budgétaires des institutions, a entraîné l'émergence d'une crise éditoriale sur le marché des revues scientifiques (Chartron 2016). L'enjeu initial du libre accès est donc de *« lever [...] les barrières économiques »*[35], qui constituent un *« obstacle à l'accès »*[35], et de réintroduire, grâce à des modèles de publication alternatifs, une diffusion rapide et large de la littérature scientifique.

Le libre accès aux données de la recherche s'est alors imposé comme la suite logique du mouvement d'ouverture de la littérature scientifique. Refusant que les éditeurs s'emparent des données de la recherche et en fassent leur nouveau « *business* », les partisans du libre accès militent pour que soient mises en place des infrastructures publiques permettant de gérer et de partager les données.

1.2. Recommandations de l'OCDE (2004)

Parallèlement au mouvement du libre accès, ont émergé des recommandations de l'Organisation de Coopération et de Développement Économiques (OCDE) concernant l'accès aux données de la recherche.

1.2.1. Déclaration sur l'accès aux données de la recherche publique

En 2004, le Comité de la politique scientifique et technologique de l'OCDE publie une déclaration, dans laquelle les gouvernements signataires s'engagent à *« œuvrer à l'établissement de régimes d'accès aux données numériques de la recherche financée sur*

35 Budapest Open Access Initiative 2002, op. cit.

fonds publics »[36]. Ces régimes d'accès s'appuient en premier lieu sur le principe d'ouverture. Quoique soucieux de *« protéger [les] intérêts sociaux, scientifiques et commerciaux »*, les gouvernements signataires s'accordent sur le fait qu'*« un accès ouvert aux données [permet] d'accroître la qualité et l'efficacité de la recherche et de l'innovation »*.

Suite à cette déclaration, l'OCDE publie en 2007 des *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*[37]. Ceux-ci s'appuient toujours sur la notion centrale d'ouverture des données, définie comme *« l'accès dans des conditions d'égalité de la communauté scientifique internationale, à un coût le plus bas possible »*. C'est dans ce document également que l'OCDE définit pour la première fois ce qu'elle entend par « données de la recherche »[38]. A ce principe d'accessibilité des données sont associées des questions techniques (infrastructures d'accès), juridiques (respect des droits de propriété intellectuelle, protection de la vie privée, sécurité nationale) et économiques (les données scientifiques *« recueillies en vue de commercialiser les résultats de la recherche »* et celles *« qui appartiennent à une entité du secteur privé »* sont exclues des Principes et Lignes directrices).

La Déclaration de 2004 et les Principes et Lignes directrices de 2007 constituent une étape significative dans le développement des politiques publiques en matière d'accès aux données de la recherche. A la différence de la Déclaration de Berlin, qui a réuni des institutions de la recherche (ministères, agences de financement, établissements de recherche, bibliothèques, associations scientifiques…), les recommandations de l'OCDE ont quant à elles, pour la première fois, fédéré des gouvernements. Trente pays[39] ont ainsi adhéré à des *« normes*

36 ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2004). *Déclaration sur l'accès aux données de la recherche financée par des fonds publics*. https://legalinstruments.oecd.org/fr/instruments/157 (consulté le 18 septembre 2019).

37 ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2007). *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics.* Paris : Éditions OCDE. http://www.oecd.org/fr/science/sci-tech/38500823.pdf (consulté le 19 septembre 2019).

38 Voir première partie, 1.1, p.25

39 L'Allemagne, l'Australie, l'Autriche, la Belgique, le Canada, la Chine, la Corée, le Danemark, l'Espagne, les États-Unis, la Fédération de Russie, la Finlande, la France, la Grèce, la Hongrie, l'Irlande, l'Islande, Israël, l'Italie, le Japon, le Luxembourg, le Mexique, la Norvège, la Nouvelle-Zélande, les Pays-Bas, la Pologne, le Portugal, la République d'Afrique du sud, la République slovaque, la République tchèque, le Royaume-Uni, la Suède, la Suisse et la Turquie.

collectives » (certes non contraignantes), allant dans le sens d'une démocratisation de l'accès aux données de la recherche.

1.2.2. Une logique d'innovation

Si le mouvement du libre accès et les recommandations de l'OCDE prônent un principe identique – rendre les données accessibles –, en revanche leurs objectifs diffèrent. Le mouvement du libre accès défend des valeurs éthiques et revendique le droit à tous d'accéder à la connaissance. L'OCDE, quant à elle, inscrit l'ouverture des données scientifiques dans une logique marchande, centrée sur la croissance économique.

Dès les années 1990, l'OCDE reconnaît le rôle de plus en plus important du savoir dans l'économie. Elle publie en 1996 un rapport intitulé *L'économie fondée sur le savoir*, dans lequel elle considère la connaissance comme le *« moteur de la productivité et de la croissance économique »*[40].

1.2.2.1. Le concept d'économie de la connaissance

Le concept d'économie de la connaissance a été développé pour la première fois dans les années 1960. On pourra se reporter à l'article de Jean-Pierre Bouchez (2014), dont les deux premières parties exposent les différents courants liés à ce concept. Kenneth Arrow est considéré comme le premier à avoir développé une conception économique de la connaissance (Arrow 1962). Il distingue la connaissance des autres types de biens, la définissant comme non rivale[41] et non-exclusive[42]. L'économiste Fritz Machlup a, pour sa part, montré qu'entre 1947 et 1958 les industries américaines liées au savoir avaient connu une expansion significative en termes de richesses (Machlup 1962).

---

40 ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (1996). *L'économie fondée sur le savoir*. OCDE/GD(96)102. Paris : Editions OCDE. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=OCDE/GD(96)102&docLanguage=Fr (consulté le 18 septembre 2019). Page 3

41 Un bien non rival est un bien qui peut être consommé par plusieurs personnes simultanément. La plupart des biens non rivaux sont immatériels (une vidéo Youtube, par exemple).

42 Un bien non exclusif est un bien dont on ne peut restreindre la consommation en mettant des barrières techniques, financières ou légales.

L'OCDE véhicule une vision néolibérale du concept d'économie de la connaissance. Associant la croissance économique à la marchandisation du savoir, cette vision a notamment été développée par les économistes Dominique Foray et Bengt-Ake Lundvall (1997). Dans son rapport de 1996, l'OCDE définit l'économie de la connaissance comme une *« économie [reposant] directement sur la production, la diffusion et l'utilisation du savoir et de l'information »*[43].

1.2.2.2. Rôle de la science dans l'économie de la connaissance

Or la science joue un rôle important dans ce nouveau système de croissance économique. Par nature, l'enseignement supérieur et la recherche ont trois fonctions :

- Transmettre des savoirs (via la formation des étudiants) ;
- Produire de nouvelles connaissances ;
- Transférer ces connaissances (afin qu'elles puissent être utilisées par la société).

Ces trois fonctions sont citées par l'OCDE dans son rapport (1996, p.21). Au sein d'une économie fondée sur le savoir, la science présente en effet un double intérêt : d'une part, elle forme des ressources humaines qualifiées ; d'autre part, elle génère des connaissances susceptibles d'intéresser la société. C'est ce second point qui nous intéresse en particulier. L'enjeu pour les acteurs économiques est d'avoir accès aux connaissances scientifiques, afin de pouvoir les exploiter dans un but commercial. La science est donc ici considérée comme moteur d'innovation. Aussi les membres de l'OCDE s'engagent-ils à *« développer les liens entre le système scientifique et le secteur privé afin d'accélérer la diffusion du savoir »*[44]. Ils prônent le libre accès, parce que celui-ci permet d'accéder rapidement et gratuitement aux résultats de la recherche. Cela concerne aussi bien les publications que les données de la recherche. La recherche est désormais considérée comme partie intégrante d'un système, dans lequel la production, la diffusion et l'utilisation des connaissances procèdent d'une logique marchande.

43 Organisation de Coopération et de Développement économiques 1996

44 Organisation de Coopération et de Développement économiques 1996, op. cit., p.25

2. Influence de l'Open Data

L'Open Data est un mouvement prônant l'ouverture des données publiques, devenu aujourd'hui l'objet de politiques locales et nationales. Les données publiques peuvent être considérées comme *« les documents produits ou reçus, dans le cadre de leur mission de service public, par l'État, les collectivités territoriales ainsi que par les autres personnes de droit public ou les personnes de droit privé chargées d'une telle mission »*[45].

L'Open Data est mobilisé ici, car son histoire et ses perspectives s'entrecroisent avec celles des données de la recherche. En témoigne la législation française sur la réutilisation des informations du secteur public[46]. Les données de la recherche – en France, du moins – sont majoritairement produites par des établissements du secteur public. Or, à la différence des publications scientifiques, auxquelles s'applique le droit d'auteur, les données de recherche ne sont pas considérées comme des créations originales par la législation française. Elles ne sont donc pas protégeables par le droit d'auteur et ont le même statut que toute autre information publique (Maurel 2015). En ce sens, les données de la recherche font partie du mouvement et des politiques d'Open Data.

2.1. Philosophies du mouvement Open Data

On situe l'origine du discours sur l'Open Data aux États-Unis dans les années 2000. A cette époque, avec les progrès du web et l'informatisation croissante des services publics, l'administration s'est mise à disposer de bases de données de plus en plus nombreuses pour ses propres besoins de fonctionnement (éducation, santé, transports urbains…).

45 Article 1 de la loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal (dite « loi CADA »).

46 Notamment la loi n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public (aussi dite « loi Valter », https://www.legifrance.gouv.fr/eli/loi/2015/12/28/PRMX1515110L/jo/texte) et la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte).
Concernant ces lois, voir infra, 4.1, p.88

Le discours de l'Open Data est porté par deux courants de pensée, qui divergent quant à leur finalité :

- L'un, politique, attend de l'ouverture des données une plus grande transparence de l'action publique ;
- L'autre, économique, voit dans l'exploitation des données ouvertes un potentiel d'innovation.

2.1.1. Transparence et plus grande participation des citoyens

Cette philosophie est portée par des associations comme Civicus[47], Sunlight Foundation[48] ou Open Society Foundations[49]. Elle consiste à revendiquer un droit d'accès aux informations produites par les gouvernements. L'objectif est de :

- Renforcer le contrôle des citoyens sur l'action publique : une plus grande transparence de l'action des gouvernements contribuera à davantage d'intégrité ;
- Œuvrer à l'approfondissement démocratique : fournir un accès aux données publiques permet à chacun de se les approprier et de les utiliser pour faire valoir ses droits.

2.1.2. Innovation et création de valeur

Le second courant de pensée prônant l'ouverture des données publiques a une portée d'ordre social et économique. Elle est représentée par des associations comme Fing[50], Open Knowledge Foundation[51] ou LiberTIC[52]. L'ouverture des données est perçue comme catalyseur d'innovation économique et technologique. Rendre les données publiques accessibles c'est offrir la possibilité à des personnes tierces (particuliers, entreprises privées, associations…) de les réutiliser pour créer de nouveaux produits ou services.

47 https://www.civicus.org/

48 https://sunlightfoundation.com/

49 https://www.opensocietyfoundations.org/

50 http://fing.org

51 https://okfn.org/

52 https://libertic.wordpress.com/libertic/

Cette vision provient essentiellement du mouvement du logiciel libre (O*pen Source Movement*), apparu dans les années 1980 et qui fonde l'élaboration des biens et des connaissances sur les principes d'ouverture, de participation et de collaboration. *« Chaque programmateur qui contribue est invité à le faire publiquement, via des plateformes de partage des codes sources. Il peut bénéficier du travail des autres, mais doit en échange republier sa production, permettant ainsi la création d'une expertise collective »* (Chignard 2012, p.22). Une des innovations les plus connues de ce mouvement est l'encyclopédie participative Wikipédia[53].

Les acteurs des technologies de l'information et de la communication, en particulier les entrepreneurs du web, ont d'ailleurs joué un rôle prééminent dans la théorisation de l'Open Data. Tim O'Reilly, entrepreneur à l'origine du concept de web 2.0, en est un des activistes. Selon lui, c'est en s'appuyant sur les technologies ouvertes et collaboratives mises au point par les entreprises du web que la puissance publique pourra se réformer et devenir réellement démocratique (O'Reilly 2010). L'État ne dispose actuellement ni des moyens financiers ni des compétences nécessaires pour développer les services innovants de demain. Il a donc tout intérêt à libérer les données publiques et à externaliser la conception de ces nouveaux services numériques, en en tirant bénéfice par la suite via l'imposition de ces activités. Les promoteurs de l'Open Data tels Tim O'Reilly conçoivent donc le rôle de l'État comme générateur de croissance économique.

Cette vision s'est traduite par l'*Open Government Working Group Meeting* à Sebastopol (Californie) en 2007. Les 7 et 8 décembre 2007 se sont réunis trente acteurs du web et des technologies[54]. L'objectif de cette réunion était d'élaborer une définition du concept de « donnée publique ouverte » (*open government data*). Le workshop était coordonné par Tim

53 https://www.wikipedia.org/

54 Y ont participé : Carl Malamud (Public.Resource.Org), Tim O'Reilly (O'Reilly Media), Greg Elin (Sunlight Foundation), Micah Sifry (Sunlight Foundation), Adrian Holovaty (EveryBlock), Daniel X. O'Neil (EveryBlock), Michal Migurski (Stamen Design), Shawn Allen (Stamen Design), Josh Tauberer (GovTrack.us), Lawrence Lessig (Stanford), Dan Newman (MapLight.Org), John Geraci (outside.in), Edwin Bender (Inst. for Money), Tom Steinberg (My Society), David Moore (Participatory Politics), Donny Shaw (Participatory Politics), JL Needham (Google), Joel Hardi (Public.Resource.Org), Ethan Zuckerman (Berkman), Greg Palmer (NewCo), Jamie Taylor (MetaWeb), Bradley Horowitz (Yahoo), Zack Exley (New Organizing Institute), Karl Fogel (Question Copyright), Michael Dale (Metavid), Joseph Lorenzo Hall (UC Berkeley), Marcia Hofmann (EFF), David Orban (Metasocial Web), Will Fitzpatrick (Omidyar Network), Aaron Swartz (Open Library).

O'Reilly et Carl Malamud et sponsorisé par Google, Yahoo et la Sunlight Foundation. Huit principes ont été élaborés à cette occasion.

Les données publiques sont considérées « ouvertes » si elle sont :

> *« 1. Complètes : toutes les données publiques doivent être rendues disponibles, dans les limites légales liées à la vie privée et à la sécurité ;*
>
> *2. Primaires : les données sont telles que collectées à la source, à un degré de granularité le plus fin possible, sous une forme non agrégée et non modifiée ;*
>
> *3. Délivrées rapidement : les données doivent être rendues disponibles dans des délais brefs, de manière à préserver leur valeur ;*
>
> *4. Accessibles : les données doivent être mises à la disposition du plus grand nombre d'utilisateurs, dans des perspectives d'usage les plus vastes possibles ;*
>
> *5. Lisibles par des machines : les données devraient être structurées de façon à permettre leur traitement automatisé ;*
>
> *6. Non discriminatoires : les données sont librement accessibles à tous, sans qu'il soit nécessaire de s'enregistrer au préalable ;*
>
> *7. Non propriétaires : les données doivent être disponibles dans un format dont aucune organisation n'a le contrôle exclusif ;*
>
> *8. Dotées d'une licence libre : les données ne peuvent être soumises à la réglementation du droit d'auteur, du droit de brevet, des marques déposées ou du secret industriel ; des restrictions liées à la vie privée et à la sécurité peuvent néanmoins s'appliquer. »*[55]

55 Traduction de : « Government data shall be considered open if it is made public in a way that complies with the principles below:
1. Complete: All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. Primary: Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. Timely: Data is made available as quickly as necessary to preserve the value of the data.
4. Accessible: Data is available to the widest range of users for the widest range of purposes.
5. Machine processable: Data is reasonably structured to allow automated processing.
6. Non-discriminatory: Data is available to anyone, with no requirement of registration.
7. Non-proprietary: Data is available in a format over which no entity has exclusive control.
8. License-free: Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed. »
Open Government Data Principles (2007). https://public.resource.org/8_principles.html (consulté le 19

Ces huit principes s'en tiennent donc à une définition technique de l'ouverture des données publiques : comment rendre les données réutilisables (sous quel format) ? Car c'est bien la réutilisation des données qui intéresse les participants à ce workshop. Nous verrons que les principes FAIR, élaborés dans le cadre de l'ouverture des données scientifiques neuf ans plus tard, s'apparentent aux huit principes des données publiques ouvertes[56].

La présence de Google et Yahoo comme sponsors de ce workshop montre que les GAFAM ont eux aussi un intérêt à l'ouverture des données publiques. Fidelia Ibekwe-Sanjuan et Françoise Paquienséguy (2015) qualifient d'« ultra-libérale » la philosophie des géants du web, la différenciant de celle « libertaire » du mouvement Open Source. La philosophie ultra-libérale valorise une marchandisation généralisée. Les grandes masses de données occupent une place importante dans l'économie des géants du web. Elles alimentent des algorithmes qui seront utilisés par des sociétés commerciales, afin de calculer par exemple des grilles tarifaires d'assurance. L'utilisation de données ouvertes s'avère donc extrêmement rentable pour ces acteurs.

2.2. L'Open Data devenu objet politique

2.2.1. Les États-Unis : précurseurs des politiques d'Open Data

La première figure politique à s'être emparée de la question de l'Open Data est Barack Obama aux États-Unis. Le candidat à l'élection présidentielle de 2008 a fait de l'ouverture des données publiques une promesse de campagne. Le jour de sa prise de fonction, le 21 janvier 2009, il signe un mémorandum sur la transparence et le gouvernement ouvert[57], dans lequel il invite les agences fédérales et les ministères à appliquer les trois principes suivants :

septembre 2019).

56 Les principes FAIR seront présentés en 3.2.3.3, p.76.

57 OBAMA, B. (2009). *Transparency and Open Government. Memorandum for the Heads of Executive Departments and Agencies*. https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government (consulté le 19 septembre 2019).

- Transparence : *« Exploiter les nouvelles technologies pour mettre en ligne et rendre facilement accessibles au public des informations sur leurs activités et les décisions prises »*[58] ;
- Participation : *« Offrir [aux citoyens] des possibilités accrues de contribuer à l'élaboration des politiques »*[59] ;
- Collaboration : *« Utiliser des outils innovants [...] pour coopérer entre eux, à tous les niveaux de gouvernement, mais aussi avec des ONG, des entreprises et des particuliers du secteur privé »*[60].

Ce mémorandum pose la base d'une stratégie d'*open government* conjuguant à la fois la vision politique de transparence et la vision économique de création de valeur.

En mai 2009, le gouvernement américain lance le portail data.gov, premier portail donnant accès à des bases de données administratives, mises gratuitement à la disposition des citoyens et des entrepreneurs.

Le modèle américain sera ensuite repris dans de nombreux pays, se traduisant notamment par la création de portails *open data*. En France, le portail data.gouv.fr a été mis en ligne en décembre 2011 dans le cadre de la mission Etalab[61].

2.2.2. L'Open Data dans la politique : une démarche consensuelle mais des objectifs originels affaiblis

Des partis de tous bords ont mis en place des politiques d'Open Data. Cet essor interpelle par son caractère consensuel.

58 « Executive departments and agencies should harness new technologies to put information about their operations and decisions online and readily available to the public » (Obama 2009)
Traduction de Simon Chignard (2012, p.20-21)

59 « Executive departments and agencies should offer Americans increased opportunities to participate in policymaking » (Obama 2009)
Traduction de Simon Chignard (2012, p.20-21)

60 « Executive departments and agencies should use innovative tools, methods, and systems to cooperate among themselves, across all levels of Government, and with nonprofit organizations, businesses, and individuals in the private sector » (Obama 2009)
Traduction de Simon Chignard (2012, p.20-21)

61 https://www.etalab.gouv.fr/

> *« L'open data est-il simplement le support plastique d'une communication politique en mal de renouvellement, ou bien l'objet d'un formidable malentendu quant aux véritables effets politiques et sociaux qu'il est susceptible de produire ? »* (Peugeot 2012)

Le principe d'ouverture sert à la fois l'approfondissement démocratique et la stimulation économique – deux objectifs très différents, réunis au sein d'une même politique. En 2011, les 8 pays signataires de l'Open Government Partnership[62] se sont ainsi engagés à :

- *« rendre leurs gouvernements plus transparents, plus réactifs, plus responsables et plus efficaces »* ;
- *« [défendre] la valeur d'ouverture dans [leur] engagement auprès des citoyens pour améliorer les services, gérer les ressources publiques, promouvoir l'innovation et créer des communautés plus sûres »*[63].

Yu et Robinson (2012) dénoncent cette confusion des objectifs :

> *« Les gouvernements donnent accès à un panel de plus en plus large de données du secteur public, dans des formats numériques lisibles par des machines, rendant celles-ci plus faciles à réutiliser. Des informations améliorant la responsabilité des citoyens, comme la législation en instance au Congrès et la réglementation fédérale, sont désormais plus facilement accessibles. Mais des informations gouvernementales plus banales et pratiques, allant des horaires de bus aux données de l'inspection sanitaire des restaurants, sont également fournies dans des formats plus faciles d'utilisation. Ce type de données peut être utilisé pour améliorer la qualité de vie et le service public, mais a peu d'incidence sur la responsabilité politique. Des initiatives récentes, promouvant ou renforçant cette tendance ont été décrites comme des projets de « gouvernement ouvert ». »*[64]

62 La France a rejoint l'Open Government Partnership un peu plus tard, en 2014.

63 *Déclaration du gouvernement ouvert* (2011). https://www.opengovpartnership.org/fr/process/joining-ogp/open-government-declaration/ (consulté le 19 septembre 2019).

64 Traduction de : « Governments have made a growing range of public sector data available in machine-processable electronic formats that are easier for others to reuse. Information that enhances civic accountability, including pending congressional legislation and federal regulations, is indeed more readily

Selon ces deux auteurs, les politiques d'ouverture des données publiques ont rendu floue la frontière entre technologies et mouvement d'Open Data.

> *« Gouvernement ouvert et données ouvertes peuvent exister l'un sans l'autre. Un gouvernement peut être ouvert, au sens de transparent, même s'il n'a pas recours aux nouvelles technologies […]. A l'inverse, un gouvernement peut fournir des données ouvertes sur des sujets politiquement neutres, tout en restant profondément opaque et peu enclin à rendre des comptes. »*[65]

La réunion d'objectifs différents dans une même politique a affaibli le terme d'« *open data* » et, de fait, chaque objectif particulier.

2.3. Problématiques de réutilisation des données ouvertes

La question de la réutilisation n'a pas forcément été anticipée par les politiques d'Open Data, qui étaient parties du postulat que les données trouveraient naturellement leur public. En effet, malgré les 38 355 jeux de données disponibles sur le portail data.gouv.fr, la réutilisation n'est pas toujours au rendez-vous. Face à la faible réutilisation des données, la mission Etalab s'est vu assigner dès 2015 des objectifs d'accroissement du nombre d'utilisateurs actifs de la plateforme data.gouv.fr[66]. Il est donc revenu aux agents des politiques d'Open Data, dans les ministères et les collectivités territoriales, la mission de trouver des publics aux données ouvertes. Samuel Goëta étudie trois types de dispositifs mis en place pour stimuler cette réutilisation (Goëta 2018) :

available. But more mundane and practical government information, from bus schedules to restaurant health inspection data, is also being provided in friendlier formats. Such data can be used to improve quality of life and enhance public service delivery, but may have little impact on political accountability. Recent policy initiatives that promote or reinforce this trend have been described as "open government" projects. »

65 Traduction de : « Open government and open data can each exist without the other: A government can be an open government, in the sense of being transparent, even if it does not embrace new technology […]. And a government can provide open data on politically neutral topics even as it remains deeply opaque and unaccountable. »

66 « La mission Etalab assure la promotion de la réutilisation des données publiques par des acteurs de l'économie réelle. A cette fin, elle multiplie les démarches afin d'intéresser et de fédérer une communauté d'utilisateurs actifs qui partagent des données ou des projets sur le site data.gouv.fr. », dans SERVICES DU PREMIER MINISTRE (2015). *Objectifs et indicateurs de performance du programme Coordination du travail gouvernemental*. Annexe au projet de loi de règlement et RAP 2015 – Mission Direction de l'action du Gouvernement. https://www.performance-publique.budget.gouv.fr/sites/performance_publique/files/farandole/ressources/2015/rap/html/DRGPGMOBJINDPGM129.htm (consulté le 14 octobre 2019). Indicateur 7.2, « Ouverture et diffusion des données publiques »

- Les métadonnées. *« Les métadonnées sont des informations standardisées et structurées qui qualifient les données et sont censées apporter toutes les informations nécessaires à leur compréhension et leur appropriation. […] Sans elles, les ré-utilisateurs doivent comprendre par eux-mêmes les conditions de production des données et les catégories employées par les agents ».*
- La visualisation des données. *« Pour éviter que les données ne s'adressent uniquement à des publics disposant de connaissances techniques avancées, les portails open data proposent donc souvent des fonctionnalités de visualisation qui permettent aux usagers d'afficher les données sous la forme de tableaux, de graphiques ou de cartes ».*
- Les évènements de type concours, aussi appelés « hackathons ». Ils consistent à *« [inciter], de manière financière ou symbolique, les développeurs et les entrepreneurs à réutiliser les données sous la forme de services et d'applications ».* Les hackathons sont issus de la communauté du logiciel libre (Coleman 2013).

Open Research Data et Open Data étant étroitement liés, peut-on lire l'un à la lumière de l'autre ? Les données de la recherche seront-elles spontanément réutilisées ? Quels sont les écueils à éviter et qu'est-ce que cela suppose en termes d'infrastructures et de médiations à instaurer pour garantir la réutilisation de ces données ?

3. Politiques et initiatives de l'Union européenne

3.1. La vision d'une science au service de l'économie

3.1.1. Création d'un Espace Européen de la Recherche (2000)

La politique de la Commission européenne en matière d'accès à l'information scientifique s'inscrit principalement dans une logique socio-économique.

Au début des années 2000, la Commission prend en effet un tournant néolibéral, impactant le domaine de la recherche. Ce tournant se matérialise par la stratégie de Lisbonne, élaborée lors d'une réunion extraordinaire du Conseil européen les 23 et 24 mars 2000. Pour faire face à la compétitivité internationale (notamment celle des États-Unis et du Japon)[67], l'Union européenne s'est fixé un *« nouvel objectif stratégique »* : *« devenir l'économie de la connaissance la plus compétitive et la plus dynamique du monde, capable d'une croissance économique durable, accompagnée d'une amélioration quantitative et qualitative de l'emploi et d'une plus grande cohésion sociale »*[68].

Pour assurer la transition vers une économie de la connaissance, plusieurs lignes d'action ont été définies. Parmi elles, la création d'un *« espace européen de la recherche et de l'innovation »* (EER). Ce projet avait été formalisé en janvier 2000, dans une communication de la Commission européenne intitulée *Vers un espace européen de la recherche*[69]. La Commission considère que l'organisation actuelle de la recherche dans l'Union européenne (*« isolement et cloisonnement des systèmes nationaux de recherche »*, *« disparité des*

67 « Depuis 20 ans, le taux potentiel de croissance annuelle de l'économie européenne a baissé (d'environ 4 % à environ 2,5 %). Le chômage augmente de manière continue de cycle en cycle. Le taux d'investissement a diminué de 5 points. Notre position relative face aux États-Unis et au Japon s'est détériorée en ce qui concerne : l'emploi ; les parts de marché à l'extérieur ; la recherche-développement et l'innovation ainsi que leur traduction dans l'offre immédiate ; le développement des produits nouveaux. », dans COMMISSION EUROPÉENNE (1993). *Croissance, compétitivité, emploi : les défis et les pistes pour entrer dans le XXIe siècle. Livre blanc*. https://publications.europa.eu/fr/publication-detail/-/publication/0d563bc1-f17e-48ab-bb2a-9dd9a31d5004 (consulté le 19 septembre 2019). Page 9

68 CONSEIL EUROPÉEN (2000). *Conclusions de la présidence*. Conseil du 23 et 24 mars 2000 à Lisbonne. http://www.europarl.europa.eu/summits/lis1_fr.htm (consulté le 19 septembre 2019).

69 COMMISSION EUROPÉENNE (2000). *Communication de la Commission au Conseil, au Parlement européen, au Comité économique et social et au Comité des régions : Vers un espace européen de la recherche*. https://eur-lex.europa.eu/legal-content/fr/TXT/?uri=CELEX:52000DC0006 (consulté le 19 septembre 2019).

régimes réglementaires et administratifs »…[70]) pourrait conduire l'Europe à *« une perte de croissance et de compétitivité dans l'économie mondialisée »*[71]. Aussi propose-t-elle la création d'un espace européen de la recherche. L'objectif est de *« décloisonner »*, en coordonnant les politiques et les activités nationales de recherche au niveau de l'Union européenne.

Le projet de l'EER s'éloigne de la tradition scientifique européenne, fondée sur la quête de connaissances nouvelles, indépendamment de tous intérêts mercantiles. Pour la politiste Isabelle Bruno (2008), qui a entrepris de décortiquer les mécanismes de la stratégie de Lisbonne en matière de recherche, l'objectif de l'EER est de *« bâtir un « marché commun de la recherche »* qui concerne aussi bien les chercheurs, les laboratoires ou les universités que les droits de propriété intellectuelle, et plus largement tous les éléments constitutifs des systèmes nationaux dits d'innovation » (Bruno 2008, p.17). Ces éléments *« influent sur le potentiel présumé de compétitivité [d'un pays], c'est-à-dire sur sa capacité à gagner la faveur des entrepreneurs et investisseurs »* (Bruno 2008, p.17).

3.1.2. Renforcement de l'EER par le libre accès aux résultats scientifiques (2012)

Même si la communication sur le projet d'EER mentionne la nécessité de *« lever davantage encore les barrières […] qui freinent la circulation des connaissances et des personnes entre le monde académique et celui des entreprises »*[72], il n'est pas encore question, au début des années 2000, de libre accès aux résultats de la recherche.

C'est en 2012, lorsque la Commission planifie la réalisation finale de l'EER[73], que le principe d'accès libre aux publications et aux données scientifiques est explicitement adopté. Le projet

70 Commission européenne 2000, op. cit., p.7

71 Commission européenne 2000, op. cit., p.4

72 Commission européenne 2000, op. cit., p.9

73 COMMISSION EUROPÉENNE (2012). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au comité des régions. Un partenariat renforcé pour l'excellence et la croissance dans l'Espace européen de la recherche*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:52012DC0392 (consulté le 19 septembre 2019).

d'EER doit être achevé avant 2014[74]. L'objectif est de créer un *« marché unique pour la connaissance, la recherche et l'innovation »*, susceptible d'*« attirer les talents et les investissements »*[75]. Pour atteindre ce but, cinq grandes mesures ont été définies. L'une d'elles consiste à *« optimiser la diffusion, l'accessibilité et le transfert des connaissances scientifiques »*[76], notamment par des moyens numériques et un accès plus large et plus rapide aux publications et aux données. Rendre les résultats scientifiques accessibles constitue donc un *« défi majeur »*[77] pour un espace de recherche reposant sur le principe de libre circulation des chercheurs, des connaissances scientifiques et des technologies.

3.1.2.1. Première prise de position en faveur du libre accès aux résultats scientifiques (2007)

L'intérêt d'un accès large aux publications et aux données pour l'économie de la connaissance avait été évoqué pour la première fois en 2007, dans une communication de la Commission européenne sur l'information scientifique à l'ère numérique[78]. Dans cette communication, la Commission s'engageait à soutenir les *« initiatives [conduisant] à un accès plus large et à une meilleure diffusion de l'information scientifique [...], en particulier en ce qui concerne les publications et les données brutes générées par les activités de recherche soutenues par un financement public »*.

« Avancer sur ces enjeux aura un impact direct sur la capacité de l'Europe à utiliser les connaissances pour faire face à ses concurrents internationaux, un facteur déterminant pour atteindre les objectifs de compétitivité de l'agenda de Lisbonne ». Si la Commission mobilise certes le discours du libre accès pour justifier la nécessité de préserver et de diffuser l'information scientifique, elle met également en avant les deux bénéfices suivants :

74 A la demande du Conseil européen, dans ses conclusions de février 2011 et de mars 2012 (Commission européenne 2012b, op. cit., p.2).

75 Extrait des conclusions du Conseil européen de février 2011, cité dans : Commission européenne 2012b, op. cit., p.2

76 Commission européenne 2012b, op. cit., p.4

77 Commission européenne 2012b, op. cit., p.15

78 COMMISSION EUROPÉENNE (2007). *Communication de la Commission au Parlement européen, au Conseil et Comité économique et social européen sur l'information scientifique à l'ère numérique : accès, diffusion et préservation*. https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52007DC0056&from=EN (consulté le 19 septembre 2019).

- Le retour sur investissement des sommes engagées par les instances publiques dans le financement de la recherche : *« les enjeux pour la Communauté Européenne sont particulièrement conséquents : pour la période 2007-2013, la Communauté a décidé d'investir quelques 50 milliards d'euros dans le 7ème PCRD [Programme Cadre de Recherche et de Développement] »*[79] ;
- La contribution à l'innovation : *« pour être une économie de la connaissance véritablement compétitive, l'Europe doit renforcer la production des connaissances par la recherche, leur diffusion par l'éducation et leur application grâce à l'innovation »*[80].

3.1.2.2. Les raisons d'une politique de libre accès aux résultats de la recherche

Il a fallu attendre 2012, avant que ne se multiplient les mesures de la Commission en matière de préservation et d'accès aux données scientifiques. Plusieurs décisions ont eu un effet catalyseur.

La planification d'objectifs pour 2020

En 2010, l'Union européenne se fixe une série d'objectifs pour la décennie à venir. Ces derniers sont formalisés dans la communication *Europe 2020 : une stratégie pour une croissance intelligente, durable et inclusive*[81]. A travers la notion de croissance intelligente, la Commission fait du développement d'une économie fondée sur la connaissance une priorité. Elle s'appuie notamment sur le contexte de crise économique pour justifier un redoublement des efforts en matière de recherche et d'innovation. Les objectifs fixés par la stratégie *Europe 2020* sont présentés plus en détail, notamment, dans les communications *Une Union de*

79 Commission européenne 2007, op. cit.

80 Commission européenne 2007, op. cit.

81 COMMISSION EUROPÉENNE (2010a). *Communication de la Commission. Europe 2020 : Une stratégie pour une croissance intelligente, durable et inclusive*. https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A52010DC2020 (consulté le 19 septembre 2019).

l'innovation[82] et *Une stratégie numérique pour l'Europe*[83]. Dans la première, la Commission s'engage à réaliser l'Espace européen de la recherche, qui *« mettra en place les structures nécessaires pour une véritable libre circulation de la connaissance »*[84]. Elle s'engage également à soutenir l' *« innovation ouverte et coopérative »* et à favoriser le libre accès aux résultats de la recherche publique[85]. La seconde communication, *Une stratégie numérique pour l'Europe*, vise à stimuler l'innovation dans le domaine des technologies de l'information et de la communication (TIC). Pour ce faire, la Commission définit une politique de diffusion large de la recherche publique, par la publication en libre accès des données et des articles scientifiques.

Une politique d'ouverture des données publiques

La réflexion sur le libre accès aux données de la recherche a également été catalysée par le déploiement d'une politique d'ouverture des données publiques.

En 2003, l'Union européenne a adopté la Directive 2003/98/CE concernant la réutilisation des informations du secteur public (dite « Directive PSI », pour « *Public Sector Information* »)[86]. Celle-ci fixe les règles et les conditions de réutilisation des données des administrations nationales et européennes. En 2011, la Commission européenne annonce des mesures renforcées pour ouvrir les données publiques[87]. Ces mesures se traduiront en 2013 par la révision de la Directive PSI[88], qui instaurera un véritable droit à la réutilisation des données du secteur public.

82 COMMISSION EUROPÉENNE (2010c). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions. Une Union de l'innovation*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2010:0546:FIN (consulté le 19 septembre 2019).

83 COMMISSION EUROPÉENNE (2010b). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions. Une stratégie numérique pour l'Europe*. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52010DC0245R(01) (consulté le 19 septembre 2019).

84 Commission européenne 2010c, op. cit., p.3

85 Commission européenne 2010c, op. cit, p.21

86 *Directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public. 345*. (32003L0098). http://data.europa.eu/eli/dir/2003/98/oj/eng (consulté le 9 septembre 2019).

87 COMMISSION EUROPÉENNE (2011). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions. L'ouverture des données publiques: un moteur pour l'innovation, la croissance et une gouvernance transparente.* https://eur-lex.europa.eu/legal-content/fr/TXT/?uri=CELEX:52011DC0882 (consulté le 19 septembre 2019).

Depuis 2019, la Directive (EU) 2019/1024 (dite « Directive Open Data »)[89] remplace la Directive PSI. Elle inclue notamment un droit de réutilisation gratuite (à des fins commerciales ou non commerciales) des données de la recherche déjà rendues publiques par des scientifiques ou des établissement de recherche. Cette directive devra être transposée dans les droits nationaux d'ici 2021.

Le développement d'infrastructures de recherche

La question des infrastructures de recherche est également au cœur de la réflexion sur l'Espace européen de la recherche. En 2010, la Commission européenne constitue un groupe d'experts, chargé d'identifier les principaux défis d'une infrastructure de recherche, en termes d'accès, de traitement et de préservation des données scientifiques. A l'issue de sa mission, le groupe délivre un rapport intitulé *Riding the wave : How Europe can gain from the rising tide of scientific data*[90]. Ce livrable confortera la Commission européenne dans la nécessité d'instaurer des mesures pour améliorer l'accès aux données – et non plus seulement aux publications.

3.1.2.3. La mise en place de moyens d'action (2012)

Jusqu'en 2012, la Commission a en effet concentré ses efforts sur le libre accès aux ouvrages et articles scientifiques. En sa qualité d'organisme de financement de la recherche[91], elle a notamment mis en place en 2008 un projet pilote couvrant certaines thématiques de son 7ème programme-cadre (FP7). Les bénéficiaires de subventions étaient invités à mettre leurs

88 *Directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013 modifiant la directive 2003/98/CE concernant la réutilisation des informations du secteur public. 175*. http://data.europa.eu/eli/dir/2013/37/oj (consulté le 19 septembre 2019).

89 *Directive (EU) 2019/1024 du Parlement européen et du Conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public. 172*. http://data.europa.eu/eli/dir/2019/1024/oj (consulté le 19 septembre 2019).

90 HIGH LEVEL-EXPERT GROUP ON SCIENTIFIC DATA (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data*. Rapport final. https://ec.europa.eu/digital-single-market/en/news/digital-agenda-unlock-full-value-scientific-data-high-level-group-presents-report (consulté le 19 septembre 2019).

91 Depuis 1983, la Commission européenne offre aux chercheurs des pays membres la possibilité de répondre à des appels à projets dans le cadre de programmes de financement quadriennaux puis septennaux, dénommés Programmes-cadres de recherche et de développement (PCRD).

publications en libre accès, soit en les déposant dans des archives ouvertes, soit en les éditant dans des revues en libre accès.

Le 17 juillet 2012, la Commission délivre une communication[92], dans laquelle elle réitère son engagement en faveur du libre accès aux informations scientifiques. Par rapport à la communication de 2007[93], les données ont gagné en importance. Une place équivalente à celle des publications leur est consacrée.

Cet engagement en faveur du libre accès aux informations scientifiques est mû par la conviction que leur réutilisation par les chercheurs, les entreprises et les citoyens peut *« contribuer de manière significative à la croissance économique de l'Europe »*[94]. Sont notamment citées l'*« accélération des découvertes scientifiques »*, le *« développement de nouvelles formes de recherches à forte intensité de données »* et l'*« adoption systématique des résultats de la recherche par les entreprises et l'industrie européennes »*[95]. L'accent est mis sur l'origine et la vocation publique des informations scientifiques : en tant qu'activité réalisée sur la base de moyens publics (c'est-à-dire financée par l'État, donc par la société), la recherche doit faire en sorte que ses résultats soient accessibles et réutilisables par tous gratuitement.

Dans cette communication, la Commission détaille les mesures qu'elle compte prendre pour améliorer l'accès aux informations scientifiques. Elle annonce qu'elle *« donner[a] l'exemple »* en demandant aux équipes scientifiques, financées par son programme de subvention, de rendre les résultats de leurs recherches librement accessibles[96]. Elle annonce également qu'elle financera la création d'infrastructures de données[97].

---

92 COMMISSION EUROPÉENNE (2012a). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au comité des régions. Pour un meilleur accès aux informations scientifiques: dynamiser les avantages des investissements publics dans le domaine de la recherche*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:52012DC0401 (consulté le 19 septembre 2019).

93 Commission européenne 2007, op. cit.

94 Commission européenne 2012a, p.13

95 Commission européenne 2012a, op. cit., p.3

96 Commission européenne 2012a, op. cit., p.11

97 Commission européenne 2012a, op. cit., p.12

3.1.2.4. Des recommandations aux États membres (2012)

La communication de 2012 s'accompagne d'une recommandation aux États membres[98], qui préconise une *« amélioration des politiques et pratiques relatives à l'accès aux informations scientifiques et à leur conservation »* à l'échelle des pays.

La Commission recommande notamment aux États membres :

- de *« définir des politiques claires en matière de diffusion des données de la recherche financée par des fonds publics »* et de *« veiller à ce que les données de la recherche financée par des fonds publics deviennent accessibles, utilisables et réutilisables par le public au moyen d'infrastructures électroniques »*[99] ;
- de *« renforcer la conservation des informations scientifiques »* ;
- de *« développer davantage les infrastructures électroniques sous-tendant le système de diffusion des informations scientifiques » et de « créer des synergies, aux niveaux européen et mondial, entre les infrastructures électroniques nationales »*.

La Commission cherche à créer une émulation parmi les États membres, en les exhortant à participer à l' *« entreprise mondiale »* du libre accès et à y *« faire figure d'exemple »* par la *« mise en place d'un environnement de recherche ouvert et collaboratif »*.

C'est donc entre 2007 et 2012 que se structure, au sein de l'Union européenne, une politique en matière de libre accès aux données de la recherche. Celle-ci s'insère dans un objectif plus large de croissance économique, fondée sur la connaissance et l'innovation. Comme le souligne Hans Dillaerts dans un article sur la libre circulation de l'information scientifique et technique, *« l'engagement croissant des politiques européennes en faveur de l'innovation ouverte et la science ouverte au cours de ces dix dernières années trouve directement son*

98 COMMISSION EUROPÉENNE (2012c). *Recommandation de la Commission relative à l'accès aux informations scientifiques et à leur conservation*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=uriserv:OJ.L_.2012.194.01.0039.01.FRA&toc=OJ:L:2012:194:TOC (consulté le 19 septembre 2019).

99 La Commission introduit néanmoins la limite suivante : « Il y a lieu de tenir dûment compte des questions relatives, notamment, au respect de la vie privée, aux secrets industriels, à la sûreté nationale, aux intérêts commerciaux légitimes et aux droits de propriété intellectuelle. Les données, le savoir-faire et/ou les informations, quelle que soit leur forme ou leur nature, qui sont détenus par des acteurs privés participant à un partenariat public-privé avant les activités de recherche et qui ont été identifiés comme tels ne sont pas soumis à ce type d'obligation ».

origine dans les théories et modèles économiques de l'économie de la connaissance » (Dillaerts, 2017, p.42). Si la Commission européenne utilise le registre incitatif, en émettant des recommandations aux États membres, elle a également recours à un registre plus prescriptif, se servant des programmes de recherche qu'elle finance comme levier d'action.

3.2. Des programmes de financement de la recherche de plus en plus prescriptifs

3.2.1. La Commission européenne, financeur de la recherche

La Commission européenne fait partie des financeurs de la recherche en Europe. Depuis 1983, elle développe des Programmes-Cadre de Recherche et de Développement (PCRD). Il s'agit d'outils de financement, permettant de subventionner des projets de recherche. Les subventions sont accordées aux équipes scientifiques sur la base d'appels à propositions suivis d'une procédure d'examen par les pairs. En tant qu'organisme de financement, la Commission européenne est donc en droit de déterminer les règles de diffusion des résultats scientifiques issus des subventions qu'elle accorde.

3.2.2. L'initiative première du Conseil européen de la recherche

En 2007, le Conseil européen de la recherche (CER)[100] demande la mise en libre accès des résultats issus des recherches qu'il finance. Cette demande a valeur d'obligation pour les publications – lesquelles doivent être déposées dans des archives ouvertes dans un délai de six mois. Pour les données, le CER ne formule pas de mandat de dépôt. Il juge seulement *« essentiel que les données brutes* (primary data) *[…] soient déposées dans les bases de données adéquates le plus rapidement possible, de préférence dès publication et au plus tard dans les six mois »*[101].

100 Le Conseil européen de la recherche est une agence de financement, instituée en 2007 par la Commission européenne. Il est chargé de financer la recherche exploratoire. Les subventions qu'il attribue sont issues d'une partie du budget du programme-cadre en cours.

101 EUROPEAN RESEARCH COUNCIL (2007). *European Research Council-Scientific Council Guidelines for Open Access*. https://erc.europa.eu/sites/default/files/document/file/erc_scc_guidelines_open_access.pdf (consulté le 19 septembre 2019). Traduction de Rémi Gaillard (2014, op. cit., p.22)

3.2.3. La politique d'ouverture des données de la recherche dans le programme Horizon 2020

Cette initiative a été reprise par la Commission européenne dans ses programmes-cadres dès 2008. Sur le même schéma que le CER, elle a mis en place un plan d'action progressif, commençant par réglementer la diffusion des publications, avant de se pencher sur la question des données de la recherche.

C'est à partir de 2013, avec le lancement du 8ème programme-cadre « Horizon 2020 », qu'elle instaure une politique de libre accès aux données issues des projets subventionnés.

3.2.3.1. Le dépôt des données en ligne

Dans la convention que signent les chercheurs bénéficiaires d'une subvention, l'article 29.3 prévoit une clause de libre accès aux données :

> *« En ce qui concerne les données numériques de la recherche produites dans le cours de l'action « données », les bénéficiaires doivent :*
>
> *(a) les déposer dans une banque de données de la recherche et prendre des mesures afin de permettre aux tiers d'accéder aux éléments suivants et de les explorer, exploiter, reproduire et diffuser gratuitement pour l'utilisateur :*
>
> *(i) les données, y compris les métadonnées, nécessaires pour valider dès que possible les résultats présentés dans des publications scientifiques;*
>
> *(ii) [OPTION A pour les actions sanitaires participant au projet pilote sur le libre accès aux données de la recherche, si cela est prévu dans le programme de travail : les données utiles pour répondre à une urgence de santé publique, si cela est demandé expressément par la Commission et dans le délai précisé dans la demande] [OPTION B : sans objet] ;*
>
> *(ii) d'autres données, y compris les métadonnées associées, spécifiées dans le « plan de gestion de données » et dans les délais qui y sont fixés ;*

> *(b) fournir des informations, par la banque de données, sur les outils et les instruments à la disposition des bénéficiaires et nécessaires pour la validation des résultats (et, si possible, fournir les outils et instruments eux-mêmes). »*[102]

Cette clause concerne tout type de données, qu'elles soient ou non sous-jacentes à une publication. Les porteurs de projets doivent déposer les données dans des entrepôts en ligne *« disciplinaires ou thématiques »*, *« institutionnel[s] ou centralisé[s] »*[103]. La Commission renvoie par défaut vers l'entrepôt indifférencié Zenodo[104]. Elle recommande par ailleurs de joindre aux données une licence d'utilisation libre de type Creative Commons (CC-BY ou CC0)[105]. Il est proposé aux porteurs de projet une prise en charge des coûts liés à la mise en libre accès.

La Commission a d'abord introduit cette clause sous forme d'action pilote (intitulée *Open Research Data Pilot*). Seules sept thématiques du programme H2020 étaient concernées, soit 20% du budget total alloué pour 2014 et 2015. A partir de 2017, l'action pilote a été étendue à tous les nouveaux projets financés par le programme-cadre, quelque soit leur thématique.

La demande de mise en libre accès des données est en réalité peu contraignante. En effet, les porteurs de projet ont la possibilité de se désengager de cette obligation à tout moment (*opt out*). La Commission reconnaît que certaines données ne se prêtent pas à une diffusion en libre accès notamment lorsqu'elles font intervenir des aspects de confidentialité (obligation de sûreté, protection des données personnelles) ou lorsqu'elles sont liées à un projet de recherche appliquée (exploitation industrielle ou commerciale envisagée). La Commission parle d'un principe d'accès *« aussi ouvert que possible, aussi fermé que nécessaire »*[106]. La mise à disposition des données n'est d'ailleurs pas soumise à évaluation, ni au moment de la sélection des propositions, ni au terme du financement du projet. En d'autres termes, les

---

102 COMMISSION EUROPÉENNE (2017c). *H2020 Programme: Mono-Beneficiary General Model Grant Agreement*. Version 5.0. http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-mono_en.pdf (consulté le 19 septembre 2019). Page 68

103 COMMISSION EUROPÉENNE (2017b). *H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Version 3.2. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (consulté le 19 septembre 2019). Page 6

104 https://zenodo.org/

105 https://creativecommons.org/

106 Commission européenne 2017b, op. cit., p.8

chercheurs bénéficiaires n'ont pas plus intérêt à respecter la clause de l'article 29.3 qu'à en sortir.

3.2.3.2. La rédaction de plans de gestion de données

Les porteurs de projets s'étant engagés à mettre les données à disposition doivent également remettre un plan de gestion de données (PGD)[107]. Il s'agit d'un document décrivant les données générées ou collectées au cours du projet de recherche et la manière dont celles-ci seront traitées. Un tel document comporte notamment des informations sur :

- La nature des données ;
- Les moyens permettant d'assurer leur stockage et leur sécurisation ;
- Leur possible archivage ;
- Leur mode de diffusion ;
- D'éventuels aspects éthiques ;
- Les coûts engendrés par l'ensemble de ces traitements.

Les plans de gestion de données sont considérés comme des documents évolutifs, amenés à être révisés et enrichis au fur et à mesure de l'avancement du projet de recherche.

Aussi la Commission demande-t-elle aux porteurs de projets de soumettre une première version de leur plan de gestion dans les six premiers mois suivants le début du projet. Une version révisée est ensuite fortement recommandée pour chaque évaluation périodique du projet.

3.2.3.3. Des données compatibles avec les principes FAIR

La Commission incite, par ailleurs, les porteurs de projets à se conformer aux principes FAIR. Élaborés par un groupe de travail entre 2014 et 2016 (Wilkinson et al. 2016), les principes FAIR sont un ensemble de lignes directrices visant à rendre les données trouvables (*findable*), accessibles (*accessible*), interopérables (*interoperable*) et réutilisables (*reusable*), aussi bien par des humains que par des machines.

107 En anglais, « *data management plan* » (DMP).

En vue de l'application de ces recommandations, la Commission a produit un guide intitulé *Lignes directrices pour la gestion des données FAIR dans Horizon 2020*[108], dans lequel est proposé un modèle de plan de gestion de données. Son usage est facultatif ; les porteurs de projets restent libres de produire le plan qu'ils souhaitent.

3.2.4. Vers une politique renforcée dans le prochain programme Horizon Europe ?

Les dispositions prises par la Commission européenne pour mettre en libre accès les données issues des recherches qu'elle finance ont donc été introduites de manière progressive, ciblant d'abord des projets pilotes avant d'être généralisées à l'ensemble des projets financés. L'application de ces mesures se voulait relativement flexible. Le taux de participation en 2017 a été estimé à 62% des projets financés[109]. Jugeant ces progrès limités[110], la Commission souhaite aboutir à davantage d'ouverture dans son prochain programme-cadre « Horizon Europe » (2021-2027). Elle a annoncé dans un communiqué de presse en date du 7 juin 2018[111] que le principe de science ouverte deviendrait le « mode opératoire » de ce nouveau programme.

La Commission entend aller plus loin que Horizon 2020, en durcissant les exigences de mise en libre accès. La rédaction de plans de gestion de données, notamment, deviendra obligatoire pour l'ensemble des projets générant des données. De même, il ne sera plus aussi facile pour les bénéficiaires de subventions de se désengager de l'action d'ouverture des données ; les demandes devront être dûment motivées. La Commission entend par ailleurs développer les compétences des chercheurs en matière de science ouverte et de gestion de données[112]. En

108 COMMISSION EUROPÉENNE (2016b). *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020*. Version 3.0. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (consulté le 19 septembre 2019).

109 COMMISSION EUROPÉENNE (2018b). *Commission Staff Working Document. Impact Assessment of Horizon Europe*. SWD(2018) 307 final, partie 2/3. https://ec.europa.eu/info/publications/horizon-europe-impact-assessment-staff-working-document_en (consulté le 19 septembre 2019). Page 104

110 Ibid.

111 COMMISSION EUROPÉENNE (2018a). *Budget de l'Union: La Commission propose le programme de recherche et d'innovation le plus ambitieux à ce jour*. Communiqué de presse. http://europa.eu/rapid/press-release_IP-18-4041_fr.htm (consulté le 19 septembre 2019).

112 Commission européenne 2018b, op. cit., p.106

revanche, elle ne prévoit pas de prendre en compte ce dispositif dans l'évaluation des projets. Cette démarche pose question quant aux résultats attendus. Car, à l'exception de l'obligation de rédiger un plan de gestion de données et de justifier une demande d'*opt out*, le dispositif prévu pour Horizon Europe reste le même que celui du précédent programme. Le taux d'*opt out* est-il réellement résorbable, au vu des contraintes de confidentialité qui peuvent peser sur les données de recherche ?

3.3. Des infrastructures de recherche mises au service de l'Open Science

Si la Commission européenne se sert des programmes de financement pour tenter de transmettre à la communauté scientifique des pratiques de libre accès, elle y a également recours pour créer des infrastructures de recherche destinées à préserver et diffuser les données.

3.3.1. Les infrastructures de recherche financées par la Commission européenne

3.3.1.1. La question des données dans les infrastructures de recherche

<u>Qu'est-ce qu'une infrastructure de recherche ?</u>

Par « infrastructures de recherche », la Commission européenne entend *« des équipements fournissant des ressources et des services aux communautés scientifiques pour mener des recherches et favoriser l'innovation »*[113]. Une infrastructure peut être à site unique, distribuée ou virtuelle. Ce sont par exemple de *« grands équipements scientifiques »*, des *« collections, archives ou données de recherche »*, des *« systèmes de calcul et réseaux de communication »*[114].

113 Traduction de : « Research Infrastructures are facilities that provide resources and services for research communities to conduct research and foster innovation » (source : https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures_en).

114 Traduction de : « They include : major scientific equipment or sets of instruments ; collections, archives or scientific data ; computing systems and communication networks ; any other research and innovation

Différents types d'infrastructures

Toutes les infrastructures que finance la Commission européenne ne sont pas dédiées à la préservation et à la diffusion des données de recherche. Des infrastructures comme Euro-BioImaging[115] ou EU-SOLARIS[116] ont plutôt pour fonction de collecter ou de générer des données. Pour la plupart, ce sont des infrastructures disciplinaires très spécialisées. Initialement, elles n'ont pas vocation à rendre accessibles les données qu'elles produisent.

Avec le développement de la science ouverte et de la science des données dans le discours de la Commission européenne, ces infrastructures sont aujourd'hui confrontées à des exigences croissante de structuration et de mise à disposition des données. Dans le cadre de leur financement par l'Europe, elles sont notamment évaluées sur la qualité des dispositifs numériques qu'elles proposent (banques de données, serveurs sécurisés...)[117].

Mais la majorité des infrastructures que finance la Commission européenne sont des infrastructures dites « de données » (*data infrastructures*). Il s'agit d'infrastructures dont la vocation est de centraliser des jeux de données et de les rendre accessibles. Dans certains cas, la fourniture de données n'est qu'une composante de l'infrastructure. On peut citer :

- DARIAH[118], qui tend à faciliter l'accessibilité et l'utilisation à long terme des données de recherche produites en Europe dans le domaine des sciences humaines et sociales.

---

infrastructure of a unique nature which is open to external users » (source : https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures_en).

115 Euro-BioImaging (*European Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences*) est une infrastructure distribuée proposant des équipements d'imagerie spécialisés dans le domaine de la biologie. Elle est composée de différents « nœuds » physiques, répartis dans 11 pays d'Europe, dans lesquels les chercheurs peuvent se rendre pour acquérir des données. Site web de l'infrastructure : http://www.eurobioimaging.eu/

116 EU-SOLARIS est une infrastructure distribuée, qui entrera en fonctionnement en 2020. Elle a pour but l'élaboration de dispositifs expérimentaux à énergie solaire concentrée et à énergie solaire thermique. Site web de l'infrastructure : http://www.eusolaris.eu/

117 « ESFRI Research Infrastructures are evaluated, selected, monitored and reviewed with much emphasis on their e-Infrastructure component that is considered a basis for excellent science and excellent data services to the broadest community », dans EUROPEAN STRATEGY FORUM ON RESEARCH INFRASTRUCTURES (2018). *Roadmap 2018. Strategy Report on Research Infrastructures*. http://roadmap2018.esfri.eu/ (consulté le 10 septembre 2019). Page 19
ESFRI est le comité chargé de la sélection des infrastructures de recherche européennes (voir infra, 3.3.1.2, p.80).

118 Digital Research Infrastructure for the Arts and Humanities (https://www.dariah.eu/)

- ELIXIR[119], dont le but est de coordonner les initiatives de partage de données en biologie et en médecine.

La Commission européenne finance également des infrastructures multidisciplinaires, souvent appelées « e-infrastructures ». Ces infrastructures sont le fruit du Big Data appliqué à la science. Parmi elles :

- EUDAT[120], une infrastructure dédiée à la préservation des données scientifiques ;
- PRACE[121] pour le calcul intensif ;
- OpenAIRE[122] et Zenodo pour la diffusion de publications, données et autres produits de la recherche en libre accès.

3.3.1.2. La politique européenne de financement des infrastructures de recherche

Depuis 2003, une partie des programmes-cadres (PCRD) est consacrée au financement d'infrastructures de recherche. Cette part représente environ 3 % du budget de chacun des programmes-cadres qui se sont succédés entre 2002 et 2020[123].

La politique européenne de financement des infrastructures s'est structurée de manière progressive. Pour la première fois en 2006, la Commission a chargé le European Strategy Forum on Research Infrastructures (ESFRI), un comité composé de délégués aux ministères de la recherche des différents pays membres, de rédiger une feuille de route des infrastructures de recherche européennes. L'objectif de cette feuille de route est de contribuer à l'élaboration d'une politique cohérente. ESFRI sélectionne des propositions d'infrastructures, qu'il soutient pendant une période d'implémentation de dix ans maximum. Ces infrastructures ont alors le statut de ESFRI Projects. Lorsqu'elles parviennent à maturité,

119 https://elixir-europe.org/

120 European Data (https://eudat.eu/)

121 Partnership for Advanced Computing in Europe (http://www.prace-ri.eu/)

122 https://www.openaire.eu/

123 Dans le 6ème PCRD (2002-2006), elle correspondait à une enveloppe de 715 millions d'euros ; dans le 7ème PCRD (2007-2013), à une enveloppe de 1 715 millions d'euros ; et dans le 8ème PCRD Horizon 2020 (2014-2020), à une enveloppe de 2 488 milliards d'euros – les budgets des trois PCRD étant respectivement de 19, 56 et 77 milliards d'euros.

elles sont classées comme ESFRI Landmarks. La feuille de route de 2018 (dernière version en date) répertorie 18 ESFRI Projects et 37 ESFRI Landmarks[124].

Dans sa feuille de route de 2018, ESFRI identifie trois enjeux à court terme :

- La conformité des infrastructures de recherche avec les principes FAIR ;
- La pérennité des infrastructures ;
- L'engagement des infrastructures dans l'innovation, avec la fourniture de services à l'industrie et la collaboration avec le secteur privé pour des recherches pré-compétitives.

En termes de données, ESFRI se donne pour mission de veiller à l'adoption de principes d'accessibilité et d'interopérabilité dans les infrastructures qu'il soutient. Ce faisant, il tente de positionner les infrastructures de recherche sur un autre projet très en vogue actuellement : le cloud européen pour la science ouverte.

3.3.2. Le cloud européen pour la science ouverte (EOSC[125])

3.3.2.1. Cadre d'émergence de l'EOSC: le passage de l'économie mondiale au numérique

L'idée d'un cloud pour la recherche a été mentionnée pour la première fois en 2015 dans une communication de la Commission européenne pour un « marché unique numérique en Europe »[126]. L'établissement d'un marché numérique unique a pour origine l'utilisation croissante du numérique comme support des activités économiques. L'objectif est de créer un *« espace dans lequel la libre circulation des biens, des personnes, des services et des capitaux est garantie »*. La stratégie de la Commission repose sur trois piliers, dont un vise à *« maximiser le potentiel de croissance de [l']économie numérique européenne »*. Pour ce faire, la Commission entend investir dans des infrastructures et technologies d'information et

124 European Strategy Forum on Research Infrastructures 2018, op. cit.

125 Acronyme anglais de « *European Open Science Cloud* ».

126 COMMISSION EUROPÉENNE (2015). *Communication de la Commission au Conseil, au Parlement européen, au Comité économique et social et au Comité des régions : Stratégie pour un marché unique numérique en Europe*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52015DC0192 (consulté le 19 septembre 2019).

de communication, comme l'informatique en nuage, les mégadonnées et le calcul haute performance. Parmi les initiatives annoncées, la création d'un *« nuage pour la recherche consacré à la science ouverte »*[127]. La Commission pose ainsi clairement le contexte de création de ce cloud pour la recherche, à savoir : la construction d'infrastructures compétitives, visant à instaurer un *« climat d'investissement favorable pour les réseaux numériques, la recherche et les entreprises innovantes »*[128].

3.3.2.2. Structure du cloud européen pour la science ouverte

Le projet de cloud européen a été proposé en avril 2016 par la Commission dans une communication intitulée *European Cloud Initiative – Building a competitive data and knowledge economy in Europe*[129].

Le cloud est une technologie tripartite, composée (1) d'une infrastructure pour stocker et gérer les données, (2) de réseaux haut-débit pour transférer ces données et (3) de serveurs de calcul haute performance pour les analyser.

Le projet européen a été intitulé EOSC (*European Open Science Cloud*) et devrait voir le jour d'ici 2020. Il s'apparente aux technologies de cloud classiques, dans la mesure où il associera des dispositifs de calcul et de stockage à des infrastructures de données préexistantes (nationales, régionales et institutionnelles) issues de la recherche. Il donnera donc accès à un ensemble de ressources distribuées (données, services, logiciels…) à partir d'un guichet unique : l'EOSC Portal[130]. Destiné aux chercheurs européens, il offrira à terme également ses services aux secteurs public et privé.

Le projet est financé par le programme-cadre Horizon 2020 à hauteur de 600 millions d'euros. Il a d'abord fait l'objet d'un processus de consultation et de réflexion de deux ans, avant

---

127 Commission européenne 2015, op. cit, p.17

128 Commission européenne 2015, op. cit., p.21

129 COMMISSION EUROPÉENNE (2016a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: European Cloud Initiative - Building a competitive data and knowledge economy in Europe*. https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe (consulté le 19 septembre 2019).

130 https://www.eosc-portal.eu/

d'être présenté aux États membres le 23 novembre 2018[131]. A cette occasion a été adoptée une déclaration[132], énonçant les principes directeurs pour l'implémentation de l'EOSC (les principaux points concernaient la culture des données et les données FAIR, l'architecture et les services de données, la gouvernance et le financement de l'EOSC). Soixante-dix acteurs de la sphère scientifique (des infrastructures de recherche, des éditeurs scientifiques…) se sont ainsi engagés à s'y conformer.

En 2019 et 2020 :

- Sera proposé un modèle économique de l'EOSC ;
- Seront répertoriées les infrastructures de recherche partenaires, souhaitant proposer leurs services via l'EOSC ;
- Sera définie l'architecture de l'EOSC, s'accompagnant d'un ensemble de normes pour l'implémentation des principes FAIR (interopérabilité technique et juridique, certification des infrastructures de données partenaires, standards de données…)[133].

Les partenaires de l'EOSC seront dans un premier temps les infrastructures de recherche de la feuille de route de l'ESFRI[134] et celles des différentes feuilles de route nationales[135], sélectionnées sur la base du volontariat. Pour ces infrastructures, le défi majeur sera l'intégration des normes de données FAIR qui seront définies par l'EOSC. Le comité exécutif du projet prévoit d'ailleurs d'instaurer une certification FAIR : les infrastructures de données partenaires seront accréditées « FAIR », c'est-à-dire qu'elles devront satisfaire à des normes de qualité prédéfinies[136].

---

131 *Le nuage européen pour la science ouverte devient une réalité* (2018). Actualités du site de la Commission européenne. https://ec.europa.eu/commission/news/european-open-science-cloud-becomes-reality-2018-nov-23_fr (consulté le 19 septembre 2019).

132 *EOSC Declaration* (2017). https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf (consulté le 19 septembre 2019).

133 DIRECTION GÉNÉRALE POUR LA RECHERCHE ET L'INNOVATION DE LA COMMISSION EUROPÉENNE (2019). *European Open Science Cloud (EOSC) Strategic Implementation Plan*. https://publications.europa.eu/s/m1qV (consulté le 19 septembre 2019).

134 Cf. supra, 3.3.1, p.78

135 La feuille de route française des infrastructures de recherche est disponible sur le site web du Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation : http://www.enseignementsup-recherche.gouv.fr/cid70554/la-feuille-route-nationale-des-infrastructures-recherche.html#fr.

136 « Data infrastructures would operate in the EOSC according to FAIR data principles and seek to become FAIR-accredited/certified entities, meaning that their data services would meet over time infrastructural and quality standards under a quality-assurance scheme » (Direction générale pour la Recherche et l'Innovation de la Commission européenne 2019, op. cit., p.15)

Si l'EOSC peut bénéficier d'une mise en place rapide (2016-2020) grâce à l'existence d'infrastructures déjà en place, la principale difficulté sera donc probablement d'ordre politique (modèle économique, droits d'accès, interopérabilité…).

3.3.2.3. Engouement autour du projet

Le projet EOSC suscite l'engouement de nombreux militants de l'Open Science. Il est considéré comme le dispositif qui va révolutionner la diffusion et la réutilisation des données scientifiques.

- Lors de la réunion du Conseil de compétitivité le 29 mai 2018, les ministres européens chargés de la recherche et de l'innovation ont déclaré que l'EOSC allait *« changer la donne en matière de science ouverte en Europe »*, poursuivant : *« nous avons besoin d'initiatives concrètes rapidement pour faire de l'Europe le véritable leader de la science ouverte et le lieu le plus propice à l'innovation »*[137].
- Dans un énoncé de position, cinq associations et e-infrastructures européennes (LIBER[138], EUDAT, OpenAIRE, EGI[139] et GEANT[140]) ont donné leur soutien à la proposition de cloud et exposé leur vision. Dans ce document, l'EOSC est présenté comme le *« véhicule »* (le moyen) qui permettra d'ouvrir la science. L'ouverture de la recherche est elle-même considérée comme le *« moteur du progrès scientifique et de l'innovation économique et sociale »*[141].

Cet engouement autour de l'EOSC tient notamment au discours tenu par la Commission européenne. Dès son annonce en 2016, le projet de cloud a été présenté comme la réponse

137 « The EOSC is a game changer for open science in Europe, as we need urgent and concrete actions to make Europe the true open science leader and the best place for innovation », dans SKORDAS, T. (2018). 'European Open Science Cloud Council Conclusions', *Digital Single Market Blog Posts*. https://ec.europa.eu/digital-single-market/en/blogposts/european-open-science-cloud-council-conclusions (consulté le 19 septembre 2019).

138 Ligue des Bibliothèques Européennes de Recherche (https://libereurope.eu/)

139 https://www.egi.eu/

140 https://www.geant.org/

141 « Open Science is a key driver, not only of scientific progress, but also of economic and societal innovation. [...] The Open Science Cloud is the vehicle to achieve this vision », dans : EUDAT, LIBER, OPENAIRE, EGI, ET GEANT (2015). *Position Paper: European Open Science Cloud for Research*. https://doi.org/10.5281/zenodo.32915 (consulté le 19 septembre 2019). Page 1

européenne à l'enjeu d'une utilisation la plus large possible des données, entre disciplines scientifiques et entre secteur public et privé[142]. La Commission européenne s'appuie sur un registre à la mode – l'*open science* et la *data-driven science* – pour s'assurer de l'adoption du projet par les acteurs politiques de la recherche. L'EOSC fédère ainsi probablement parce qu'il véhicule l'idéal de la science ouverte des données.

3.3.2.4. Quels utilisateurs pour l'EOSC ?

L'EOSC est présenté comme un environnement qui sera dédié avant tout aux chercheurs, leur permettant de stocker, d'analyser et de réutiliser des données. Trouvera-t-il son public cible ?

Des domaines scientifiques ciblés

La recherche a certes besoin d'équipements toujours plus performants pour le calcul intensif, le transfert et le stockage de données. Mais cela concerne un nombre somme toute restreint de communautés de recherche. L'EOSC permettra de répondre à des problématiques qui touchent des domaines ciblés comme l'intelligence artificielle, la modélisation climatique ou les techniques biomédicales in silico. Autrement dit, il répondra aux besoins d'une recherche de pointe. L'EOSC sera-t-il réservé à la recherche d'« excellence » ? C'est une hypothèse probable, si l'on considère les conditions actuelles d'accès aux infrastructures de recherche. La plupart d'entre elles sélectionnent en effet leurs utilisateurs selon des critères d'excellence, en soumettant les propositions des candidats à l'avis d'un comité scientifique. Le périmètre de l'EOSC dépendra donc du modèle économique et de la politique d'accès qui seront retenus.

Quant à la diffusion et la réutilisation de données, elles constituent une stratégie politique de la Commission européenne avant d'être une tendance globale de la communauté scientifique.

142 « To fully exploit the potential of data as a key driver of Open Science and the 4th industrial revolution, Europe needs to answer several specific questions:
-How to maximise the incentives for sharing data and to increase the capacity to exploit them?
-How to ensure that data can be used as widely as possible, across scientific disciplines and between the public and the private sector?
-How better to interconnect the existing and the new data infrastructures across Europe?
-How best to coordinate the support available to European data infrastructures as they move towards exascale computing?
[…] This Communication proposes as a direct response a European Cloud Initiative which can secure Europe's place in the global data-driven economy » (Commission européenne 2016a, op. cit., p.2)

Comment s'assurer alors que les services de l'EOSC seront utilisés, si l'heure n'est pas à l'ouverture des données dans les équipes de recherche ? Un dispositif technique suffit-il à impulser de nouvelles pratiques ?

Des utilisateurs européens

Par ailleurs, il peut s'avérer délicat d'implémenter une infrastructure qui limite le périmètre géographique de la recherche. Comme le soulignait un groupe d'experts dans un livre blanc du réseau Science | Business[143], la recherche scientifique est globale, elle se fait dans des réseaux internationaux[144]. L'EOSC peut présenter l'inconvénient de segmenter la recherche à un niveau européen. Le groupe d'experts recommande de penser en amont la manière dont l'EOSC interagira avec les infrastructures des autres régions du monde et s'intégrera dans les réseaux d'échanges scientifiques.

Une implémentation top-down

Par ailleurs, l'EOSC prend-il suffisamment en compte les pratiques scientifiques, pour garantir son utilisation future par les communautés de recherche ?

Selon Ghislaine Chartron (2018), *« l'adhésion des chercheurs à des dispositifs numériques peut difficilement s'envisager sans l'implication de leur communauté »*. Cibler un public scientifique suppose pour l'EOSC de répondre à un besoin des communautés et d'y subvenir en tenant compte des pratiques et spécificités de chacune.

Les communautés se distinguent entre elles non seulement par la nature des données qu'elles collectent et l'utilisation qu'elles en font, mais aussi dans la façon dont elles gèrent et partagent les données.

> *« Il est admis que des efforts spécifiques sont nécessaires pour rendre les données découvrables et réutilisables, mais le degré de préparation au partage des*

143 SCIENCE|BUSINESS CLOUD CONSULTATION GROUP (2018). *Priorities for the European Open Science Cloud. White paper*. Science|Business. https://sciencebusiness.net/report/priorities-european-open-science-cloud (consulté le 16 septembre 2019).

144 On peut citer l'exemple de Euro-BioImaging (voir note de bas de page n°115, p.79). Euro-BioImaging est une des infrastructures de la feuille de route européenne. Elle fait partie d'un réseau international, le Global BioImaging Project, qui fédère des infrastructures du même ordre en Australie, en Inde...

> *données diffère encore beaucoup, y compris au sein des disciplines. Les infrastructures de données développées par les infrastructures de recherche disciplinaires sont souvent adaptées au projet ou au domaine de recherche concerné ; elles n'ont pas été conçues au départ pour être utilisées au-delà du projet ou de la discipline. A vrai dire, plusieurs des grandes infrastructures de recherche européennes existantes pourraient être considérées comme des e-infrastructures disciplinaires dédiées à l'interopérabilité et à la diffusion des données au sein de la discipline. »*[145]

Il semble donc difficile de faire l'économie d'une co-construction avec les chercheurs, tant les données et leur utilisation sont spécifiques à chaque thématique de recherche.

145 Traduction de : « It is recognised that specific efforts are needed for making data discoverable and reusable, but data sharing preparedness even within disciplines still differs a lot. The data infrastructures developed by disciplinary Research Infrastructures are often, for natural reasons, customised for the concerned project or research discipline domain and not primarily aimed at use beyond the project or discipline borders. In fact, several of the existing European large-scale Research Infrastructures could be classified as disciplinary e-Infrastructures focussing on disciplinary interoperability and access to data. » (European Strategy Forum on Research Infrastructures 2018, op. cit., p.121)

4. Politiques et initiatives de l'État français

4.1. Législation relative à la diffusion des données scientifiques

4.1.1. A l'origine : la loi CADA (1978 / 2005)

Il existe en France une législation qui régit l'accès et la réutilisation des informations du secteur public. En 1978 a été promulguée une loi relative aux relations entre l'administration et les usagers (dite « loi CADA »)[146], instaurant un droit d'accès aux documents administratifs. Elle a été modifiée en 2005, lorsqu'a été transposée en droit français la directive européenne du 17 novembre 2003 concernant la réutilisation des informations du service public (dite « directive PSI »)[147]. Cette modification a établi un droit à la libre réutilisation des informations du secteur public. Un régime dérogatoire a néanmoins été accordé aux établissements d'enseignement et de recherche, qui restaient libres de fixer les conditions de réutilisation des informations qu'ils produisaient. Jusqu'au tournant des années 2000, la politique de recherche française était en effet davantage tournée vers la valorisation économique des résultats scientifiques que vers leur ouverture. C'est en partie pourquoi un régime dérogatoire avait été accordé aux établissements de recherche, ceux-ci pouvant réguler la réutilisation des informations qu'ils produisaient. Jusqu'en 2016, c'est donc la loi CADA qui, sur le plan juridique, régissait les conditions de diffusion des données de la recherche française.

4.1.2. La loi Valter et la loi pour une République numérique (2015 / 2016)

Quinze ans plus tard, la rhétorique de l'ouverture (*openness*) est devenue omniprésente, s'étendant du domaine scientifique aux domaines politique et économique. Elle s'est traduite dans la législation française par l'introduction de la loi Valter en 2015 et de la loi pour une République numérique en 2016.

146 *Loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal* (dite « loi CADA »). https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000339241 (consulté le 19 septembre 2019).

147 Directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public 2003, op. cit.

Promulguée le 28 décembre 2015, la loi relative à la gratuité et aux modalités de la réutilisation des informations du secteur public (dite « loi Valter »)[148] est la transposition de la directive européenne du 26 juin 2013[149]. Elle rend gratuite la réutilisation des informations issues du secteur public.

La loi pour une République numérique[150] a, quant à elle, été promulguée le 7 octobre 2016. Elle impose aux administrations de plus de 2 500 agents de mettre en ligne de manière spontanée et dans un format ouvert les documents produits dans le cadre de leurs activités. Ces documents deviennent alors librement réutilisables, y compris à des fins commerciales. La loi pour une République numérique instaure ce que Manuel Valls, alors Premier ministre, a qualifié de *« principe d'open data par défaut »*[151]. Même si les données de la recherche n'y sont que brièvement mentionnées, elles ne font désormais plus exception et doivent répondre à ce même principe d'ouverture et de libre réutilisation. L'article 11 de l'ancienne loi CADA, qui accordait un statut dérogatoire aux informations issues des établissements de recherche, a en effet été abrogé. Seule la protection de droits appartenant à des tiers (propriété intellectuelle, vie privée, confidentialité et secrets) peut désormais justifier la non diffusion de ces informations (Maurel 2018b).

L'unique mention du terme de données de recherche concerne l'article 30 du texte de loi :

> *« Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre.*

148 *Loi n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public* (dite « loi Valter ») https://www.legifrance.gouv.fr/eli/loi/2015/12/28/PRMX1515110L/jo/texte (consulté le 19 septembre 2019).

149 Directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013 modifiant la directive 2003/98/CE concernant la réutilisation des informations du secteur public, op. cit.

150 *Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique*. https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte (consulté le 19 septembre 2019).

151 VALLS, M. (2015). *Présentation de la stratégie numérique du Gouvernement*. https://www.gouvernement.fr/partage/4972-discours-de-manuel-valls-lors-de-la-presentation-de-la-strategie-numerique-du-gouvernement-a-la (consulté le 19 septembre 2019).

> *L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication.*
>
> *Les dispositions du présent article sont d'ordre public et toute clause contraire à celles-ci est réputée non écrite. »*

Cet article vise à protéger le principe de libre réutilisation des données de recherche. L'objectif est d'empêcher des éditeurs privés de s'arroger l'utilisation exclusive des données de recherche. Au moment de la publication d'un article scientifique, il arrive que des éditeurs demandent aux chercheurs de leur remettre les données sous-jacentes et qu'ils insèrent dans le contrat d'édition une clause d'utilisation exclusive. L'article 30 a rendu nulle toute clause de ce type.

L'article 30 de la loi pour une République numérique est en partie dû aux acteurs de la science ouverte (principalement des professionnels de l'information scientifique et technique), qui ont contribué à l'élaboration du texte législatif, dans le cadre de la consultation publique ouverte en 2015. Le projet de loi avait en effet été soumis à une contribution en ligne des citoyens[152]. L'article 30 introduit donc le terme de « données de recherche » dans le Code de la recherche (article L533-4), lui conférant ainsi une existence juridique, quoique sans en donner de définition. De la même manière, le terme synonyme de « données scientifiques » avait été introduit dans l'article L112-1 par la loi relative à l'enseignement supérieur et à la recherche de 2013. Celle-ci avait en effet ajouté aux objectifs de la recherche publique française *« l'organisation de l'accès libre aux données scientifiques »*[153].

Le droit français fait donc des données de recherche un type d'information publique et scientifique à part entière. Cela témoigne de l'émergence, depuis la loi CADA, du concept de donnée de recherche. Ce dernier a commencé à être pensé de manière transdisciplinaire sous cette appellation à partir des années 2000 environ. Cela témoigne également de la reconnaissance par l'État de ce type d'information et de ses spécificités. D'après un rapport du Digital Curation Centre et de SPARC Europe publié en 2018[154], 13 pays européens, dont la France, disposent actuellement d'une politique d'ouverture des données de la recherche (sur

152 La chronologie du projet de loi est disponible sur https://www.gouvernement.fr/action/pour-une-republique-numerique

153 *Article L112-1 du Code de la recherche*. https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006071190&idArticle=LEGIARTI000006524135&dateTexte=&categorieLien=cid (consulté le 19 septembre 2019).

un total de 28 États membres). La France est le seul pays avec la Lituanie[155] à avoir entériné dans la loi cette question de l'ouverture des données. Dans les autres pays, l'ouverture des données fait plutôt l'objet de politiques venant de financeurs de la recherche, de feuilles de route nationales ou de codes d'éthique. Le rapport qualifie la politique mise en place par la France de « douce » (*soft*). Celle-ci se révèle en effet peu contraignante, se concentrant davantage sur des droits en matière d'ouverture que sur des obligations.

4.2. Initiatives du Ministère de la Recherche en matière de données de recherche

4.2.1. La Bibliothèque Scientifique Numérique (2009-2017) et son successeur, le Comité pour la Science Ouverte (2018-)

Au niveau gouvernemental, c'est le Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI) qui porte la question des données scientifiques. Celle-ci a été intégrée à une structure préexistante : la Bibliothèque Scientifique Numérique (BSN). Créée en 2009, la BSN est un dispositif national de coordination des actions en matière d'information scientifique et technique. Elle a vu le jour dans un contexte d'évolution des modèles de communication scientifique et de leurs acteurs. Placée sous la tutelle du Ministère, elle se compose de différents groupes de travail (appelés « segments »), axés sur des thématiques précises (l'édition scientifique, les archives ouvertes, l'archivage pérenne…). Au sein de ces groupes, il avait été choisi de réunir les représentants de différentes structures de l'enseignement supérieur et de la recherche (universités, organismes de recherche, associations professionnelles, infrastructures de recherche…). En rassemblant les différents acteurs de l'ESR, le Ministère imaginait renforcer les pratiques de coopération et favoriser l'utilisation d'outils nationaux mutualisés. Chaque segment avait pour mission : d'instruire

154 SPARC EUROPE (2018). *An Analysis of Open Data and Open Science Policies in Europe*. Version 3. https://sparceurope.org/latest-update-to-european-open-data-and-open-science-policies-released/ (consulté le 19 septembre 2019).

155 En Lituanie, la loi relative à l'enseignement supérieur et à la recherche (adoptée en 2009 et révisée en 2015 puis 2016) stipule que les résultats des travaux de recherche réalisés par les institutions publiques d'enseignement supérieur et de recherche doivent être rendus publics.
Voir le texte de loi : https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/548a2a30ead611e59b76f36d7fa634f8?jfwid=rp9xf47k7

des sujets thématiques ; de proposer des orientations ; d'accompagner des actions associées. La gouvernance était assurée par un comité de pilotage, secondé par un comité technique, composé des pilotes des différents segments et chargé d'instruire les décisions du comité de pilotage.

Au départ, les données de la recherche ne faisaient pas partie des thématiques « officielles » de la Bibliothèque Scientifique Numérique. C'est en 2013, quatre ans après le lancement de la BSN, qu'a été créé un segment dédié : BSN10. Ce nouveau groupe répondait à la même structure que les 9 autres d'ores et déjà en place. Parmi ses membres figuraient des personnes travaillant déjà sur la question des données à l'échelle de leur établissement de recherche. Ces personnes étaient principalement issues d'organismes de recherche comme le CNRS[156], le CEA[157], l'INRA[158], l'INSERM[159] et l'IRSTEA[160]. Une des raisons est que ces organismes étaient considérés comme plus avancés que les universités, ayant à cette époque d'ores et déjà mis en place une politique institutionnelle de gestion et d'ouverture des données au sein de leur établissement. Par ailleurs, les membres du groupe BSN10 étaient pour la plupart des professionnels de l'information scientifique et technique.

De 2013 à 2017, le groupe a réalisé plusieurs études et émis différentes recommandations sur des sujets comme la fouille de données ou l'article 30 de la loi pour une République numérique[161]. Il a également coordonné des actions collectives, dont Cat OPIDoR, catalogue référençant l'offre de services française en matière de données de recherche[162]. L'action de BSN10 s'est donc concentrée davantage sur l'étude du paysage français que sur la mise en place de dispositifs nouveaux pour gérer et partager les données de la recherche. Cela tient notamment à son mode de fonctionnement. Le fait d'être composé de membres venant d'horizons divers lui apportait certes une richesse en termes de points de vue et de

156 Centre National de la Recherche Scientifique (http://www.cnrs.fr/)

157 Commissariat à l'énergie atomique et aux énergies alternatives (http://www.cea.fr/)

158 Institut National de la Recherche Agronomique (http://www.inra.fr/)

159 Institut National de la Santé et de la Recherche Médicale (https://www.inserm.fr/)

160 Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture (https://www.irstea.fr/)

161 Voir supra, 4.1.2, p.88

162 Voir infra, troisième partie, p.101

propositions, mais créait une sorte d'inertie due au peu de temps que chacun de ses membres pouvait lui consacrer (la plupart d'entre eux occupaient en effet des postes à responsabilités dans leurs propres établissements).

En 2018, la Bibliothèque Scientifique Numérique a cédé la place au Comité pour la Science Ouverte (CoSO). Ce dernier a été créé dans le cadre de la mise en place d'un plan national pour la science ouverte (voir ci-dessous). Fondé sur une structure similaire à la BSN, le CoSO se veut composé d'experts de toutes professions et disciplines concernées par la science ouverte. Il est organisé non plus en dix segments mais en quatre collèges :

- Un collège Publications
- Un collège Données de la recherche
- Un collège Compétences
- Un collège Europe & International

Cette nouvelle répartition montre l'importance qu'ont prise les données de la recherche dans la politique nationale d'ouverture de la science entre 2009 et 2018.

4.2.2. Un plan national pour la science ouverte (2018)

Le 4 juillet 2018, la Ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation, Frédérique Vidal, a annoncé le lancement d'un plan national pour la science ouverte. Elle a présenté la transition de la France vers la science ouverte comme un enjeu de compétitivité[163].

Doté d'un budget de 5,4 millions d'euros pour l'année 2018-2019, ce plan se compose de trois axes : le premier axe vise à généraliser le libre accès aux publications ; le deuxième axe à structurer et ouvrir les données de la recherche ; le troisième axe à inscrire la France dans une dynamique européenne et internationale[164].

163 Vidal 2018, op. cit.

164 Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation 2018, op. cit.

Ce plan entend répondre à un double objectif :

- Celui de l'*Amsterdam Call for Action on Open Science*[165], lancé sous la présidence néerlandaise au Conseil de l'Union européenne et appelant à une démocratisation de l'accès aux savoirs ;
- Celui de l'Open Government Partnership[166], visant à développer la transparence de l'action publique.

En voulant à la fois rendre la science plus transparente, accélérer le progrès scientifique, nourrir la formation et contribuer à l'innovation, le plan national pour la science ouverte semble donc nourrir de grandes ambitions. On peut se questionner sur la possibilité de les honorer toutes. Est-il possible de mettre en place des solutions concrètes, qui tiennent compte de tous ces objectifs ?

Pour ce qui est de l'axe consacré aux données scientifiques, l'objectif est que *« les données produites par la recherche publique soient progressivement structurées en conformité avec les principes FAIR (Faciles à trouver, Accessibles, Interopérables, Réutilisables), préservées et, quand cela est possible, ouvertes »*[167]. Le plan national recommande également *« l'adoption d'une politique de données ouvertes associées aux articles et le développement des* data papers[168] *»*[167].

Afin d'accélérer cette transition, a été planifié le lancement d'un appel à projets. Intitulé « ANR Flash », cet appel a été publié le 28 mars 2019 par l'Agence Nationale de la Recherche (ANR)[169]. Il vise à *« demander à la communauté scientifique elle-même de proposer, domaine par domaine, discipline par discipline, spécialité par spécialité, comment elle peut appliquer les principes de la science ouverte à propos des données de la recherche »*[170]. Le Ministère met donc entre les mains de la communauté scientifique la question de comment ouvrir les données.

165 *Amsterdam Call for Action on Open Science* (2016). https://www.ouvrirlascience.fr/amsterdam-call-for-action-on-open-science/ (consulté le 19 septembre 2019).

166 Déclaration du gouvernement ouvert 2011, op. cit.

167 Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation 2018, op. cit., p.6

168 Un *data paper* est « une publication qui décrit un jeu de données scientifiques, notamment à l'aide d'informations structurées appelées métadonnées. Contrairement aux articles de recherches classiques, les data papers fournissent une voie formalisée au partage des données plutôt que tester des hypothèses ou présenter de nouvelles analyses » (source : https://doranum.fr/data-paper-data-journal/fiche-synthetique/).

169 https://anr.fr/

Le plan national pour la science ouverte prévoit également plusieurs autres mesures.

- Sur le plan administratif : sera créée la fonction d'administrateur des données au ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI). Cette fonction existe déjà dans d'autres ministères. Elle avait été instituée en 2014 par le décret n°2014-1050[171], dans le cadre de la politique nationale d'Open Data. L'administrateur du MESRI sera chargé de coordonner l'action publique en matière de données scientifiques et d'animer un réseau d'administrateurs des données dans les établissements de recherche.
- Sur le plan financier : les dépenses de traitement des données seront rendues éligibles dans les appels à projets.
- Sur le plan pédagogique : des offres de formation seront proposées à la communauté scientifique, pour développer les compétences sur les données de la recherche.
- Sur le plan technique : des centres de données thématiques et disciplinaires seront développés ; la mise en place de plans de gestion de données dans les appels à projets de recherche sera généralisée (depuis 2019, l'ANR impose ainsi un plan de gestion de données aux projets de recherche qu'elle finance[172]).

Ces différentes mesures se focalisent sur l'ouverture des données, c'est-à-dire sur la manière de les rendre accessibles et réutilisables. La question de leur réutilisation effective n'est pas abordée. Le plan pour la science ouverte ne prévoit pas de stratégie pour stimuler la réutilisation des données qui seront mises en ligne. Peut-être est-ce encore trop tôt ? C'est en tout cas un aspect qu'il faudra aborder, car le défi est aussi économique. Ce problème a déjà été identifié par les acteurs de l'Open Data, qui ont constaté que la réutilisation des informations publiques mises en ligne demeurait limitée[173].

---

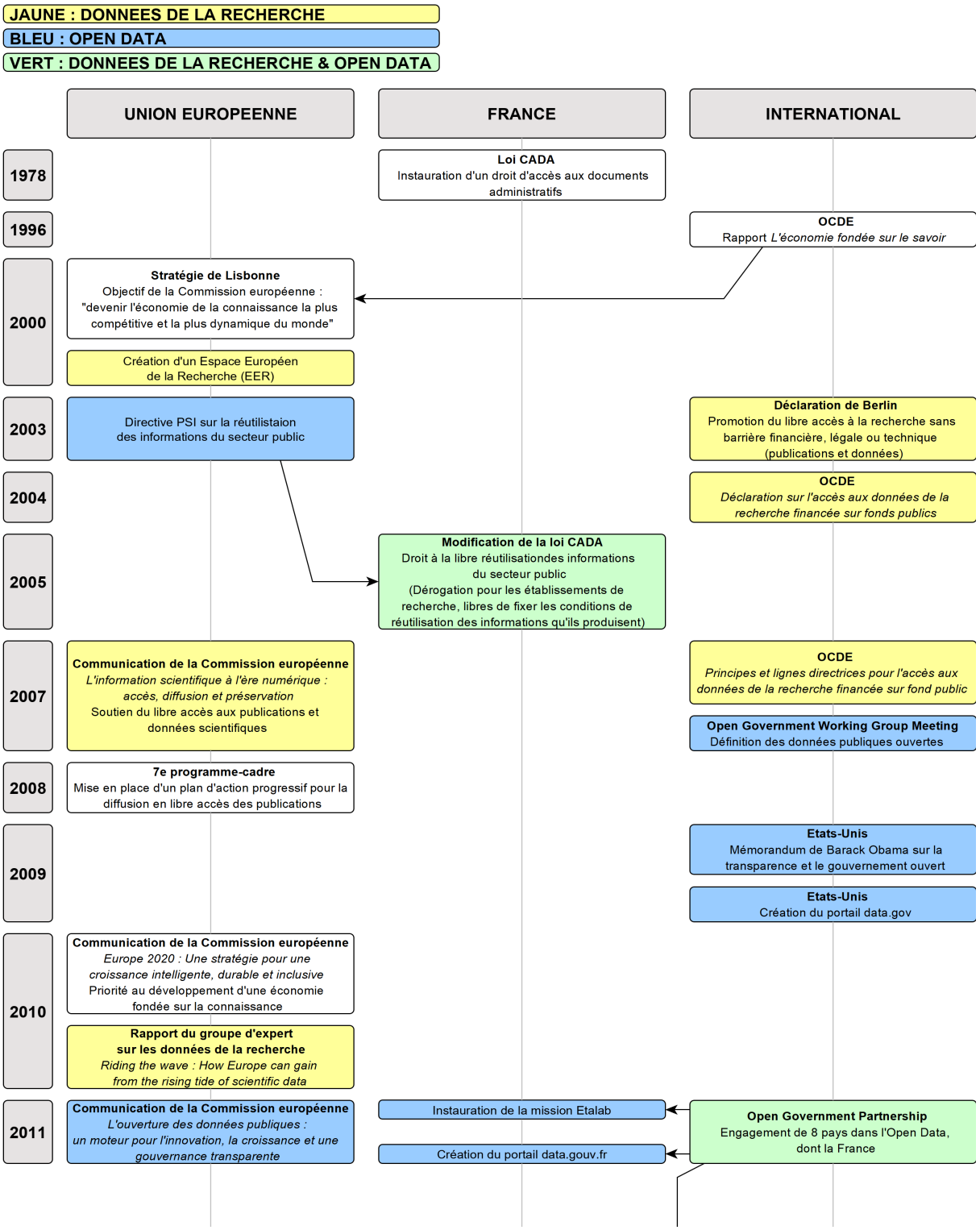
170 AGENCE NATIONALE DE LA RECHERCHE (2019). *Appel FLASH Science ouverte : pratiques de recherche et données ouvertes*. http://www.agence-nationale-recherche.fr/fileadmin/aap/2019/aap-data-2019.pdf (consulté le 19 septembre 2019).

171 *Décret n°2014-1050 du 16 septembre 2014 instituant un administrateur général des données*. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000029463482&categorieLien=id (consulté le 19 septembre 2019).

172 AGENCE NATIONALE DE LA RECHERCHE (2018). *Plan d'action 2019*. https://anr.fr/fileadmin/documents/2018/Plan-d-action-ANR-2019.pdf (consulté le 19 septembre 2019).

173 Voir supra, 2.3, p.63

5. Conclusion



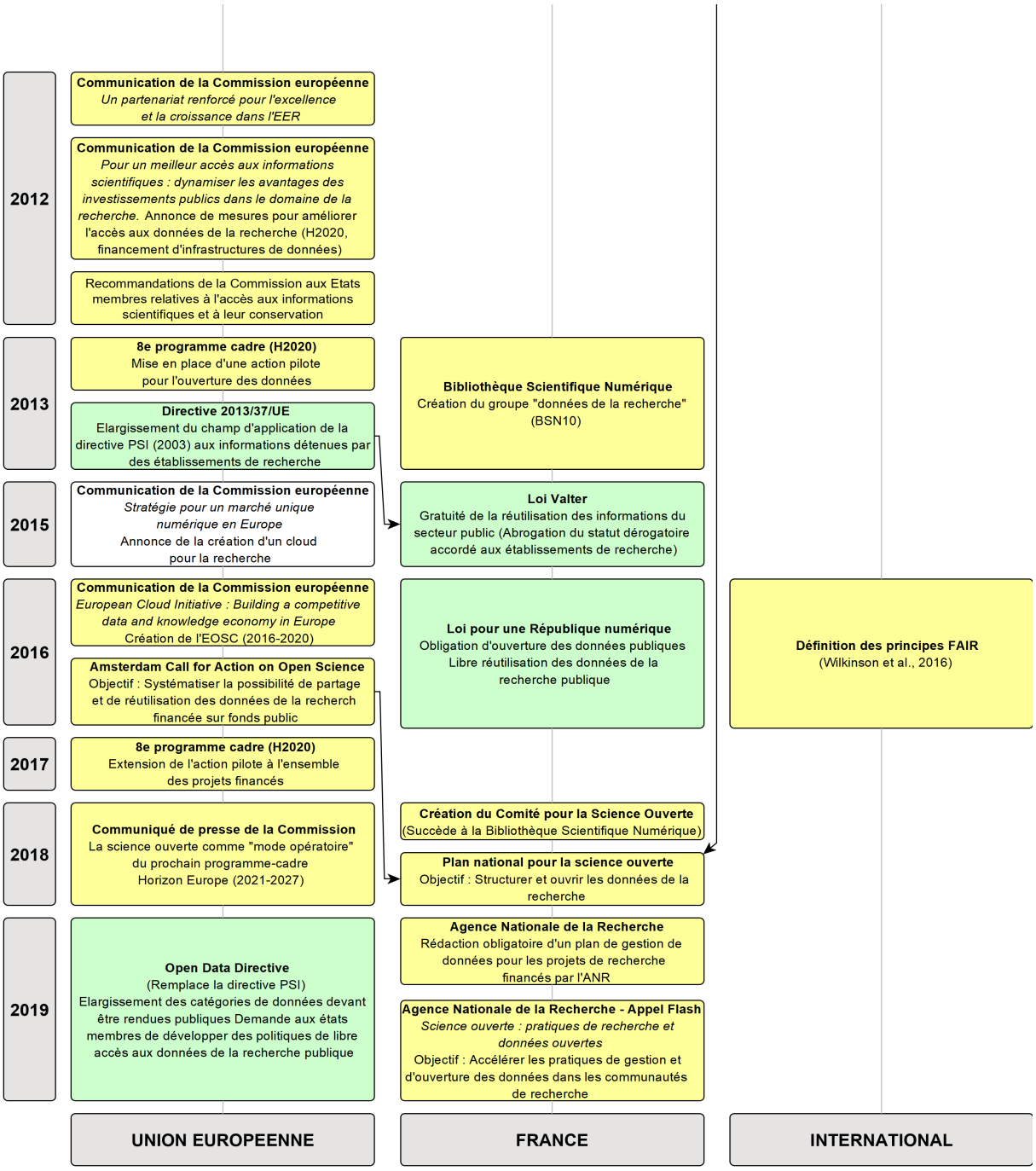


Figure 4 : Chronologie des politiques publiques d'ouverture des données

A l'instar de l'Open Data, le principe d'ouverture des données de la recherche est mobilisé dans les politiques publiques pour servir simultanément plusieurs objectifs. De la Déclaration de l'OCDE (2004) au plan national français pour la science ouverte (2018), en passant par la création d'un espace européen de la recherche (2012), l'ouverture des données est prônée à la fois pour accélérer le progrès scientifique et stimuler l'innovation. Les politiques d'ouverture conjuguent donc une vision libertaire et une vision néolibérale de la science (Chartron 2016, p.2). L'orientation de la Commission européenne semble pourtant sous-tendre fondamentalement la seconde vision : celle d'une science au service de l'innovation. Comment expliquer alors cette double argumentation ? On peut s'appuyer, pour y répondre, sur le concept de *buzzword*, décrit par l'historienne Bernadette Bensaude Vincent (2014). L'expression d'« ouverture des données » fonctionne en effet comme un *buzzword*, c'est-à-dire comme un mot ou groupe de mots capable de fédérer des personnes d'horizons différents. La puissance fédératrice des *buzzwords* vient de leur connotation positive et du fait qu'ils véhiculent des valeurs présentes dans la société en général. Ils pointent un objectif, un but à atteindre et, ce faisant, créent une dynamique autour de l'élaboration de principes d'action (dans le cas des données, les principes FAIR en sont un exemple). De manière générale, le terme d'« ouverture » connote la libération de quelque chose qui jusqu'alors ne bénéficiait qu'à quelques uns. La notion d'ouverture des données de la recherche fait par ailleurs appel à des valeurs scientifiques (la connaissance comme bien commun[174]) ainsi qu'aux valeurs d'un Internet sans frontières, dans lequel les données sont des biens communs numériques et informationnels (Latrive 2000). Les *buzzwords*, et l'ouverture des données en particulier, se rapprochent de ce que Star et Griesemer (1989) nomment des « objets-frontières » (*boundary objects*). Ce sont des mots malléables, dont le sens est suffisamment imprécis et à la fois suffisamment accessible, pour susciter des interprétations diverses. Ils peuvent ainsi être adoptés par des catégories d'acteurs différentes, qui chacune les adaptent au cadre de référence qui leur est propre. C'est probablement ce qui explique le succès de l'ouverture des données de la recherche (et de la science ouverte en général). Leur flexibilité sémantique permet de réunir autour d'un même principe d'action – l'ouverture – des personnes aux

174 Dans l'« ethos de la science » de Merton (1973), une des quatre normes régulant l'activité scientifique est le « communisme ». Selon cette norme, toute découverte scientifique, en tant que produit de la collaboration entre chercheurs, est un bien commun et doit donc systématiquement être publiée.

perspectives divergentes, les unes plaçant la recherche au cœur de la croissance économique, les autres ayant une vision plus traditionnelle et humaniste de la science.

Bernadette Bensaude-Vincent attribue la prolifération des *buzzwords* dans le langage de l'innovation scientifique et technologique au régime actuel de la recherche, qui est de plus en plus lié à l'économie et à la société.

> *« L'entreprise scientifique est dépendante de la politique de recherche et est explicitement destinée à la société et à la compétitivité économique. Son orientation étant déterminée par la politique scientifique et les demandes du marché, un nombre croissant d'acteurs s'y trouvent donc impliqués (la Commission européenne, les ministères, les banques privées, les entreprises, les chercheurs et les ingénieurs, les sociétés civiles, les ONG). Les buzzwords sont essentiels pour rassembler une telle variété d'acteurs. Ils créent une « zone d'échange » qui permet à ces différents acteurs de communiquer. »*[175] (Bensaude Vincent 2014)

En articulant les différentes conceptions de l'ouverture des données, les politiques publiques parviennent ainsi à fédérer une grande diversité d'acteurs. Ce faisant, elles catalysent l'engagement de ces derniers dans l'élaboration de solutions concrètes pour la structuration et la mise à disposition des données. C'est le cas des acteurs qui se sont impliqués dans le développement d'une offre de services, comme nous le verrons dans la partie suivante. Qu'en est-il cependant de la sphère scientifique ? Les politiques d'ouverture parviennent-elles à toucher les communautés de recherche ? Cette interrogation sera étayée par les résultats de l'enquête sur les pratiques et stratégies des chercheurs, dans la quatrième partie de la thèse.

175 Traduction de : « Scientific endeavours are dependent on science policy and explicitly aimed at society and economic competitiveness. As science policy and market demands are driving the orientation of scientific efforts an increasing number of actors are involved – the European Commission, national agencies and ministries, private banks, industrial companies, scientists and engineers, civil societies and NGOs (Bensaude Vincent, 2009). Buzzwords are crucial to bring together such a variety of people. As shallow linguistic units deprived of substantial meanings, they create a 'trading zone' that allows different stakeholders to communicate. » (Bensaude Vincent 2014)

Troisième partie

-

Les services d’appui à la gestion et au partage des données de recherche

Les politiques publiques de gestion et d'ouverture des données de la recherche ont généré la création de services dédiés. Des corps intermédiaires, comme les professionnels de la documentation (bibliothécaires, archivistes…), se sont mis à développer des services destinés à accompagner les chercheurs dans l'application de ces recommandations politiques.

Face à la multiplication de ces services de gestion et d'ouverture, des études ont été menées afin de les répertorier : Tenopir et al. (2012 ; 2015b ; 2017) ont notamment étudié quels types de services de données étaient délivrés par les bibliothèques universitaires européennes et nord-américaines ; ils ont mené des enquêtes et montré que les bibliothèques offraient plus fréquemment des services d'information et de conseil que des services techniques, tels que la préparation de données en vue de leur dépôt dans un entrepôt. Cécile Delay-Artous (2017) s'est quant à elle concentrée sur les services de données de recherche dans le domaine des sciences humaines et sociales à l'échelle internationale. Elle a répertorié les initiatives et les acteurs sous forme de représentation graphique, soulignant la rapidité avec laquelle cette représentation pouvait devenir obsolète. Cette remarque explique peut-être pourquoi certaines cartographies des services de données prennent la forme de catalogues actualisés de manière régulière. Le répertoire d'entrepôts de données scientifiques Re3data (*Registry of Research Data Repositories*)[176] est un des plus connus à l'échelle internationale (Kindling et al. 2017). Il recense les infrastructures dédiées à la diffusion et la conservation des données scientifiques. Aux Pays-Bas, l'Université de Leyde a créé un catalogue des équipements destinés aux chercheurs pour la gestion des données : le *Leiden Research Data Information Sheets*[177]. Son périmètre est plus large que celui du Re3data : il inclut non seulement les entrepôts de données mais aussi les archives de données et les outils d'aide à la rédaction de plans de gestion de données.

Un recensement du même type que le *Leiden Research Data Information Sheets* a été commandité en 2015 par le segment 10 de la Bibliothèque Scientifique Numérique (BSN10). L'objectif était de *« réaliser une étude préliminaire de cartographie des centres et services de données dans les différentes institutions et organismes de recherche nationaux »*[178]. J'y ai contribué, en parallèle de la thèse, à titre de chargée de mission. Cette étude m'a donné à voir

176 http://www.re3data.org/

177https://vre.leidenuniv.nl/vre/lrd/Pages/information-Sheets.aspx

178Extrait de la demande de subvention au Ministère de la Recherche, dans le cadre de l'action BSN10 « Cartographie nationale des centres et services de données » [document à diffusion confidentielle].

tout un pan des données de la recherche, que je n'aurais peut-être pas exploré avec autant d'attention, si je n'avais pas eu l'opportunité de ce contrat. La thèse, qui au départ devait se concentrer sur le rapport des chercheurs aux données de la recherche, s'est ainsi vue élargie à la question des services de données. Cette partie vise à dresser un portrait des dispositifs mis à la disposition de la communauté de recherche française pour conserver et diffuser les données scientifiques, dans la lignée des politiques publiques sur le sujet. L'objectif est d'alimenter la réflexion sur l'adéquation des services de données avec les pratiques des chercheurs (notre seconde question de recherche). Cette partie constitue donc une première phase de contextualisation, permettant de comprendre quelles sont la nature et les caractéristiques des services de données en France.

1. Terrain et méthodologie

1.1. Un recensement des services de données par la Bibliothèque Scientifique Numérique

BSN10 est un groupe de travail sur les données de la recherche. Il constituait une des entités de l'ancienne Bibliothèque Scientifique Numérique, devenue aujourd'hui Comité pour la Science Ouverte[179]. Investi d'une mission de conseil auprès du Ministère de la recherche, le groupe était chargé de proposer des orientations en matière de gestion et d'ouverture des données de la recherche. Ses premiers travaux ont été d'ordre prospectif, visant à dresser un état des lieux du paysage français.

Le groupe s'est notamment intéressé aux services dédiés aux données de recherche, donnant lieu à la commande d'une cartographie nationale des services[180]. Initiée en 2015, celle-ci visait à :

- Mieux connaître les services de données existant en France ainsi que leurs modes d'organisation ;
- Identifier d'éventuelles lacunes et informer les acteurs politiques des secteurs où l'investissement de ressources pourrait être nécessaire.

La commande initiale était peu claire. Il a donc d'abord fallu définir le périmètre de la cartographie à réaliser. Les données de recherche sont une réalité complexe, dont le caractère se répercute sur les infrastructures qui leur sont liées. Le paysage des services de données s'est en effet révélé être un paysage hétérogène et peu visible. Des difficultés ont été rencontrées à plusieurs niveaux pour établir le périmètre de la cartographie.

- Premier niveau de difficulté, lié à la notion de donnée de recherche : les services de données utilisent-ils le terme de données de recherche ? Si non, faut-il tout de même inclure ces services dans la cartographie ?

179 Voir supra, deuxième partie, 4.2, p.91

180 Voir l'offre de poste publiée à cette occasion en annexe 1

- Deuxième niveau de difficulté, lié à la définition de ce qu'est un service de données : quelles sont ses missions ? Se limitent-elles à la diffusion des données ? Si oui, doit-on uniquement répertorier les services de type entrepôts de données ? Si non, quelles sont les autres fonctions possibles d'un service de données ?
- Troisième niveau difficulté, lié à l'expression « services de données nationaux » : dans quel sens doit-elle être entendue ? S'agit-il des services dont ont besoin les chercheurs qui travaillent en France ? S'agit-il des centres de données implantés en France ? Ou bien s'agit-il des centres de données ayant pour public cible (entre autres) les chercheurs de la recherche française ?

Avant d'entamer le travail de cartographie, ces différentes interrogations ont dû être tranchées.

- Les entités recensées ont été regroupées sous le terme générique de « services dédiés aux données de recherche », bien qu'elles-mêmes ne se dénomment pas ainsi. Par « service », on entend la fourniture de ressources humaines et/ou techniques pour gérer les données à une ou plusieurs étapes d'un projet de recherche. Dans « service », il y a l'idée d'accomplir une tâche pour autrui, ici dans le sens d'une ressource humaine ou technique, offerte aux chercheurs pour la gestion et l'ouverture des données.
- L'équipe projet[181] a par ailleurs décidé que la cartographie s'appuierait sur le cycle de vie des données[182] (figure 5) et qu'elle couvrirait, par conséquent, aussi bien les services de diffusion que les services de collecte, d'analyse et d'archivage des données.
- L'expression de « services de données nationaux » a été entendue dans le sens de « services fournis par des structures françaises ». De plus, seuls les services proposés en tout ou partie par des structures publiques, c'est-à-dire affiliés à une institution ou

181 La composition de l'équipe projet est détaillée en annexe 2.

182 Dans une perspective d'ouverture de la science, « le cycle de vie des données de la recherche est l'ensemble des étapes de gestion, conservation, diffusion et réutilisation des données scientifiques liées aux activités de recherche » (Deboin 2018). Chaque étape du cycle a donné lieu à la définition de normes (schémas de métadonnées pour décrire les données, modèles de plan de gestion de données…) et à la mise en place d'infrastructures (entrepôts de données, centres d'archivage…).

une infrastructure de recherche financée par l'État français, ont été répertoriés. Les services fournis par des structures privées ont donc été exclus. Tel n'a pas été le choix des répertoires Re3data et *Leiden Research Data Information Sheets*, dans lesquels sont référencés à la fois des initiatives publiques et des initiatives privées. Il existe en effet des structures privées, certes peu nombreuses et souvent implantées à l'étranger, qui proposent des services de gestion et d'ouverture des données aux communautés de recherche. L'entrepôt Figshare[183] de la société Digital Science en est probablement l'exemple le plus connu.

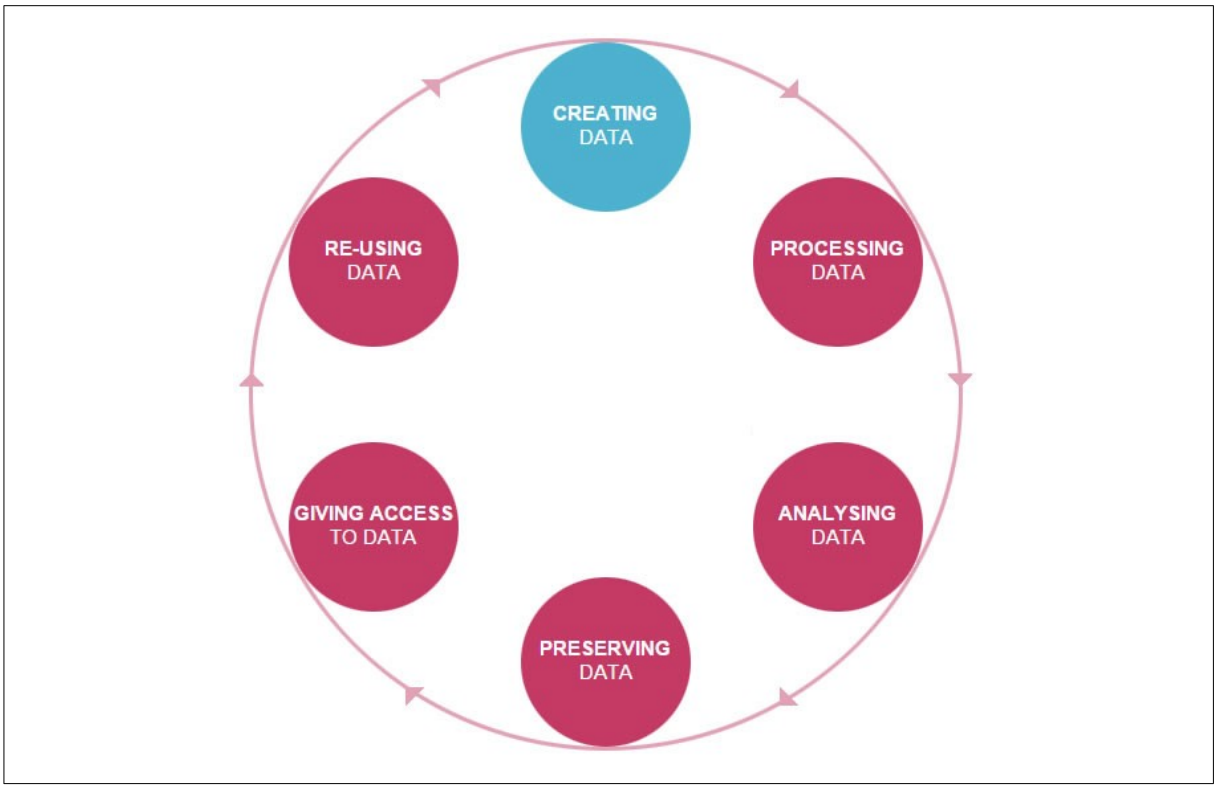


Figure 5 : Cycle de vie des données de la recherche[184]

183 https://figshare.com/

184 Source : UK Data Archive, University of Essex (le document original n'est plus en ligne à ce jour)

1.2. Phase 1 : une première cartographie

Une fois le périmètre de la cartographie déterminé, une méthodologie a été définie pour identifier les services de données et les analyser. Elle était composée de quatre étapes : identifier l'offre de services ; élaborer une typologie des services ; concevoir une grille d'analyse pour chaque type de service ; recenser et analyser les services.

1.2.1. Identification de l'offre de services

La première étape a consisté à identifier les structures fournissant des services de gestion de données. Pour ce faire, plusieurs approches ont été employées :

- Une veille a été mise en place pour repérer les colloques ayant lieu sur le thème des données de recherche. Leur programme était alors consulté, afin d'identifier d'éventuels services y étant présentés.
- Les membres du groupe BSN10, ceux de l'association EPRIST[185] et le groupe de travail pour l'accès ouvert du consortium COUPERIN[186] ont été interrogés sur l'existence de services de données dans leurs établissements d'appartenance.
- Par la suite, au cours des entretiens avec les fournisseurs de services, ceux-ci ont parfois fait mention de l'existence de services qui n'avaient pas été repérés jusque là.

1.2.2. Élaboration d'une typologie des services

La deuxième étape a eu pour but d'établir une typologie des services identifiés. Ceux-ci ont été classés selon leur fonction (du moins selon leur fonction principale).

185 EPRIST est l'association des responsables de l'information scientifique et technique des organismes de recherche français publics ou d'utilité publique (https://www.eprist.fr/).

186 COUPERIN est le consortium unifié des établissements universitaires et de recherche pour l'accès aux publications numériques (https://www.couperin.org). Son groupe de travail pour l'accès ouvert (GTAO) a pour mission de promouvoir et d'implémenter le libre accès dans les établissements membres du consortium (https://www.couperin.org/services-et-prospective/open-access/gtao).

La typologie finale se compose des 9 catégories suivantes :

- Information : site web agrégeant des informations et des actualités sur le thème des données de recherche.
- Formation : service de formation voire d'auto-formation, en présentiel ou à distance, portant sur un ou plusieurs aspects de la gestion des données de recherche.
- Accompagnement : service ayant pour mission d'offrir aux personnels de recherche une aide personnalisée dans la gestion des données scientifiques (accompagnement dans la rédaction d'un plan de gestion de données, aide au dépôt des données ou à la création d'une base de données…). Les membres de l'équipe d'accompagnement disposent d'une expertise informatique, documentaire, archivistique et/ou juridique.
- Outil de gestion de données : outil permettant de planifier ou de mettre en œuvre la gestion, diffusion ou réutilisation de données de recherche (aide à la rédaction de plans de gestion de données, éditorialisation de données, attribution d'identifiants pérennes...).
- Plateforme d'acquisition : infrastructure mettant à la disposition des équipes de recherche des moyens techniques et humains pour la collecte de données.
- Plateforme de calcul : infrastructure mettant à la disposition des équipes de recherche des moyens informatiques et humains pour le calcul intensif, à des fins de simulation, de modélisation et/ou d'analyse numériques.
- Annuaire de données : base de données en ligne décrivant des jeux de données scientifiques et permettant la saisie, la consultation voire l'export de fiches descriptives.
- Entrepôt de données : plateforme de dépôt et d'accès à des données numériques, garantissant la conservation et l'accessibilité aux jeux de données à plus ou moins long terme.
- Plateforme d'archivage : plateforme accueillant des données de recherche au format numérique, dans le but de les conserver sur le long terme, tout en préservant la lisibilité et l'intelligibilité des fichiers dans le temps.

1.2.3. Conception d'une grille d'analyse pour chaque type de service

Au cours de la troisième étape, une grille d'analyse a été conçue pour chacun des 9 types de service (annexe 3). L'objectif était de collecter des informations sur :

- L'identité du service (nom, date de création, adresse de contact…) ;
- Sa gestion (structure d'appartenance, tutelles, modèle économique, ressources humaines…) ;
- Ses caractéristiques fonctionnelles (conformité avec des standards et autres aspects techniques spécifiques à chaque type de service) ;
- Son utilisation (discipline et public cible, conditions d'accès, fréquence d'utilisation…).

1.2.4. Recensement et analyse des services

La quatrième étape a été l'analyse à proprement parler des services identifiés. Pour chacun, une recherche d'informations en ligne a été menée (sur le site web du service, dans des articles, rapports et communications). Des entretiens ont ensuite été réalisés avec les fournisseurs des services, afin d'obtenir des informations plus détaillées. Les renseignements collectés ont finalement été enregistrés dans la grille d'analyse.

1.3. Phase 2 : la conception d'un répertoire en ligne, Cat OPIDoR

A l'issue de l'étude cartographique, il a été décidé de développer un répertoire en ligne, permettant de consulter les services recensés et d'en répertorier de nouveaux de manière collaborative.

1.3.1. Objectifs du répertoire

Répondre à un manque de visibilité des services

Le projet de répertoire est né du constat d'un manque de visibilité des services proposés aux chercheurs pour gérer et partager leurs données. Dans un cadre d'incitation national et international à l'ouverture de la science, le rôle du segment BSN10 était de faire connaître les outils pensés et développés sur le territoire national, au sein des institutions de recherche. Aussi le groupe a-t-il commandité la conception d'un catalogue dédié.

Anticiper l'évolution du paysage des services

L'étude cartographique a permis de constater le caractère émergent et évolutif des services de données. Nombre d'entre eux étaient encore en construction en 2016. D'autres étaient à l'état de projet. L'Université de Lorraine[187], celle de Lyon 3[188] ainsi que Sciences Po[189] étudiaient à l'époque l'opportunité de mettre en place un service de gestion des données au sein de leur établissement. Le paysage des services de données semblait donc amené à se développer au cours des prochaines années. L'idée d'un répertoire, qui puisse être alimenté au fur et à mesure que de nouveaux projets voyaient le jour, se voulait adaptée à ce paysage en mouvement des services de données.

1.3.2. Caractéristiques du répertoire

Le répertoire a été conçu en partenariat avec l'Inist-CNRS[190], qui héberge et modère aujourd'hui l'outil. Celui-ci a pris le nom de Cat OPIDoR, pour « Catalogue pour une Optimisation du Partage et de l'Interopérabilité des Données de Recherche ». Lancé en septembre 2017, Cat OPIDoR[191] affiche l'objectif de répertorier les services français dédiés aux données scientifiques (de leur acquisition à leur archivage, en passant par leur diffusion).

187 https://www.univ-lorraine.fr/

188 https://www.univ-lyon3.fr/

189 http://www.sciencespo.fr/

190 Institut de l'Information Scientifique et Technique (https://www.inist.fr/)

191 https://cat.opidor.fr

Il fait aujourd'hui partie d'un ensemble de services proposés par l'Inist-CNRS : le portail OPIDoR[192].

Cat OPIDoR a été conçu pour un public multiple.

- Il est destiné aux communautés de recherche, afin que celles-ci puissent identifier facilement les services dont elles ont besoin dans le cadre de leurs projets de recherche.
- Il cible aussi les services d'appui à la recherche, qui peuvent l'utiliser lorsqu'ils sont amenés à guider les chercheurs dans la gestion de leurs données.
- Il est également destiné aux décideurs nationaux, leur permettant de dresser un état des lieux des services de données, d'identifier les services manquants ou de financer de nouveaux services.

L'élaboration du répertoire a suivi trois grandes phases :

1) Première grande phase : l'énumération des cas d'utilisation pouvant être faits du répertoire. Autrement dit, quelles recherches souhaiteront faire les utilisateurs dans Cat OPIDoR ? De quelles informations auront besoin les utilisateurs ? Les cas d'utilisation ont permis d'identifier comment décrire les services dans le catalogue.

2) Deuxième grande phase : l'élaboration du modèle de données.
Le modèle de données est fondé sur l'articulation entre les services et leur structure d'appartenance. Le terme « structure d'appartenance » désigne l'infrastructure ou le département qui opère ou héberge le service. Une structure peut fournir plusieurs services. Cette articulation entre services et structures permet de visualiser tous les services proposés par une structure et d'identifier des réseaux de relation entre les structures. Les champs descriptifs définis pour les services sont présentés dans l'annexe 7.
Le modèle de données a également été conditionné par le choix des modes de navigation dans le répertoire. Outre la recherche classique d'un service via le moteur de recherche et l'index

192 https://opidor.fr/

complet des services et structures, l'utilisateur peut utiliser les quatre autres modes de navigation suivants :

- Recherche par géolocalisation du service ;
- Recherche par type de service ;
- Recherche par discipline scientifique (la classification disciplinaire utilisée est inspirée de celle du European Research Council, qui s'articule en 3 grands domaines et 25 sous-domaines[193]) ;
- Par phase du cycle de vie des données : souvent cité dans les guides de bonnes pratiques de gestion des données[194], le cycle de vie (figure 5, p.107) rassemble les étapes par lesquelles sont censées passer des données ouvertes (planification, collecte, analyse, documentation, stockage, conservation, exposition, réutilisation).

A chaque service est ainsi associé un ensemble de champs descriptifs prédéfinis (annexe 7).

3) Troisième grande phase : le paramétrage de l'outil. La solution logicielle retenue est un wiki sémantique ouvert. Il s'agit de la solution Semantic Media Wiki (SMW)[195], une extension du logiciel Media Wiki, utilisé par Wikipedia. Cette extension permet de gérer des données structurées. L'objectif était par ailleurs de permettre une alimentation collaborative du répertoire. Toute personne, à la condition de se créer un compte utilisateur, a la possibilité de contribuer au contenu de Cat OPIDoR, soit en ajoutant un service, soit en modifiant un service déjà référencé.

193 https://cat.opidor.fr/index.php/Nomenclature_ERC

194 Le guide de l'IRSTEA ou celui de la UK Data Archive (en Grande-Bretagne) par exemple. IRSTEA (2017). *Guide pratique pour la gestion des données de recherche*. Version 2. https://donnees-recherche.irstea.fr/preambule-au-guide/ (consulté le 6 octobre 2019). Page 9
CORTI, L., VAN DEN EYNDEN, V., BISHOP, L. ET WOOLLARD, M. (2014). *Managing and Sharing Research Data: a Guide to Good Practice*. Londres : SAGE Publications Ltd. Chapitre 2

195 https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki

2. Le paysage national des services de gestion et d'ouverture des données : constats

Les résultats présentés ici doivent être considérés comme un aperçu du paysage français des services de données. Ils ne représentent pas ce paysage dans sa totalité, mais seulement l'échantillon des services qui ont pu être répertoriés entre novembre 2015 et avril 2016, au cours de l'étude cartographique pour BSN10.

L'analyse qui suit tient compte uniquement des services de gestion et d'ouverture des données. Elle exclut les services d'acquisition et d'analyse de données, qui ont été répertoriés dans Cat OPIDoR sous les catégories « plateformes d'acquisition » et « plateformes de calcul ». L'objectif est de révéler les dispositifs qui sont mis à la disposition des communautés de recherche pour conserver et diffuser les données scientifiques, conformément aux exigences politiques[196]. Si l'on ne prend en compte que les services de gestion et d'ouverture des données, qui ont été identifiés et analysés entre novembre 2015 et avril 2016, on en répertorie 44 au total[197]. Ces 44 services sont fournis par 34 structures différentes. Une structure peut proposer plusieurs services. C'est le cas notamment de l'Inist-CNRS, qui a développé 5 services ayant trait à la gestion et l'ouverture des données. La majorité des structures (27 d'entre elles) proposent un seul service ; 6 autres structures couplent deux services.

Il convient de noter que les données d'analyse ne sont pas uniformes pour tous les services. Pour certains, les informations recueillies sont lacunaires. Cela s'explique de deux manières.

- Première raison : le mode de recensement. Les 44 services n'ont pas été analysés selon la même méthode. Durant les premiers mois du recensement, l'analyse reposait sur des entretiens menés avec les fournisseurs des services identifiés. Les informations obtenues étaient alors plus détaillées et plus riches que les seuls renseignements relevés sur le web. Par la suite, lorsque le recensement a été couplé à l'élaboration de Cat OPIDoR, les informations collectées ont été réduites aux champs descriptifs

196 Cf. supra, deuxième partie, p.47

197 Le tableau analytique de ces 44 services est présenté dans l'annexe 8.

utilisés dans le catalogue, dans un souci d'efficacité et de rapidité. L'analyse ne passait alors plus par la conduite d'entretiens ; seules les informations disponibles en ligne étaient relevées.

- Seconde raison à l'inégalité des données d'analyse : le niveau de renseignement de l'interlocuteur et/ou le degré de précision des informations disponibles en ligne. Parfois l'information recherchée n'a tout simplement pas pu être trouvée, soit parce qu'elle ne figurait pas sur le site web du service, soit parce qu'elle n'était pas connue de l'interlocuteur avec qui l'entretien était mené.

Pour chaque résultat quantitatif, il sera donc précisé le nombre de services pour lesquels l'information est disponible.

2.1. Des services relativement récents

Les services étudiés dessinent un paysage émergent et évolutif. Les chiffres avancés ici se fondent sur un ensemble de 27 services, dont la date de création nous était connue.

- Leur date moyenne de création est 2011.
- 18 d'entre eux ont été créés entre 2013 et 2016, dont 4 en 2016 (l'outil de gestion de données DMP OPIDoR[198] et les annuaires IrsteaData, ArchiPolis Catalogue et ECOSCOPE).
- 7 étaient en cours de création au moment de l'étude en 2016.

Le service le plus ancien est celui du Centre de Données astronomique de Strasbourg (CDS), créé en 1972, d'abord sous forme de catalogues imprimés, puis sous forme de collections numériques à partir des années 1990.

Une majeure partie des services étudiés ont été créés en réponse au contexte politique d'ouverture des données de la recherche. C'est notamment le cas des 9 services de type accompagnement, formation, information et outil de gestion de données. Or, en France, le mouvement d'ouverture des données de recherche reste relativement récent. Cela expliquerait la nouveauté des services dédiés à la gestion et au partage des données.

198 Les services et structures cités dans cette partie et les suivantes sont répertoriés, avec leur URL, dans la sitographie p.293-307.

2.2. Des services extrêmement divers

Des études sur le paysage international des entrepôts de données (Kindling et al. 2017 ; Marcial et Hemminger 2010) font état d'un paysage hétérogène. Un constat similaire peut être fait à l'échelle de la France, pour les entrepôts de données comme pour les autres types de services (Rebouillat 2017).

Les 44 services répertoriés se différencient en effet :

- Par leur nature (il existe différents types de services) ;
- Par le périmètre de leur public cible ;
- Par leur degré d'utilisation (certains sont davantage utilisés que d'autres, bien qu'il soit difficile de le mesurer objectivement).

2.2.1. Divers par leur nature

Les services étudiés ont tout d'abord des fonctions différentes. Certains ont vocation à conserver et à diffuser les données ; d'autres à former et accompagner à la gestion et l'ouverture ; d'autres encore à informer sur la thématique des données de recherche. Pour les étudier, il a fallu les regrouper en différentes catégories. Une typologie a ainsi été créée dans le cadre de l'étude cartographique (elle reste une proposition parmi d'autres).

Les 44 services étudiés ici relèvent de 7 types distincts (figure 6) :

- 3 sont des services de type information ;
- 1 est un service de formation ;
- 3 sont des services d'accompagnement ;
- 2 sont des outils de gestion de données ;
- 25 sont des entrepôts de données ;
- 9 des annuaires de données ;
- 1 est une plateforme d'archivage.

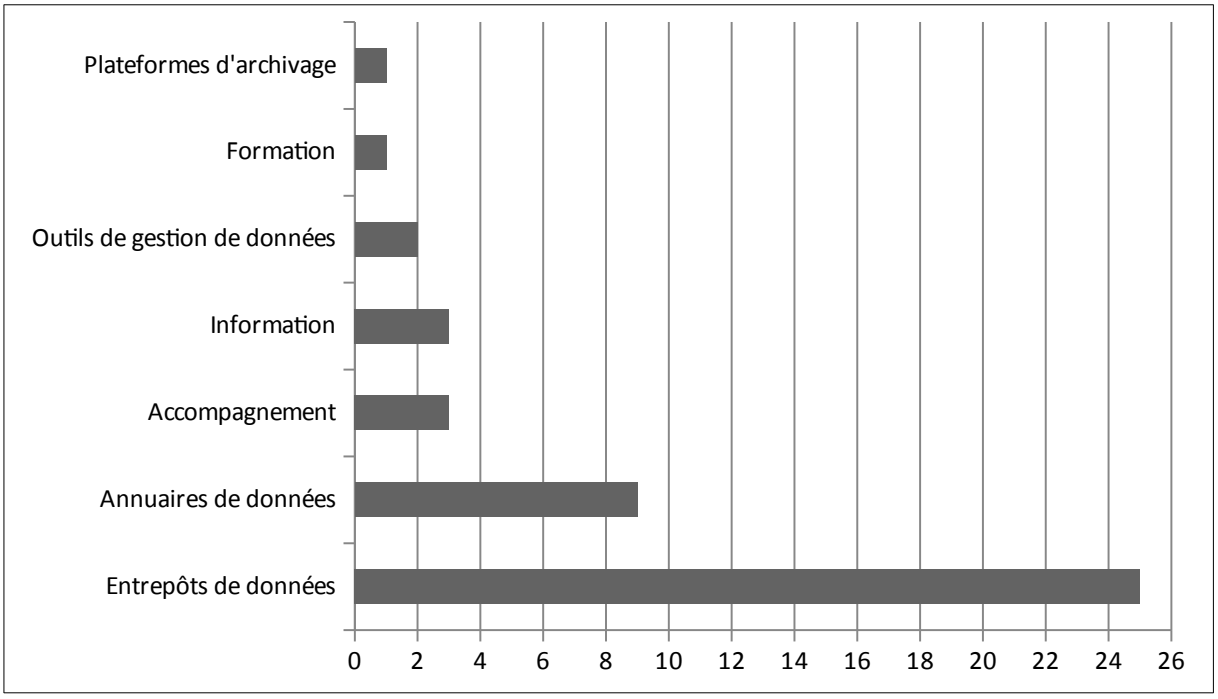


Figure 6 : Nombre de services analysés par type

Les services les plus nombreux sont les services de type entrepôts et annuaires de données. Ils sont nombreux, probablement parce qu'ils répondent aux spécificités d'une discipline ou d'un type de données. Les formats de fichiers et les modèles de métadonnées (c'est-à-dire la manière de décrire les données) varient en fonction des disciplines et de la nature des données. On trouve en effet des entrepôts dédiés à une thématique particulière ou à un projet de recherche particulier. Le RESIF Seismic Data Portal en est un exemple. Développé par le Réseau Sismologique et Géodésique Français (RESIF), il donne accès aux données sismologiques des réseaux permanents et mobiles des institutions de recherche françaises. Les données sont librement accessibles. Elles sont distribuées selon les standards et formats internationaux spécifiques à chaque type de données (notamment le s*tandard for the exchange of earthquake data*, dit SEED). L'annuaire du consortium ArchiPolis est un autre exemple de service dédié à une thématique très spécifique. Y sont inventoriées des enquêtes qualitatives en sciences sociales du politique. Ces enquêtes proviennent des 8 unités de recherche dont est composé le consortium, à savoir : le Centre de Données Socio-Politiques (CDSP), le Centre d'Études Européennes (CEE), le Centre de Recherches Politiques (CEVIPOF), l'Observatoire Sociologique du Changement (OSC), le Centre de Sociologie des

Organisations (CSO), le Centre d'Études et de Recherches Administratives, Politiques et Sociales (UMR CERAPS), le laboratoire Politiques publiques, Action politique, Territoires (UMR PACTE), et le laboratoire Action, discours, pensée politique et économique (UMR Triangle). L'objectif est de rendre ces enquêtes intelligibles grâce à une documentation et une mise en contexte conséquentes. La multiplicité des entrepôts et annuaires s'explique peut-être également par l'absence d'une solution nationale pour le dépôt et la diffusion des données scientifiques. A défaut d'entrepôt national, les acteurs de la recherche ont dû se positionner : soit en renvoyant vers des entrepôts existants (au niveau international notamment) ; soit en créant leur propre outil. D'où une variété d'initiatives dispersées.

2.2.2. Divers par le périmètre de leur public cible

Le paysage des services est également hétérogène en terme de public cible. Cela s'observe à trois niveaux :

- Au niveau de leur portée géographique ;
- Au niveau de leur portée disciplinaire ;
- Pour les entrepôts et les annuaires de données spécifiquement, au niveau de leurs modalités d'accès.

Portée géographique des services

Un service peut avoir une vocation institutionnelle, nationale ou internationale. Certains sont ouverts uniquement aux membres de l'établissement de recherche qui fournit le service ; d'autres sont ouverts à l'ensemble de la communauté scientifique française ; d'autres encore à la communauté élargie des chercheurs français et étrangers.

Parmi les 44 services étudiés, la plupart ont une portée institutionnelle ou nationale (36 d'entre eux).

- Les services à vocation institutionnelle sont au nombre de 20 ;
- Les services à vocation nationale au nombre de 16.

Les services à vocation internationale sont plus rares. Dans le présent échantillon, ils sont au nombre de 8 :

- Le CDS Portal, qui donne accès aux produits élaborés par le Centre de Données astronomiques de Strasbourg (CDS). Il fonctionne comme un méta-moteur permettant de rechercher dans les bases de données SIMBAD (la base de données de référence pour l'identification et la bibliographie des objets astronomiques hors système solaire), VizieR (la base de données répertoriant les grands relevés du ciel, les catalogues et les tables publiées dans les journaux scientifiques) et Aladin (un atlas interactif du ciel permettant d'accéder, de visualiser et d'analyser la collection d'images de référence du CDS, ainsi que les images disponibles dans les archives des observatoires spatiaux et au sol).
- SEANOE, qui est une solution de publication des données scientifiques marines. Développée par l'Institut Français de Recherche pour l'Exploitation de la Mer (Ifremer), elle est accessible à l'ensemble de la communauté scientifique du domaine. Les chercheurs peuvent y publier un jeu de données, en libre accès ou avec un embargo d'une période de deux ans.
- ArkeoGIS, qui est un Système d'Information Géographique (SIG) multilingue, initialement développé dans le but de mutualiser les données archéologiques et paléoenvironnementales de la vallée du Rhin. Il permet aujourd'hui de mettre en commun les données scientifiques spatialisées concernant le passé, depuis la Préhistoire jusqu'à nos jours. Les bases de données sont issues de travaux de chercheurs, de doctorants, d'étudiants en master, de sociétés privées et de services d'archéologie. En raison du caractère sensible des données, qui pourrait conduire à un pillage des gisements archéologiques, l'accès à l'outil est réservé aux professionnels de l'archéologie, issus d'institutions de recherche ou d'organisations à but non lucratif.
- EELS Data Base, qui est un entrepôt en libre accès de spectres d'excitation, géré par l'Institut des Matériaux Jean Rouxel (IMN). Les données sont issues d'expériences en spectroscopie des pertes d'énergie et en spectroscopie d'absorption des rayons X. L'entrepôt est ouvert au dépôt de nouvelles données.

- ESTHER DB, qui est une base de données recensant les données relatives à la famille des protéines à repliement de type alpha/beta hydrolase. Libre d'accès, elle cible des domaines de recherche comme l'agronomie et la chimie pharmaceutique. Elle a été conçue par deux unités de recherche, le laboratoire Architecture et Fonction des Macromolécules Biologiques (AFMB) et l'unité Dynamique Musculaire et Métabolisme (DMEM).
- VRP-REP, qui est un entrepôt de données dédié aux problèmes de tournées de véhicules. Hébergé à l'Université Catholique de l'Ouest, il est piloté par un comité international. Les utilisateurs peuvent consulter ou déposer des exemples de problèmes-types ainsi que leur solution. Le signalement de nouveaux cas se fait sur demande auprès du comité de pilotage de l'entrepôt.
- ILL Data Portal, qui est un entrepôt de l'Institut Laue-Langevin, permettant aux chercheurs internationaux, ayant utilisé les spectromètres de l'institut, de déposer les données acquises. L'institut est en effet consacré à l'étude des matériaux à partir de faisceaux neutroniques, touchant des domaines comme la biologie, la chimie, la physique nucléaire ou la science des matériaux.
- L'entrepôt du Centre de données de la physique des plasmas (CDDP), qui assure la conservation à long terme des données pertinentes à la physique des plasmas naturels dans le système solaire et les rend accessibles et exploitables à la communauté internationale. Les données archivées sont obtenues à bord des satellites et dans les observatoires terrestres. Le CDPP met à disposition des outils de visualisation, de manipulation et d'analyse des ensembles de données hétérogènes.

Portée disciplinaire des services

Les services de données se différencient également par le ou les champ(s) disciplinaire(s) qu'ils se proposent de couvrir. Dans le schéma ci-dessous (figure 7), ils ont été classés en quatre grandes catégories : les services relevant du domaine des sciences et technologies ; ceux relevant du domaine vie et santé ; ceux relevant du domaine des sciences humaines et sociales ; et les services multidisciplinaires, s'adressant de manière indifférenciée aux trois domaines précités.

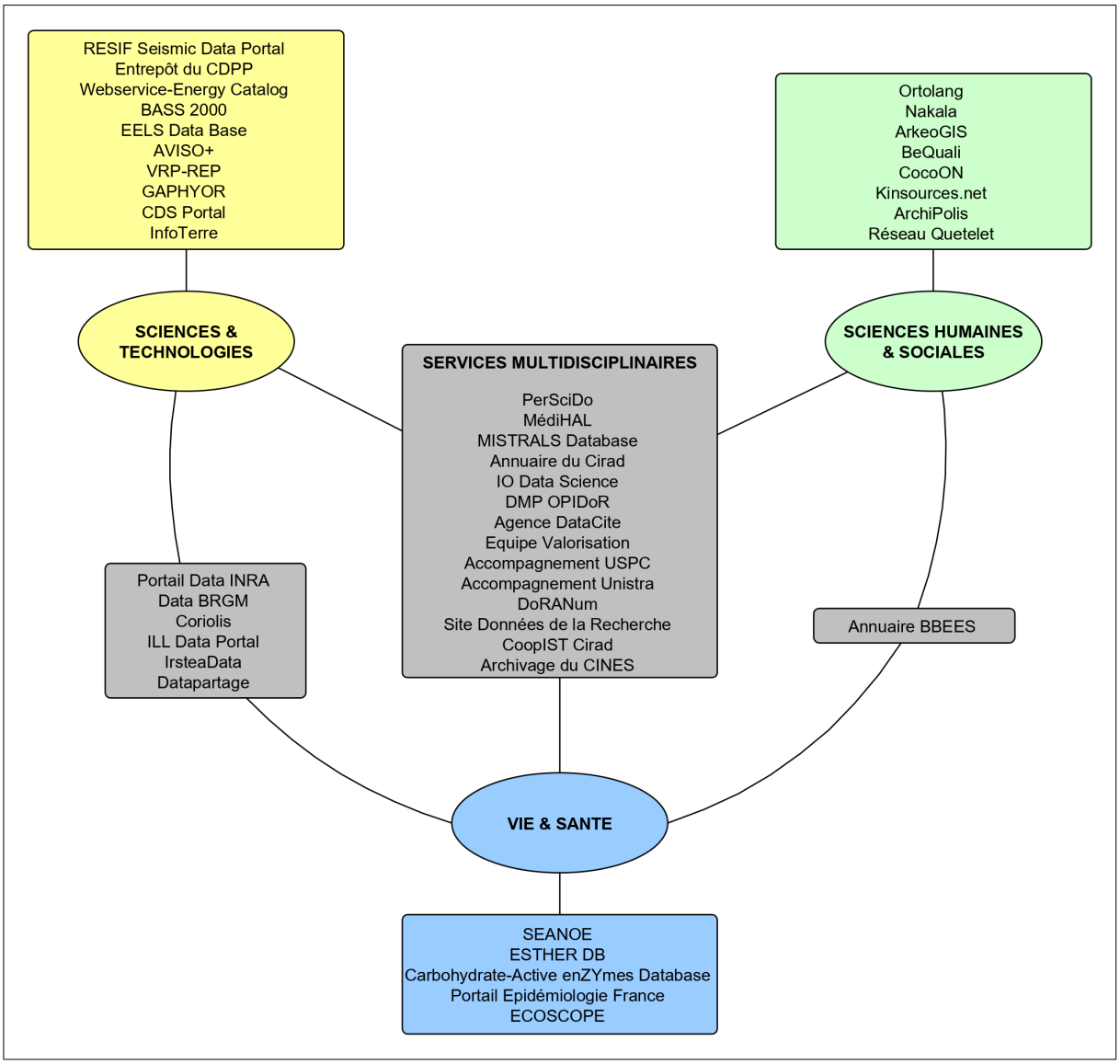


Figure 7 : Répartition des services par domaine scientifique

En termes de répartition, on dénombre 10 services spécifiques aux sciences et technologies, auxquels s'ajoute un nombre relativement important de services à la frontière avec le domaine vie et santé (6 services). Dans le champ vie et santé, les services répertoriés sont moins nombreux qu'en sciences et technologies : 5 au total. En sciences humaines et sociales, on en dénombre 8. Les services multidisciplinaires représentent la plus grande proportion : 14 services sur 44. Parmi eux, figure l'ensemble des services de type information, formation, accompagnement, outil de gestion de données et plateforme d'archivage. En revanche, les

entrepôts généralistes (c'est-à-dire couvrant tous les domaines disciplinaires) sont minoritaires par rapport aux entrepôts disciplinaires : ils représentent 3 des 25 entrepôts répertoriés. Cette proportion tend vers celle mise en évidence par Kindling et al. (2017) : dans une étude des entrepôts répertoriés par Re3data, les auteurs identifiaient 5,8% d'entrepôts généralistes sur un total de 1 381. Ce résultat conforte l'hypothèse émise plus haut[199], selon laquelle les entrepôts de données répondent à des spécificités disciplinaires.

Modalités d'accès aux entrepôts et annuaires de données

Pour ce qui est des entrepôts et des annuaires de données, ceux-ci proposent plusieurs niveaux d'ouverture. Il est d'ailleurs complexe de le déterminer, car un entrepôt ou un annuaire peut être ouvert ou fermé à différents niveaux :

- L'interface de consultation de l'entrepôt ou de l'annuaire peut en elle-même être plus ou moins librement accessible (1) ;
- Le dépôt ou le signalement de données peut être plus ou moins ouvert à tous (2) ;
- Enfin, pour les entrepôts, un troisième niveau d'ouverture intervient : celui de l'accès aux fichiers de données (3). Autrement dit, qui peut accéder aux fichiers de données, voire les télécharger et les réutiliser ?

Parmi les 34 entrepôts et annuaires étudiés, aucun n'est totalement ouvert (1, 2, 3). Il faut a minima demander la création d'un compte utilisateur pour déposer ou signaler des données.

Une majorité d'annuaires et d'entrepôts (28 sur 34) sont en accès libre, c'est-à-dire que tout internaute peut en parcourir le site web, y faire une recherche et lire les informations descriptives des différents jeux de données (1).

Les autres annuaires et entrepôts (6 sur 34) sont en accès restreint.

- Les entrepôts ArkeoGIS et ILL Data Portal sont accessibles après création d'un compte utilisateur.

199 Cf. supra, 2.2.1, p.116

- Les entrepôts Data BRGM, GAPHYOR et l'annuaire du Cirad sont consultables uniquement par les membres de la structure qui les a conçus. Pour Data BRGM, il s'agit des personnels, en particulier des scientifiques, affiliés au Bureau de Recherches Géologiques et Minières (BRGM) ; pour GAPHYOR, des membres du Laboratoire de Physique des Gaz et des Plasmas (LPGP) ; pour l'annuaire du Cirad, des membres du Centre de Coopération Internationale en Recherche Agronomique et pour le Développement (Cirad).
- Nakala, l'entrepôt de données de l'infrastructure Huma-Num, est un cas un peu particulier. Il ne s'agit pas d'un outil de consultation des données, mais plutôt d'un entrepôt sécurisé permettant de déposer, décrire et conserver les données. Il est donc accessible uniquement aux déposants, qui eux-mêmes disposent chacun d'une interface de gestion privée. Les données peuvent tout de même être rendues visibles via d'autres outils d'Huma-Num, comme le portail Isidore et le service de création de CMS[200] Nakalona.

Pour ce qui est du signalement ou du dépôt des données (2), aucun des 34 entrepôts et annuaires ne propose d'enregistrement libre de contraintes.

- Il faut a minima créer un compte utilisateur (pour 6 des 34 entrepôts et annuaires), voire soumettre une demande aux administrateurs et attendre leur aval avant de pouvoir contribuer (pour 6 autres d'entre eux).
- Dans d'autres cas (13/34), le dépôt ou le signalement dans l'entrepôt ou l'annuaire est réservé aux membres d'un laboratoire de recherche, d'un établissement de recherche ou d'un réseau thématique[201].
- Un troisième cas de figure existe, où l'entrepôt ou annuaire n'est pas voué à une alimentation collaborative de son contenu (9/34). Celle-ci est réalisée par une équipe

200 Content Management System (CMS)

201 Par réseau thématique, on entend le regroupement de plusieurs laboratoires de recherche autour d'un thème commun. Ont notamment été rassemblés sous cette désignation l'EquipEx RESIF (https://www.resif.fr/), le LabEx PERSYVAL-Lab (https://persyval-lab.org/), le programme MISTRALS (http://www.mistrals-home.org/), le consortium ArchiPolis (https://archipolis.hypotheses.org/) et les observatoires de la biodiversité représentés par le pôle de données ECOSCOPE (http://ecoscope.fondationbiodiversite.fr/ecoscope-portal/).

dédiée. L'entrepôt ou l'annuaire se présente donc plutôt comme une base de données, dont l'objectif est de fournir des données dans une perspective de réutilisation. C'est le cas du CDS Portal, qui est géré par une équipe de 30 personnes, chargée de recenser, structurer et diffuser les données sur les objets du ciel. C'est également le cas des données diffusées par le réseau Quetelet. Le réseau Quetelet permet aux chercheurs français et étrangers de commander des données extraites : de grandes enquêtes, recensements et autres bases de données issues de la statistique publique française ; de grandes enquêtes françaises provenant de la recherche ; et de grandes enquêtes internationales. Ces enquêtes sont mises à disposition par quatre unités partenaires du réseau Quetelet : l'ADISP du Centre Maurice Halbwachs ; le CDSP de Sciences Po ; le service des enquêtes de l'INED ; le CASD.

Enfin, pour les entrepôts, il existe différents niveaux d'accès aux fichiers de données (3).

- La majorité des entrepôts étudiés (10 sur 25) proposent un accès libre aux fichiers de données (éventuellement après une période d'embargo, pendant laquelle le fichier n'est pas accessible aux utilisateurs). Il s'agit d'entrepôts, dans lesquels le déposant ou le fournisseur de service choisit sciemment de mettre ses données à disposition. Pour ce profil d'entrepôt, le choix de l'ouverture est une politique clairement assumée.
- Une part importante également d'entrepôts (8/25) couple deux niveaux d'accès : certains fichiers de données sont en accès libre, d'autres sont en accès restreint. Cette double modalité d'accès englobe des catégories d'entrepôts très diverses. Il peut s'agir d'entrepôts institutionnels ou communautaires qui, tout en donnant à voir la production scientifique de l'établissement ou de la communauté, souhaite préserver la confidentialité de certaines données. Celles-ci sont alors réservées à un partage en interne ou à une utilisation exclusive par leur producteur. C'est le cas de Data Inra, de PerSCiDO et de l'ILL Data Portal. Il peut aussi s'agir d'entrepôts thématiques qui, en raison du type de données, se doivent de proposer des modalités d'accès restreint. C'est le cas de CoCoON et d'Ortolang, qui accueillent notamment des corpus oraux, pouvant contenir des données sensibles car personnelles.

- D'autres entrepôts fournissent un accès exclusivement restreint aux fichiers de données qu'ils hébergent. L'accès peut être réservé aux membres d'un établissement de recherche ou d'un réseau thématique (2/25). Ou bien il peut être soumis à la création d'un compte utilisateur, avec fourniture d'informations professionnelles et personnelles plus ou moins détaillées (4/25). L'accès aux enquêtes déposées dans l'entrepôt BeQuali se fait selon des modalités particulières, étant donné le caractère sensible des données. L'utilisateur intéressé par une enquête doit envoyer une demande d'accès via le réseau Quetelet. Celle-ci devra être argumentée et s'inscrire exclusivement dans un contexte de recherche ou d'enseignement. Elle sera alors soumise à validation. L'utilisateur aura ensuite accès à l'enquête pendant une durée de trois ans.
- Pour ce qui est de l'entrepôt Nakala, le mode d'accès aux fichiers de données fait figure d'exception. Les fichiers ne sont directement accessibles qu'à leurs producteurs (via l'interface de gestion personnelle de ces derniers).

Le degré d'ouverture des entrepôts et annuaires dépend par ailleurs des licences utilisées (droit d'auteur, licence Creative Commons[202]…). Certaines licences restreignent plus que d'autres la réutilisation des données et des métadonnées. Kindling et al. (2017) ont étudié cet aspect dans leur analyse du répertoire Re3data :

> *« Dans Re3data, les types de licences les plus fréquemment observés au niveau des entrepôts sont « autre » et « copyright » (cf. table 4). Cela signifie que les politiques et les avis de droit d'auteur rédigés par leurs propriétaires sont principalement utilisés pour réguler les conditions d'accès et d'utilisation des entrepôts et de leur contenu. »*[203]

Kindling et al. montrent qu'une majorité d'entrepôts offrent un accès libre à leur contenu et aux fichiers de données qui y sont liés. Pourtant ils constatent que 38,6 % des entrepôts

202 https://creativecommons.org

203 Traduction de : « The most frequent data license types according to re3data observed on the repository level are "other" and "copyright" (cf. Table 4). This means that self-written policy documents or copyright notices are mainly used to regulate access and usage conditions for the database and the content it provides. » (Kindling et al. 2017)

placent les données sous *copyright* (Kindling et al., tableau 4), ce qui contribue à entraver l'exploitation des données.

2.2.3. Divers dans leur utilisation

Il existe également une hétérogénéité des services en termes d'utilisation. Les scientifiques connaissent-ils l'existence des services de gestion et d'ouverture des données ? Les utilisent-ils ? Il a été difficile de répondre à ces questions. Des données quantitatives telles que le taux de fréquentation d'un site web étaient peu souvent disponibles en ligne au moment du relevé (au premier semestre 2016). Par ailleurs, rares étaient les fournisseurs de services interrogés à avoir mis en place un système de mesure des statistiques d'usage de leurs services. Seuls ceux, dont le service était accessible après création d'un compte utilisateur, ont pu nous fournir des données chiffrées.

En réalité, il n'est pas facile de comparer les services en fonction de leur utilisation. Pour cela, il conviendrait de prendre en compte plusieurs paramètres, tels que le caractère unique ou non du service (existe-t-il d'autres services qui lui font concurrence ?), la taille de la communauté qu'il cible (c'est-à-dire le nombre de ses utilisateurs potentiels)... Les données d'utilisation qu'il conviendrait de relever varient également en fonction du type de service. Pour les entrepôts et les annuaires, par exemple, il faudrait mesurer le nombre de données déposées ou signalées, ainsi que le nombre de consultations ou de téléchargements de chaque jeu de données. Or, comme le soulignent Marcial et Hemminger (2010) dans une étude portant sur 100 entrepôts de données scientifiques, il est rare de pouvoir relever ces informations de façon exacte et homogène, y compris via des entretiens approfondis avec les fournisseurs de services.

En termes d'utilisation, il n'est donc possible d'esquisser que quelques tendances partielles. Le nombre de jeux de données signalés et déposés dans les entrepôts et annuaires[204] a notamment pu être mesuré, de manière plus ou moins systématique. Sur 8 entrepôts dont

204 Les données, sur lesquelles se basent les calculs ci-dessous, ont été relevées au cours du premier semestre 2016.

l'information est connue, le nombre de jeux de données déposés est de 6 900 en moyenne. Sur 6 annuaires dont l'information est connue, le nombre moyen de jeux de données signalés est de 488. Leur volumétrie est néanmoins extrêmement variable : un entrepôt ou un annuaire peut contenir 5 jeux de données comme il peut en contenir 30 000 (tableaux 2 et 3). Il y a plusieurs facteurs à cela :

- Certains services sont récents ; d'autres plus anciens. Il est normal qu'un service nouveau contienne moins de jeux de données qu'un service qui existe depuis dix ans.
- La volumétrie d'un entrepôt ou d'un annuaire dépend aussi de son périmètre. Si l'entrepôt ou l'annuaire est multidisciplinaire, tous les chercheurs de toutes les disciplines vont pouvoir y signaler des données. Il contiendra donc peut-être un grand nombre de jeux de données. A l'inverse, dans un entrepôt disciplinaire, spécifique aux sciences du langage par exemple (comme Ortolang), il est possible que le nombre de jeux de données soit plus restreint. Par ailleurs, s'il existe deux entrepôts à vocation similaire, l'un au niveau national, l'autre au niveau international, ils se feront probablement concurrence. Parmi les entrepôts internationaux les plus connus, figurent Pangaea[205], en sciences de la Terre, Dryad[206], en biologie, GenBank[207], en génomique, Zenodo[208] et Figshare[209], des entrepôts sans étiquette disciplinaire.
- Autre facteur de variation de la volumétrie : la nature des données. Certains entrepôts contiennent peu de jeux de données, car ceux-ci demandent un travail de documentation très long. Dans le cas de BeQuali (5 jeux de données), les données déposées sont des enquêtes qualitatives. Une enquête est un matériau volumineux, qui demande un travail de recontextualisation très riche et très long à établir, pour pouvoir être réutilisé. La mesure de la volumétrie dépend également de ce que l'on définit comme jeu de données. Ici l'expression « jeu de données » a été entendue dans le sens d'« entité formant une unité de sens ». Cette définition englobe néanmoins une grande variété de données, allant d'une base de données à une image unique. Il est plus long

205 https://www.pangaea.de/

206 https://datadryad.org/

207 https://www.ncbi.nlm.nih.gov/genbank/

208 https://zenodo.org/

209 https://figshare.com/

de constituer une base de données que d'acquérir une image. Les enquêtes qualitatives diffusées par BeQuali requièrent par exemple un délai de préparation de six à douze mois. Il est donc normal qu'un entrepôt qui collecte des images en contienne 22 000 (MédiHAL), alors qu'un entrepôt qui collecte des bases de données n'en contienne que 5 (BeQuali). La volumétrie des entrepôts s'avère, par conséquent, difficile à mesurer.

Nom de l'entrepôt	Domaine disciplinaire cible	Nombre de jeux de données (au 1er semestre 2016)
SEANOE	Vie & Santé	200
Ortolang	Sciences humaines et sociales	58
Nakala	Sciences humaines et sociales	113
ArkeoGIS	Sciences humaines et sociales	65
Portail Data Inra	Sciences & Technologies ; Vie & Santé	NC (en cours de création au moment de l'étude)
PerSCiDO	Sciences Humaines & Sociales ; Sciences & Technologies ; Vie & Santé	NC (en cours de création au moment de l'étude)
BeQuali	Sciences Humaines & Sociales	5
MédiHAL	Sciences Humaines & Sociales ; Sciences & Technologies ; Vie & Santé	22 476
RESIF Seismic data portal	Sciences & Technologies	NC (non relevé)
CDPP: Entrepôt de données	Sciences & Technologies	NC (non relevé)
DATA BRGM	Sciences & Technologies ; Vie & Santé	NC (non relevé)
Webservice-Energy Catalog	Sciences & Technologies	1699
CoCoON	Sciences Humaines & Sociales	NC (non relevé)
BASS2000	Sciences & Technologies	NC (non relevé)
EELS Data Base	Sciences & Technologies	NC (non relevé)
Kinsources.net	Sciences Humaines & Sociales	NC (non relevé)
AVISO+	Sciences & Technologies	NC (non relevé)
VRP-REP	Sciences & Technologies	NC (non relevé)
MISTRALS database	Sciences Humaines & Sociales ; Sciences & Technologies ; Vie & Santé	NC (non relevé)
Coriolis	Sciences & Technologies ; Vie & Santé	NC (non relevé)
ESTHER database	Vie & Santé	NC (non relevé)
GAPHYOR	Sciences & Technologies	NC (non relevé)
CDS Portal	Sciences & Technologies	30587
ILL Data Portal	Sciences & Technologies ; Vie & Santé	NC (non relevé)
Carbohydrate-Active enZYmes Database	Vie & Santé	NC (non relevé)

Tableau 2 : Volumétrie des entrepôts de données en 2016

Nom de l'annuaire	Domaine disciplinaire cible	Nombre de jeux de données (au 1er semestre 2016)
BBEES : Annuaire de données	Vie & Santé ; Sciences Humaines & Sociales	180
Portail Epidémiologie France	Vie & Santé	901
IrsteaData	Sciences & Technologies ; Vie & Santé	NC (en cours de création au moment de l'étude)
Cirad : Annuaire de données	Sciences & Technologies ; Vie & Santé ; Sciences Humaines et Sociales	280
InfoTerre	Sciences & Technologies	NC (non relevé)
IO Data Science	Sciences & Technologies ; Vie & Santé ; Sciences Humaines & Sociales	NC (non relevé)
ECOSCOPE : Annuaire de données	Vie & Santé	70
ArchiPolis	Sciences Humaines & Sociales	222
Réseau Quetelet	Sciences Humaines & Sociales	1274

Tableau 3 : Volumétrie des annuaires de données en 2016

Le paysage national des services de gestion et d'ouverture des données scientifiques se caractérise donc par son hétérogénéité. Variété et diversité prévalent, tant en termes de typologie que de public cible et d'utilisation. Cette hétérogénéité reflète la proximité des services avec les communautés et les établissements de recherche, mais peut-être aussi l'absence de coordination au niveau national.

2.3. Organisation des services

Là encore c'est une grande diversité qui prévaut. Le mode d'organisation peut être très différent d'un service à l'autre. Les services se différencient :

- Par la nature de leur structure d'appartenance ;
- Par le profil professionnel des personnels fournissant le service (professionnels de l'information scientifique et technique, informaticiens, personnels de recherche…) ;
- Par leurs sources de financement.

2.3.1. Structures d'appartenance

Des structures très diverses peuvent être à l'origine de services de données. Trois grands groupes se dégagent.

- Premier groupe : les services gérés par des laboratoires de recherche. Il peut s'agir d'un unique laboratoire ou de plusieurs laboratoires associés, parfois avec une dimension internationale (ILL Data Portal ; VRP-REP). Ortolang, par exemple, est un entrepôt porté par le laboratoire Analyse et Traitement Informatique de la Langue Française (ATILF). PerSciDO est un entrepôt conçu par plusieurs laboratoires grenoblois, réunis autour du LabEx PERSYVAL-Lab. Quant à l'ILL Data Portal, il associe des laboratoires français, allemands et britanniques.
- Deuxième groupe : les services émanant des départements transversaux d'universités ou d'organismes de recherche. Ce sont souvent les départements Informatique, Systèmes d'information, Ingénierie documentaire ou Information scientifique et technique qui gèrent ces services. On peut citer l'exemple du Cirad, celui de l'IFREMER et celui de l'Université de Strasbourg.
- Troisième groupe : les services mis en place par des unités de services. Par « unité de service », il faut entendre à la fois les Unités Mixtes de Service (UMS) et les Unités de Service et de Recherche (USR), mais aussi les structures ayant une vocation de service, comme le CINES qui est un établissement public à caractère administratif (EPA), dont la mission est d'offrir à la communauté scientifique des ressources informatiques pour le calcul, l'hébergement et l'archivage. Parmi les unités de service figurent : l'Unité de Service et de Recherche PROGEDO, en charge du réseau Quetelet ; le Centre de Données Socio-Politiques (CDSP), responsable de l'entrepôt BeQuali et de l'annuaire ArchiPolis Catalogue ; ou encore l'Unité Mixte de Service Bases de Données sur la Biodiversité, l'Écologie, l'Environnement et les Sociétés (BBEES), à l'origine de l'annuaire BBEES.

Sur 44 services :

- 18 sont fournis par des départements transversaux d'établissements de recherche, dont 2 par des universités et 16 par des organismes de recherche ;
- 15 sont gérés par des laboratoires de recherche ;
- 11 par des unités de services.

Aucun type de structure ne semble donc prédominer plus qu'un autre (figure 8).

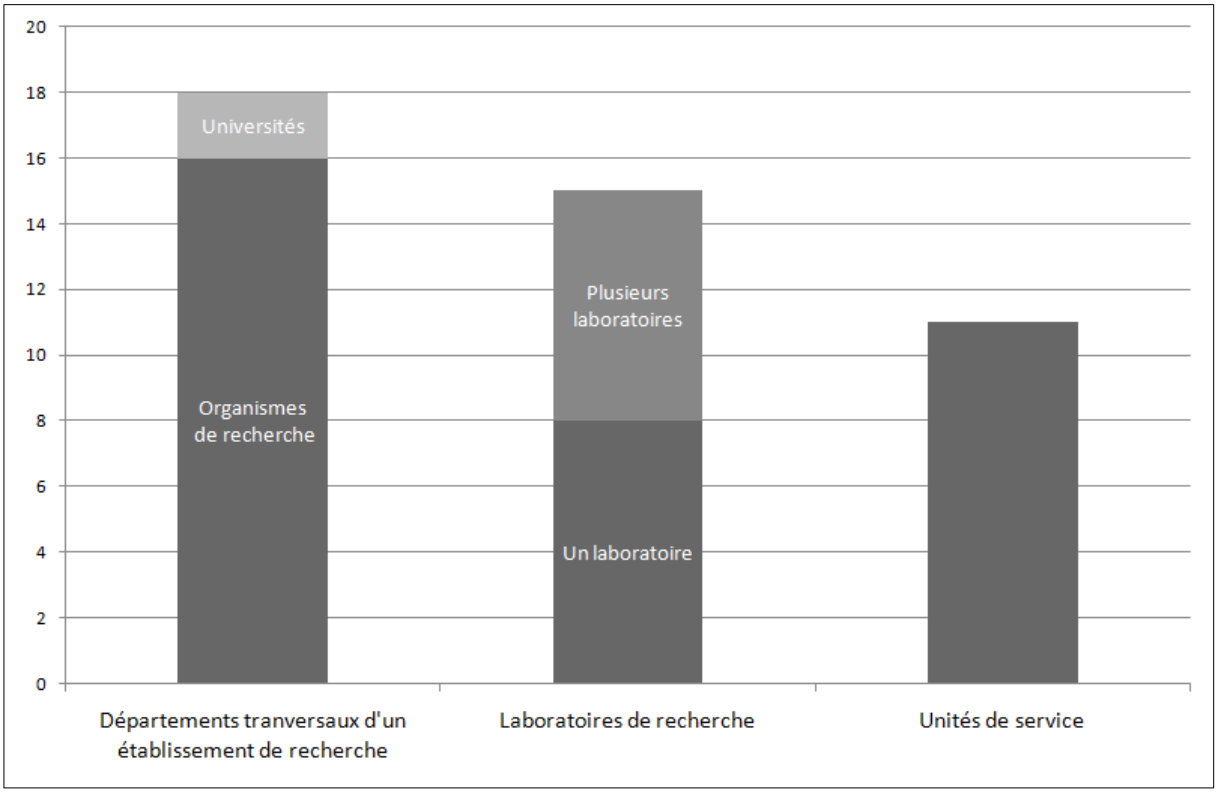


Figure 8 : Types de structures à l'origine d'un service

Une des 34 structures étudiées bénéficie du statut d'infrastructure de recherche (RESIF) ; trois autres du statut de très grande infrastructure de recherche (l'Institut Laue Langevin, Huma-Num et PROGEDO). Étant multinational, l'institut Laue-Langevin fait également partie depuis 2006 de la feuille de route des infrastructures de recherche européennes du *European Strategic Forum of Research Infrastructures* (ESFRI)[210]. Ce statut lui apporte à la fois des

210 European Strategy Forum on Research Infrastructures 2018, op. cit.

financements supplémentaires de l'État, mais aussi une visibilité plus grande au niveau national et international. Par ailleurs, les infrastructures de recherche sont des services auxquels tous les scientifiques de la recherche publique peuvent potentiellement avoir accès (dans la limite de leur champ disciplinaire).

A l'inverse, les services des laboratoires sont plus fragiles, parce qu'ils peuvent paraître moins légitimes aux yeux des instances décisionnelles et politiques (directions d'établissements, Ministère de la recherche…). Par exemple, ils ne bénéficient pas du statut de fournisseur de services dont disposent les départements transversaux des établissements de recherche et les unités de service. En revanche, ils sont souvent plus connus de la communauté scientifique dans laquelle ils s'ancrent, parce qu'ils émanent de ce milieu, sont gérés par des chercheurs et répondent à des besoins très ciblés [211].

2.3.2. Financement

Les services de gestion et d'ouverture des données disposent dans la quasi-totalité des cas d'un modèle économique fondé sur la gratuité pour l'utilisateur[212]. Leur financement repose donc sur un apport en amont. La nature de celui-ci diffère selon les services. Sur 29 services dont nous connaissons l'origine des subventions, 21 sont financés par le budget interne de leur structure d'appartenance. Les 8 autres sont financés de manière non pérenne, grâce à des programmes de type Investissement d'Avenir (EquipEx, LabEx, IdEx)[213] ou bien étaient encore, au moment de l'étude, à la recherche de financements.

Sur les 21 services qui ont un budget plus pérenne, on constate que 14 d'entre eux sont fournis par des organismes de recherche :

- Soit par des EPST[214] (Irstea, CNRS, INRA, Inserm) ;

211 Voir infra, cinquième partie, p.213

212 Les services faisant exception sont : l'agence DataCite (l'attribution de DOI est payante pour les structures – établissements de recherche ou services de données – qui en font la demande) ; la plateforme d'archivage du CINES ; le Webservice-Energy Catalog (le dépôt de données est payant pour toute personne externe au Centre Observation, Impacts, Energie).

213 http://www.enseignementsup-recherche.gouv.fr/pid24578/investissements-d-avenir.html

214 Etablissements Publics à caractère Scientifique et Technologique (EPST)

- Soit par des EPIC[215] (IFREMER, Cirad, BRGM).

Il semble donc n'y avoir aucune règle commune qui régisse le financement des services de données. Chaque service, en fonction du contexte dans lequel il s'insère, a son propre mode de financement. Il existe certes des dispositifs pour le financement des infrastructures de recherche, mais ils restent épars et relativement rares. Aussi le financement des services se caractérise-t-il par son hétérogénéité, tant dans sa nature que dans son montant et sa durée.

2.3.3. Ressources humaines

Les services étudiés mobilisent en moyenne 2 équivalents temps plein (ETP). Ce chiffre cache néanmoins de très grandes différences : l'équipe en charge du CDS Portal au Centre de Données astronomiques de Strasbourg se compose de 30 personnes[216], tandis que l'équipe d'accompagnement de l'Université Sorbonne Paris Cité, est constituée de 3 personnes se partageant la valeur d'un ETP.

En termes de profils professionnels, les entrepôts et annuaires de données impliquent systématiquement des ingénieurs informaticiens : 100 % des 17 entrepôts et annuaires, dont la nature des ressources humaines est connue, emploient des ingénieurs informaticiens (tableau 4). Ce type de services nécessite en effet des développements numériques complexes (notamment pendant la phase de création de l'outil). Ils mobilisent par ailleurs plus fréquemment des compétences scientifiques que les autres types de services. On note la présence de chercheurs dans 12 des 17 entrepôts et annuaires (70 %), pour lesquels la nature des ressources humaines a pu être relevée. De même, 9 d'entre eux font appel à des ingénieurs et techniciens de recherche (soit 53 % des 17 entrepôts et annuaires).

A l'inverse, les services de type formation, information, accompagnement, les outils de gestion de données et les plateformes d'archivage (répertoriés comme « autres types de services » dans le tableau 4) sont principalement conçus par des professionnels de l'information scientifique et technique (IST). Sur 8 services connus, tous sont gérés par des professionnels de l'IST (100 %). Aucun d'entre eux ne fait appel à des ingénieurs ou des techniciens de recherche et seulement 2 font intervenir des chercheurs (25 %).

215 Établissements Publics à caractère Industriel et Commercial (EPIC)

216 La correspondance en équivalent temps plein n'est pas connue.

De quelque type qu'ils soient, les services mobilisent parfois également les compétences de juristes et de personnels d'appui au montage de projets (12 % des 17 annuaires et entrepôts et 12 % également des 8 autres types de services). L'encadrement juridique et les exigences des financeurs en termes de gestion et d'ouverture des données peuvent être très variables, si bien qu'il est parfois fait appel aux compétences d'experts dans ces domaines.

	Entrepôts et annuaires de données (sur 17 connus)		**Autres types de services (sur 8 connus)**	
Chercheurs	12	70%	2	25%
Ingénieurs et techniciens de recherche	9	53%	0	0%
Ingénieurs développement	17	100%	3	37%
Professionnels de l'IST (dont archivistes)	9	53%	8	100%
Autres (juristes…)	2	12%	1	12%

Tableau 4 : Profil des professionnels gérant les services de données

2.4. Spécificités des services fournis par des professionnels de l'information scientifique et technique

Les professionnels de l'information scientifique et technique (IST) semblent occuper une place importante dans l'offre de service nationale en matière de gestion et d'ouverture des données de la recherche. 17 des 44 services étudiés ont en effet été initiés par des professionnels de l'IST (tableau 5). Ils couvrent notamment tous les services de type accompagnement, formation, information, outil de gestion de données et plateforme d'archivage.

Nom du service	Type de service	Structure d'appartenance
Portail Data Inra	Entrepôt de données	INRA
PerSCiDO	Entrepôt de données	PERSYVAL-Lab
BeQuali	Entrepôt de données	CDSP
MédiHAL	Entrepôt de données	CCSD
IrsteaData	Annuaire de données	Irstea
Cirad : Annuaire de données	Annuaire de données	Cirad
ArchiPolis	Annuaire de données	CDSP
DMP OPIDoR	Outil de gestion de données	Inist-CNRS
Agence DataCite	Outil de gestion de données	Inist-CNRS
Equipe Valorisation des données de la recherche	Accompagnement	Inist-CNRS
USPC : Accompagnement	Accompagnement	Université Sorbonne Paris Cité
Université de Strasbourg : Accompagnement	Accompagnement	Université de Strasbourg
DoRANum	Formation	Inist-CNRS
Site Données de la Recherche	Information	Inist-CNRS
DataPartage	Information	INRA
Cirad : Information	Information	Cirad
CINES : Plateforme d'archivage	Plateforme d'archivage	CINES

Tableau 5 : Services gérés par des professionnels de l'IST

2.4.1. A l'origine de l'offre de services des professionnels de l'IST : la défense du libre accès aux résultats de la recherche et le repositionnement de la profession

En France, les professionnels de l'IST se sont très vite positionnés sur le développement de services pour la gestion et la valorisation des données de recherche. La mission première de ces professionnels, qu'ils soient documentalistes ou personnels des bibliothèques, est de collecter l'information scientifique et de la diffuser, par le biais de médiations documentaires, à des communautés de recherche plus ou moins diversifiées. Les professionnels de l'IST

interviennent donc traditionnellement en amont du processus de recherche, fournissant aux chercheurs une information qui servira de matière première à leurs recherches. Or, en raison du développement des technologies, ces professionnels se trouvent soumis à une remise en cause régulière de leur fonction (Fabre et Gardiès 2008). C'est pourquoi ils ont tenté au cours de ces vingt dernières années de se positionner également sur la valorisation de l'information scientifique, en aval du processus de recherche. Directement concernés par l'augmentation du prix des abonnements aux revues scientifiques (Chartron 2010), ils ont notamment rallié le mouvement du libre accès aux résultats de la recherche et développé des archives ouvertes pour la diffusion des publications. Leur positionnement sur la question de la gestion et du partage des données de recherche s'est alors imposé comme une suite logique. Il en va de même des archivistes, dont la profession a évolué avec le traitement d'archives de la recherche essentiellement numériques et qui tentent de valoriser leur savoir-faire en matière de conservation des données scientifiques.

2.4.2. Nature de l'offre de services conçue par les professionnels de l'IST

Se proposant d'accompagner les chercheurs dans la documentation et la diffusion des données, les professionnels de l'IST ont développé des services de différents ordres, décrits dans plusieurs études sur le sujet (Cox et Pinfield 2014 ; Si et al. 2015 ; Tenopir et al. 2017).

L'une d'elle est le fruit d'une enquête menée en 2016 auprès des directeurs des bibliothèques membres de l'association LIBER[217] (Tenopir et al. 2017). L'enquête avait pour but d'identifier les différents types de services délivrés par les bibliothèques universitaires d'Europe en matière de données de recherche. Les résultats montrent que les bibliothèques offrent davantage des services d'information et de conseil (par exemple, aider l'utilisateur à trouver des informations sur les plans de gestion de données ou sur les standards de métadonnées) que des services techniques (comme créer des métadonnées pour un jeu de données ou préparer des données à verser dans un entrepôt). Le même constat avait été fait dans une précédente enquête sur les bibliothèques nord-américaines (Tenopir et al. 2012 ; Tenopir et al. 2015b). Pour Tenopir et al., les bibliothèques fournissent moins de services

217 Ligue des Bibliothèques Européennes de Recherche (https://libereurope.eu/)

techniques, car ceux-ci demandent des investissements non négligeables en termes de temps, de moyens et de nouvelles compétences à acquérir.

Parmi les 7 types de services identifiés au cours de l'étude cartographique, 3 sont d'ordre informatif (formation, information, accompagnement) et 4 d'ordre technique (outils de gestion de données, annuaires de données, entrepôts de données, plateformes d'archivage). Au vu des résultats, il semble que les professionnels de l'IST ne se soient pas cantonnés aux services informatifs, mais aient également proposés des services d'ordre technique. Des exemples sont présentés ci-dessous pour chacun des 7 types de service. Là encore les informations données datent de 2016. Les services ont pu évoluer, comme cela sera évoqué dans la partie suivante (2.5, p.144).

Des formations

Les professionnels de l'IST sont à l'origine de diverses formations dédiées à la gestion et l'ouverture des données de la recherche. Ces formations prennent souvent la forme de stages (l'Enssib[218] propose par exemple une formation de 2 à 4 jours, intitulée « Gérer, valoriser et préserver les données de la recherche ») ou de présentations plus courtes (organisées par exemple au moment de l'Open Access Week[219] chaque année). L'objectif est de sensibiliser les chercheurs et, pour certaines formations, d'apporter de nouvelles compétences aux professionnels de l'IST. Ces formations, en raison de leur caractère ponctuel, n'ont pas été répertoriées dans la cartographie des services de données. Les services de type fixe leur ont été préférés. C'est le cas de DoRANum dans le domaine de la formation. DoRANum, contraction de « Données de la Recherche : Apprentissage NUMérique à la gestion et au partage », était à l'état de projet en 2016. Le but était de mettre en place un dispositif de formation à distance, intégrant différentes ressources d'auto-formation sur la thématique de la gestion et du partage des données de recherche (« pourquoi et comment rédiger un plan de gestion des données », « comment et où déposer mes données », « comment décrire les données », « comment associer durablement des données à son auteur »…). Existantes ou créées dans le cadre du projet, ces ressources devaient proposer plusieurs parcours et modes

218 École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (https://www.enssib.fr/)

219 http://openaccessweek.org/

d'apprentissage, à destination des enseignants-chercheurs, des doctorants et des professionnels de l'information. Le projet était piloté par la Bibliothèque scientifique numérique. Il a été réalisé par deux structures de la sphère IST, l'Inist-CNRS et le réseau national des URFIST, avec la contribution de développeurs contractuels.

Des supports d'information et de veille

Des supports d'information ont également été élaborés par des professionnels de l'IST. Les plus connus du domaine sont les sites internet Données de la Recherche et DataPartage.

Le site Données de la Recherche est une plateforme d'information et de veille sur les données de la recherche. Conçu et alimenté par l'Inist-CNRS, il répertorie des actualités, évènements, publications, politiques de données et textes de référence sur le sujet.

Le site DataPartage a été développé par la délégation Information Scientifique et Technique de l'INRA. Il cible en premier lieu les chercheurs de l'institut, mais peut être consulté par tous. Il s'agit d'une plateforme d'information sur la gestion et le partage des données, qui recense les services, outils et bonnes pratiques recommandés par l'INRA.

Des modèles et outils de plan de gestion de données

Lorsque les financeurs de la recherche, notamment le programme Horizon 2020 de la Commission européenne, ont commencé à préconiser la rédaction de plans de gestion de données (DMP) pour les projets qu'ils financent, des professionnels de l'IST se sont mis à développer des modèles. A l'Université Sorbonne Paris Cité (USPC), une conservatrice des bibliothèques (Service Commun de la Documentation), une archiviste (Bureau des archives) et une ingénieure pour la valorisation des résultats scientifiques (Direction d'Appui à la Recherche et à l'Innovation) ont conçu un guide à l'intention des chercheurs, pour les aider à rédiger des plans de gestion de données[220]. Elles ont conçu un modèle de plan compatible avec

220 CARTIER, A., MOYSAN, M. ET REYMONET, N. (2015). *Réaliser un plan de gestion de données : guide de rédaction*. Version 1. https://www.fosteropenscience.eu/sites/default/files/pdf/2252.pdf (consulté le 20 septembre 2019).
Une deuxième version du guide a été proposée par la suite : REYMONET, N., MOYSAN, M., CARTIER, A. ET DÉLÉMONTEZ, R. (2018). *Réaliser un plan de gestion de données « FAIR » : modèle*. Version 2. https://archivesic.ccsd.cnrs.fr/sic_01690547 (consulté le 20 septembre 2019).

celui proposé par la Commission européenne. L'objectif était d'offrir une liste de champs applicables, mais également d'identifier au sein de l'USPC les différents acteurs susceptibles d'accompagner les chercheurs dans la rédaction de leur DMP. Le modèle de DMP proposé s'accompagne d'un workflow pointant, pour chaque section du plan de gestion de données, vers des personnes ressources.

Par la suite, l'Inist-CNRS a créé DMP OPIDoR, un outil d'aide à la rédaction de plans de gestion de données. Il a été développé à partir du code source de DMPonline[221], son homologue britannique. L'utilisateur a la possibilité de choisir le modèle de plan de gestion de données qui correspond à son institution d'affiliation ou au financeur de son projet de recherche. A défaut, il peut aussi utiliser le modèle proposé par la Commission européenne dans le cadre du programme de financement Horizon 2020. Chaque modèle est accompagné de recommandations, aidant l'utilisateur à répondre aux questions. Les institutions et communautés de recherche peuvent ajouter dans DMP OPIDoR leur modèle de plan de gestion de données, ainsi que leurs propres recommandations. Plusieurs établissements de recherche ont ainsi implémenté leur modèle, comme l'INRA[222] ou l'Université de Strasbourg[223].

Des équipes d'accompagnement

Plusieurs services d'information scientifique et technique ont constitué des équipes consacrées à l'accompagnement des chercheurs pour la gestion et l'ouverture de leurs données.

En 2012, l'Inist-CNRS a créé une équipe « Valorisation des données de la recherche ». Composée d'une dizaine de personnes, cette équipe propose aux chercheurs un accompagnement personnalisé, à distance ou en présentiel. La particularité de cette équipe est que ses membres possèdent une double compétence scientifique et documentaire, chacun venant d'horizons disciplinaires différents (chimie, biologie, médecine, sciences humaines et sociales…). L'équipe est notamment intervenue auprès de l'Observatoire Terre

221 https://dmponline.dcc.ac.uk/

222 https://dmp.opidor.fr/template_export/1988740616.pdf

223 https://dmp.opidor.fr/template_export/1640217907.pdf

Environnement de Lorraine (OTELo). Elle a accompagné les chercheurs de l'observatoire dans la mise en place de pratiques de gestion communes (rédaction de plans de gestion de données, utilisation de standards de métadonnées pour décrire les données, attribution d'identifiants pérennes permettant de citer les données…)[224]. L'objectif de l'observatoire était de créer une infrastructure de dépôt des données. L'équipe Valorisation a donc également contribué à la modélisation et à l'interopérabilité de l'outil.

Une initiative similaire existe à l'Université de Strasbourg, où le service des bibliothèques propose l'assistance d'un *data librarian* (ou bibliothécaire des données). Ce service d'accompagnement est conçu comme une interface entre les chercheurs, le service des bibliothèques et la direction informatique. Il propose notamment des actions de type : accompagnement à la création de bases de données et d'outils de visualisation des données en ligne ; accompagnement dans la réalisation de plans de gestion de données ; conseil sur les aspects juridiques liés aux données de la recherche ; conseil en matière de curation de données, de standards de métadonnées, de formats de fichiers ou de choix d'entrepôt de données en ligne.

Des entrepôts et annuaires de données

Des professionnels de l'IST sont également à l'initiative d'entrepôts ou d'annuaires de données. C'est le cas du Centre de Données Socio-Politique (CDSP), qui a conçu l'entrepôt BeQuali pour la diffusion des résultats d'enquêtes qualitatives en sciences sociales (Duchesne et Garcia 2014). Créé en 2012, l'entrepôt est géré par une équipe de 5 personnes, composée de professionnels de l'IST, d'archivistes et d'ingénieurs. Le développement informatique a été réalisé par des ingénieurs de l'EquipEx DIME-SHS, dont fait partie BeQuali. La description des enquêtes est disponible en libre accès. Les données elles-mêmes (guide d'entretien, transcription des entretiens, comptes-rendus…) sont accessibles sur demande justifiée, via le formulaire dédié du réseau Quetelet. Elles sont par ailleurs archivées au CINES. La numérisation, la mise en forme et la recontextualisation de l'enquête est un travail d'envergure, qui prend environ six mois. Début 2016, cinq enquêtes étaient disponibles dans BeQuali.

224 Cette collaboration a donné lieu à la rédaction d'un guide de bonnes pratiques, intitulé *Gestion et valorisation des données de recherche* (Arnould et Jacquemot 2016).

Il existe également des projets d'annuaires de données, auxquels participent des professionnels de l'IST, comme au Cirad. L'annuaire des données scientifiques du Cirad est un projet d'établissement, initié en 2015, qui associe la Délégation à l'Information Scientifique et Technique, la Délégation juridique et la Direction des systèmes d'information. Il a pour objectif de répertorier les données produites par les unités de recherche de l'institut. L'outil est accessible uniquement depuis l'intranet du Cirad. Les données sont décrites selon le standard Dublin Core[225]. A ces informations génériques s'ajoutent des métadonnées plus spécifiques sur la production scientifique et les nomenclatures du Cirad (coordonnées géographiques, formats de fichiers, mode de sauvegarde, niveau de confidentialité...). Lors de la première vague d'alimentation du portail, des correspondants ont été désignés dans chaque unité de recherche, afin de recueillir la liste des données qui y étaient produites. A terme, les pratiques d'inventaire seront amenées à se généraliser, avec la création d'un formulaire web qui permettra aux chercheurs de référencer directement leurs jeux de données sur le portail.

Une plateforme d'archivage

L'archivage des données de recherche revient normalement au service des archives de chaque établissement. Toutefois ces services proposent un archivage essentiellement physique des documents. L'archivage numérique est réalisé au niveau national par le Centre Informatique National de l'Enseignement Supérieur (CINES). La plateforme d'archivage du CINES a vocation à archiver les données et les documents numériques produits par la communauté française de l'Enseignement Supérieur et de la Recherche. Elle propose des solutions d'archivage numérique payantes, sur le moyen et le long terme, et offre à ses utilisateurs une expertise dans les domaines informatique et archivistique. Les projets à volumétrie importante sont néanmoins privilégiés, pour des raisons de rentabilité. La sécurité et l'intégrité des données sont garanties par un ensemble de procédures telles que l'attribution de métadonnées, le choix de formats de fichiers pérennes, la réplication des données et un environnement informatique protégé.

225 https://www.dublincore.org/

2.4.3. Quelles tendances : des services techniques ou informatifs ?

L'enquête de Tenopir et al. (2017) montre que les bibliothèques universitaires européennes fournissent davantage de services informatifs que de services techniques.

En 2016, en France, seules 3 bibliothèques universitaires proposaient des services pour les données de la recherche. Deux d'entre elles fournissaient des services d'information et de conseil – ce qui rejoindrait les conclusions de Tenopir et al. (toutes réserves gardées, étant donné le nombre très restreint de bibliothèques à fournir des services de données). Plusieurs autres bibliothèques, comme la bibliothèque de l'Université Lille 3, étudiaient l'opportunité de mettre en place une offre de service de ce type.

De manière plus générale, si l'on considère l'ensemble des structures spécialisées dans l'IST qui fournissent un des services répertoriés, soit 10 structures (3 bibliothèques universitaires, 4 pôles IST affiliés à des organismes de recherche, 2 unités mixtes de services et 1 établissement à caractère administratif), on constate que :

- 5 d'entre elles offrent des services techniques exclusivement ;
- 2 offrent exclusivement des services d'information et de conseil ;
- 3 offrent à la fois des services techniques et des services d'information.

Les structures de la sphère IST sont donc plus nombreuses à offrir des services techniques (tableau 6). En revanche, il est vrai que les professionnels de l'IST qui développent des services commencent par déployer des services de conseil et d'information. Les services techniques viennent dans un second temps (c'est le cas de l'Irstea, du Cirad, de l'Inist-CNRS).

Par ailleurs, lorsqu'elles développent des services techniques, ces structures le font généralement en partenariat, car elles n'ont pas les compétences informatiques nécessaires (pour le développement d'un entrepôt par exemple). Soit elles s'associent avec le service informatique de leur établissement, soit elles recrutent des ingénieurs développement contractuels le temps de la création de l'outil. La collaboration avec des services informatiques est un point évoqué dans l'enquête de Tenopir et al. (2017) :

> *« Presque toutes les bibliothèques (90,7%) qui ont répondu à la question oui/non, leur demandant si elles entretenaient des collaborations pour la fourniture de services de données de recherche, ont dit qu'elles collaboraient avec d'autres*

unités ou services à l'intérieur de leur institution. Les départements informatiques et les directions de la recherche ont été les partenaires les plus souvent cités ; les bibliothèques travaillent également en collaboration avec divers départements. Dans la catégorie « autres collaborateurs » ont été signalés les services des archives universitaires, les départements juridiques et les unités d'appui à la recherche. »[226]

Structures	Type	Services techniques	Services d'information
INRA	Organisme de recherche (Pôle IST)	Portail Data Inra	DataPartage
PERSYVAL-Lab	Université (Bibliothèque, associée à un laboratoire de recherche)	PerSCiDO	
CDSP	Unité mixte de services	BeQuali ; ArchiPolis	
CCSD	Unité mixte de services	MédiHAL	
Irstea	Organisme de recherche (Pôle IST)	IrsteaData	
Cirad	Organisme de recherche (Pôle IST)	Cirad : Annuaire de données	Cirad : Information
Inist-CNRS	Organisme de recherche (Pôle IST)	DMP OPIDoR ; Agence DataCite	Equipe Valorisation des données de la recherche ; DoRANum ; Site Données de la Recherche
CINES	Etablissement public à caractère administratif (pôle archivage)	CINES : Plateforme d'archivage	
Université Sorbonne Paris Cité	Université (Bibliothèque, associée au service des archives et à la direction de la recherche)		USPC : Accompagnement
Université de Strasbourg	Université (Bibliothèque)		Université de Strasbourg : Accompagnement

Tableau 6 : Nature de l'offre de service des 10 structures IST répertoriées

226 Traduction de : « Almost all (90.7%) libraries who answered a yes/no question on whether they collaborate say they collaborate with other units or offices within their institutions regarding RDS. The IT Center and Office of Research are the most frequent collaborators; libraries also collaborate with various subject departments. "Other" collaborators include university archives, legal offices, and research support units. » (Tenopir et al. 2017, p.35)

2.5. Evolution des services de gestion et d'ouverture entre 2016 et 2019

Parmi les 44 services analysés en 2016, deux n'existent plus à ce jour : l'entrepôt GAPHYOR et le site d'information Données de la Recherche. Le premier n'est plus accessible en ligne depuis 2018. Quant au second, il a été transposé au niveau du Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, avec le lancement en 2018 du Plan national pour la Science Ouverte[227]. Les actualités et références en matière de gestion et de partage des données de recherche sont désormais accessibles depuis le site www.ouvrirlascience.fr.

2.5.1. Evolution des entrepôts et annuaires de données

Il est possible de mesurer l'évolution d'un entrepôt ou d'un annuaire de différentes façons. C'est ici le critère de la volumétrie qui sera utilisé. A savoir : comment a évolué la volumétrie des 34 entrepôts et annuaires étudiés entre 2016 et 2019 ?

Les tableaux 7 et 8 présentent la volumétrie de 14 entrepôts et 8 annuaires, pour lesquels le nombre de données répertoriées en 2016 et en 2019 a pu être relevé.

Annuaires	**Nombre de notices en 2016**	**Nombre de notices en 2019**	**Taux d'évolution**
BBEES	180	2177	x 12,1
Portail Epidémiologie France	901	914	x 1,0
ECOSCOPE	70	53	x 0,8[228]
ArchiPolis	222	228	x 1,0
Réseau Quetelet	1274	1518	x 1,2
IrsteaData	NC (en cours de création au moment de l'étude)	34	NC
InfoTerre	NC (non relevé)	1564027	NC
IO Data Science	NC (non relevé)	59	NC

Tableau 7 : Taux d'évolution des annuaires de données entre 2016 et 2019

227 Voir supra, deuxième partie, 4.2.2, p.93

228 Le nombre de notices a diminué entre 2016 et 2019. Peut-être est-ce dû à une erreur de relevé ?

Entrepôts	Nombre de jeux de données en 2016	Nombre de jeux de données en 2019	Taux d'évolution
SEANOE	200	487	x 2,4
Ortolang	58	420	x 7,2
ArkeoGIS	64	99	x 1,5
BeQuali	5	12	x 2,4
MédiHAL	22476	42046	x 1,9
Webservice-Energy Catalog	1699	1690	x 1,0
CDS Portal	30587	18764	x 0,6 [229]
Portail Data Inra	NC (en cours de création au moment de l'étude)	78183	NC
PerSCiDO	NC (en cours de création au moment de l'étude)	34	NC
CoCoON	NC (non relevé)	11939	NC
EELS Data Base	NC (non relevé)	276	NC
Kinsources.net	NC (non relevé)	127	NC
VRP-REP	NC (non relevé)	85	NC
Coriolis	NC (non relevé)	12	NC

Tableau 8 : Taux d'évolution des entrepôts de données entre 2016 et 2019

A la date du 23 juin 2019, la volumétrie moyenne des entrepôts était de 11 012 jeux de données ; la volumétrie des annuaires de 196 126 notices.

Considérée individuellement, la volumétrie des entrepôts comme celle des annuaires reste très disparate (pour les mêmes raisons que celles évoquées en 2.2.3, page 126, à savoir qu'elle varie en fonction de l'ancienneté de l'outil, de son périmètre disciplinaire et de la typologie des données qu'il héberge).

- Parmi les entrepôts : BeQuali contient 12 enquêtes, CORIOLIS 12 bases de données, tandis que MédiHAL compte 42 046 jeux de données (images, vidéos, sons et cartes).
- Parmi les annuaires : IrsteaData contient la description de 34 jeux de données, InfoTerre la description de 1 564 027 banques de données.

229 Le nombre de données a diminué entre 2016 et 2019, probablement en raison d'une erreur de relevé en 2016. Le taux d'évolution n'a pas été pris en compte dans la moyenne globale.

Entre 2016 et 2019, la volumétrie des entrepôts a donc été multipliée par 2,7 (calcul réalisé à partir de 6 entrepôts, pour lesquels étaient connues à la fois la volumétrie de 2016 et celle de 2019). Parmi les développements les plus importants :

- Celui de SEANOE, BeQuali et MédiHAL, dont le nombre de jeux de données a quasiment doublé ;
- Celui d'Ortolang, qui contient aujourd'hui 7 fois plus de jeux de données qu'il y a trois ans.

Quant aux annuaires, entre 2016 et 2019, leur volumétrie a été multipliée par 3,2 (calcul réalisé à partir des données de 5 annuaires). La croissance de l'annuaire BBEES est la plus significative : le nombre de ses notices est passé de 180 à 2 177.

La croissance de ces dispositifs de signalement et de diffusion des données semble donc être importante. Il convient néanmoins de noter que, pour plusieurs entrepôts et annuaires, le nombre de notices est resté quasiment le même. C'est le cas du Portail Epidémiologie France, du catalogue ArchiPolis et de l'entrepôt Webservice-Energy Catalog.

On peut également se demander comment a évolué le nombre d'entrepôts et d'annuaires au niveau national et international entre 2016 et 2019. Comment, par exemple, a évolué leur nombre dans Cat OPIDoR ?

A la date du 5 août 2019, y étaient répertoriés :

- 38 annuaires (en 2016, l'étude cartographique en avait révélé 9) ;
- 51 entrepôts (contre 25 en 2016).

Les figures 9 et 10 présentent l'évolution, par grand domaine scientifique, du nombre d'entrepôts et d'annuaires recensés dans Cat OPIDoR.

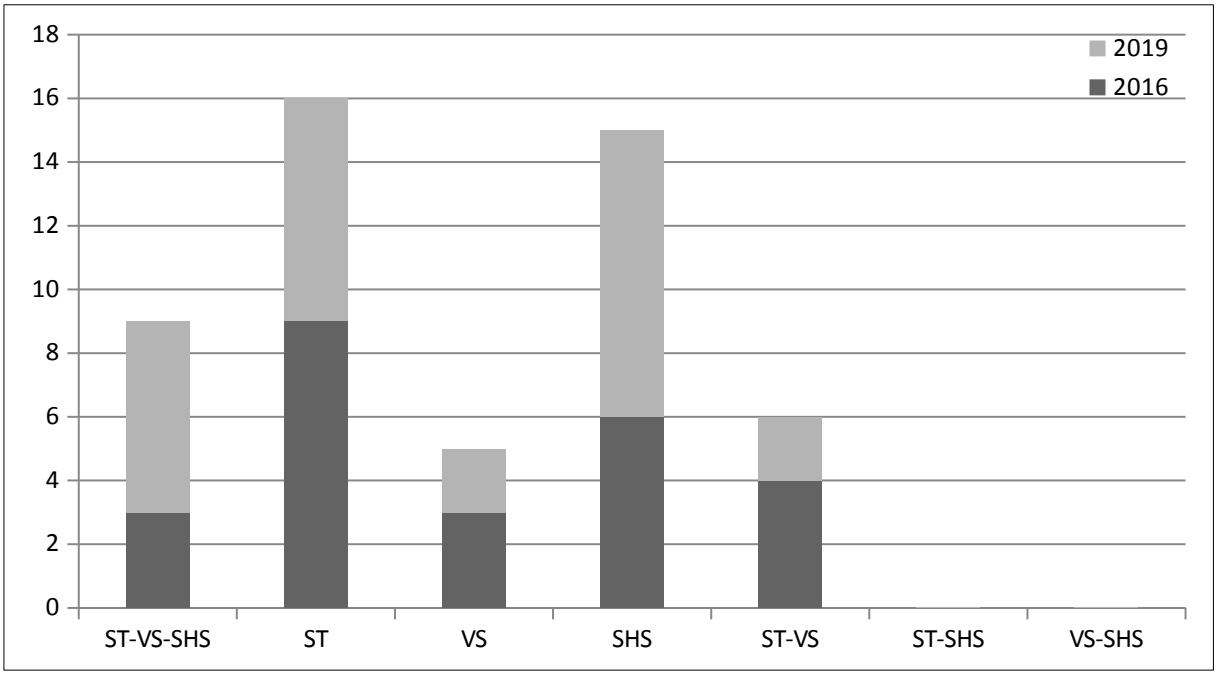


Figure 9 : Evolution du nombre d'annuaires répertoriés dans Cat OPIDoR par grand domaine scientifique

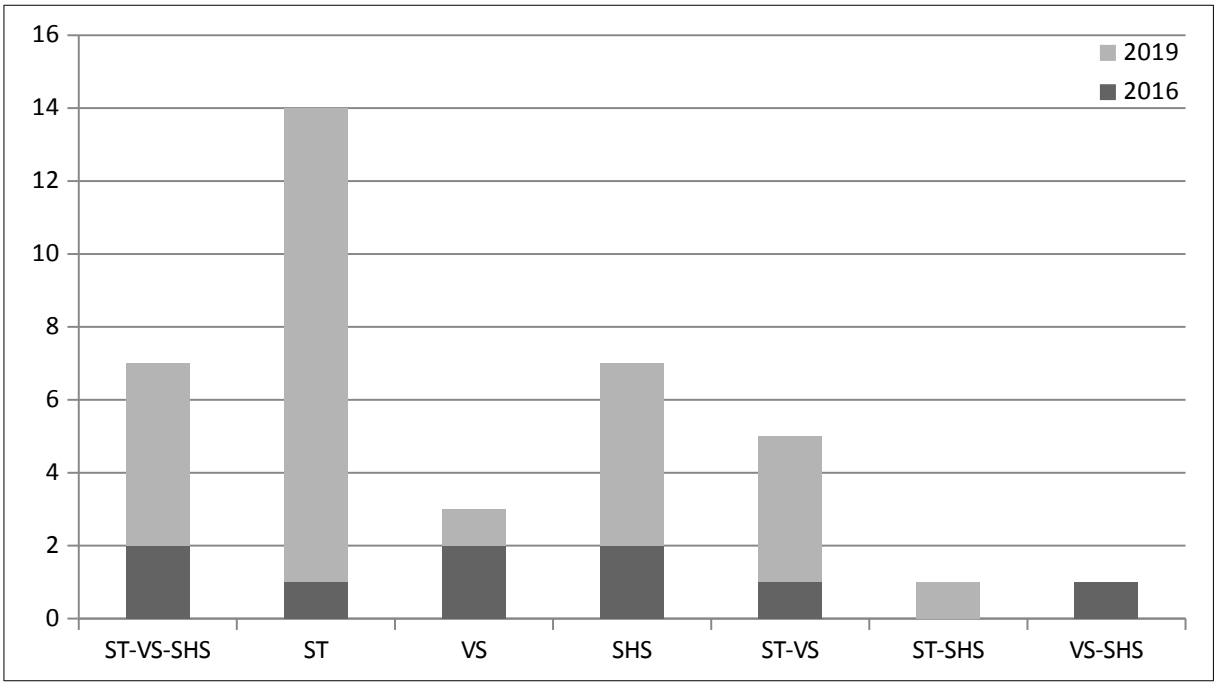


Figure 10 : Evolution du nombre d'entrepôts répertoriés dans Cat OPIDoR par grand domaine scientifique

ST : Sciences & Technologies. VS : Vie & Santé. SHS : Sciences humaines et sociales

Le nombre d'annuaires en Sciences et Technologies a beaucoup augmenté (de 9 à 16 entre 2016 et 2019). L'écart s'est creusé avec les annuaires de Vie et Santé, de Sciences humaines et sociales et les annuaires multidisciplinaires (ST-VS-SHS). On dénombre un seul annuaire supplémentaire répertorié en Vie et Santé depuis 2016.

Pour ce qui est des entrepôts en Vie et Santé, là également, leur nombre a peu évolué, contrairement aux entrepôts multidisciplinaires, de Sciences et Technologies et des Sciences humaines et sociales. Les entrepôts en Sciences et Technologies et en Sciences humaines et sociales restent les plus nombreux.

Il semble donc que la majorité des entrepôts et annuaires soient des outils spécialisés à vocation thématique.

Le même constat peut être fait à partir du répertoire Re3data, dans lequel étaient répertoriés 2371 entrepôts à la date du 5 août 2019, contre 1381 en 2017 (Kindling et al. 2017). La classification disciplinaire utilisée par le répertoire donne à voir une grande variété d'entrepôts, spécialisés dans une thématique précise. Les entrepôts multidisciplinaires (233), quant à eux, représentent seulement 10% du total.

Les résultats montrent un nombre d'entrepôts et d'annuaires en augmentation. Néanmoins, à quoi faut-il attribuer cette évolution ? Est-elle due à la création de nouveaux entrepôts et annuaires, ou bien à la progression du référencement de ces outils dans les répertoires Cat OPIDoR et Re3data ? Fin 2015, Re3data a par exemple modifié sa politique d'indexation. Auparavant seuls les entrepôts proposant une interface d'accueil en anglais pouvaient être répertoriés. Par la suite, la langue n'a plus été un critère discriminatif. Un certain nombre d'entrepôts, qui n'avaient pas pu être indexés jusque là, ont donc probablement été ajoutés. De même, grâce à la campagne de communication réalisée au moment du lancement de Cat OPIDoR en 2017, plusieurs fournisseurs de services ont souhaité répertorier leur entrepôt ou annuaire. Néanmoins, il est vrai aussi que le contexte de promotion de la science ouverte a pu donner naissance à de nouveaux entrepôts et annuaires de données. SciencesPo a par exemple développé DataSPIRE[230], un entrepôt dédié aux données produites par les chercheurs de l'établissement. En sciences de la Terre, un portail d'accès unifié aux données d'imagerie

230 https://catalogues.cdsp.sciences-po.fr/dataverse/dataspire

satellitaire est en cours de création : il a été nommé DINAMIS (Dispositif Institutionnel National d'Approvisionnement Mutualisé en Imagerie Satellitaire)[231].

2.5.2. Evolution des autres types de services

L'évolution des autres types de services est difficile à percevoir à partir d'une simple consultation de leur site web.

Les services répertoriés en 2016 sont toujours en place en 2019, à l'exception du site Données de la recherche, et semblent s'être stabilisés. Le site d'information CoopIST du Cirad publie chaque année environ deux nouvelles fiches didactiques sur la gestion et l'ouverture des données. Le site DataPartage de l'INRA est un peu plus dynamique, car il inclut un flux d'actualités et relaie des informations concernant l'entrepôt de l'institut (Data Inra). L'offre de service de l'Inist-CNRS s'est, quant à elle, structurée autour du portail OPIDoR. Celui-ci met à disposition de la communauté de l'enseignement supérieur et de la recherche un ensemble d'outils et de services facilitant la gestion et la valorisation des données. Il articule actuellement les trois services DMP OPIDoR, Cat OPIDoR et PID OPIDoR (le service d'attribution de DOI), mais est destiné à proposer davantage de services à l'avenir.

La montée en puissance du discours d'ouverture de la science a par ailleurs donné lieu à de plus nombreuses journées d'études, formations, veilles ou « guides de bonnes pratiques ».

- Diverses structures organisent désormais des journées de sensibilisation à la gestion et l'ouverture des données de recherche : l'Observatoire des Sciences de l'Univers THETA en association avec la MSHE Ledoux de Besançon par exemple[232] ; l'Atelier Données du CNRS[233] ; le Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS) et l'Observatoire Midi-Pyrénées (OMP)[234]…

231 https://dinamis.teledetection.fr/

232 Ils ont organisé le colloque DataBFC les 13, 14 et 15 novembre 2017 (https://databfc.sciencesconf.org/).

233 L'Atelier Données est un groupe de travail inter-réseaux du CNRS, regroupant les représentants des réseaux thématiques suivants : CALCUL, DEVLOG, MEDICI, RDBB, RENATIS, RESINFO, QeR, Frantiq et MASA. Le 27 novembre 2018 a été organisée une journée d'étude intitulée « Interopérabilité et pérennisation des données de la recherche : comment FAIR en pratique ? » (https://gt-donnees2018.sciencesconf.org/).

234 Ils ont été à l'initiative d'une journée d'étude, intitulée « Partager les données de la recherche : pour qui, comment, pourquoi ? », le 14 novembre 2017 (https://datalaasomp.sciencesconf.org/).

- Les URFIST proposent des formations à destination des personnels de recherche : 13 formations sur le thème des données de la recherche ont été organisées dans 6 des 7 URFIST au cours du premier semestre 2019.
- Des guides à destination de la communauté scientifique ont également été produits. On peut citer le guide « Données de la recherche »[235] du réseau Form@doct, s'adressant aux doctorants de l'Université Bretagne Loire. Un groupe de travail inter-organismes a par ailleurs rédigé, sous l'égide du Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, un guide d'analyse du cadre juridique des données de la recherche en France (Becard et al. 2017).
- Quant aux bibliothèques universitaires, elles sont de plus en plus nombreuses à se doter d'un pôle dédié aux données de recherche, se composant de personnels chargés de conseiller et d'accompagner les chercheurs. C'est le cas des bibliothèques de l'Université de Nice Sophia Antipolis[236] et de l'Université de Lorraine.

235 SERRES, A., VIGNALE, F. (2016). 'Les données de la recherche'. *Formadoct*. Rennes : Université Européenne de Bretagne. http://guides-formadoct.u-bretagneloire.fr/donnees_recherche (consulté le 20 septembre 2019).

236 https://bu.univ-cotedazur.fr/fr/utiliser-nos-services/services-a-la-recherche

3. Conclusion

Le paysage national des services de gestion et d'ouverture des données scientifiques se caractérise par son hétérogénéité. Les services qui le composent se révèlent divers tant par leurs fonctions que par leur structure et les acteurs qui en sont à l'initiative. Comme le soulignent Marcial et Hemminger (Marcial et Hemminger 2010) dans une étude sur les services de données :

> *« Il est reconnu de manière générale que les services de données de recherche sont essentiels pour l'avenir de la science et qu'il est nécessaire de trouver le moyen de les préserver et de faire valoir leur richesse à travers les disciplines. Une question est en revanche moins évidente : celle de savoir quels sont les composants essentiels des services de données qui rendront cela possible. »*[237]

Parmi les acteurs se positionnant sur le développement d'une offre de services se distinguent les professionnels de l'information scientifique et technique. Ces derniers conçoivent la fourniture de services pour la gestion et le partage des données comme une extension de leurs fonctions traditionnelles (Koltay 2016). En s'emparant de cette question, ils tentent de se positionner en tant que médiateurs entre les instances politiques qui prônent une ouverture des données et les communautés scientifiques. Vincent Liquète (2010) définit la médiation comme *« la recherche du lien entre l'énonciateur et le récepteur ». « Ce lien s'établit, grâce à une tierce personne et/ou un ensemble de techniques, d'outils, de messages ou d'interfaces accompagnant le récepteur (usager, client, citoyen) afin de lui faciliter la compréhension par la construction de sens, pouvant se solder par un changement (d'actions, de représentations, etc.) de sa part »*. En ce sens, les professionnels de l'IST développent des solutions pratiques pour gérer et ouvrir les données et transmettre cette culture aux chercheurs.

Ils sont, tout à la fois, à l'origine d'un rapprochement entre le concept de *wicked problem*[238] (Rittel et Webber 1973) et la question de la gestion des données (Awre et al. 2015 ; Cox et al. 2016), ce qui peut être révélateur de leur position de médiateurs. Le terme *wicked problem* fait

237 Traduction de : « There is an almost universal recognition that SDRs are critical to the future of science, and a means for preserving them and for leveraging their richness across disciplines is needed. It is less clear what the essential components of SDRs are that will make this possible. » (Marcial et Hemminger 2010, p.2044)

238 Que l'on peut traduire en français par « problème épineux ».

référence à une situation difficile à résoudre, pour laquelle il n'existe pas de solution unique, en raison d'exigences incomplètes, contradictoires et changeantes[239]. La gestion des données de la recherche a été qualifiée de « *wicked problem* », en raison de l'hétérogénéité et de la complexité des données, mais aussi en raison du manque de clarté sur le(s) type(s) de services à délivrer. Cette association est révélatrice de la position des professionnels de l'IST qui souhaitent s'investir dans la gestion des données scientifiques. D'une part, l'ouverture des données est un mot d'ordre politique dépourvu de directive claire[240]. D'autre part, les professionnels de l'IST tentent de le relayer, tout en étant étrangers aux pratiques de production et de communication des données au sein des équipes de recherche. Étant donné la complexité de ce que tente d'englober la notion de « donnée de recherche », il semble en effet impossible de développer une offre de service adaptée, sans connaître les besoins et la configuration des communautés de recherche, voire même sans en faire partie. Les difficultés rencontrées par les professionnels de l'IST viennent souvent du fait qu'ils sont extérieurs aux communautés scientifiques.

239 https://en.wikipedia.org/wiki/Wicked_problem

240 Cf. supra, deuxième partie, 5, p.96

Quatrième partie

-

Les données dans les pratiques de recherche

Le développement de politiques publiques et de services pour l'ouverture des données scientifiques révèle une logique descendante (*top-down*), qui questionne sur leur introduction dans les communautés de recherche. Comment la philosophie de l'ouverture est-elle perçue par les chercheurs ? Quelle place occupent les données dans les pratiques des différentes communautés de recherche ? Telles sont les questions auxquelles tente de répondre cette quatrième partie, qui mobilisera les résultats d'une enquête qualitative menée auprès de 57 chercheurs de l'Université de Strasbourg. Il s'agira de vérifier les hypothèses émises en introduction, sur l'influence du cadre épistémique (première hypothèse), institutionnel (deuxième hypothèse) et social (troisième hypothèse) sur les modes de gestion et de partage des données.

1. Terrain et méthodologie

1.1. Une enquête sous forme d'entretiens semi-directifs

Une approche qualitative a été choisie pour étudier les pratiques de recherche. *« Les méthodes qualitatives ont pour fonction de comprendre plus que de décrire systématiquement ou de mesurer »* (Kaufmann 1996). Elles permettent de rendre intelligible la complexité des actions d'un individu dans un contexte particulier, tout en donnant à voir les dimensions collective et individuelle des pratiques. Pour servir cette posture compréhensive, a été utilisée la forme de l'entretien semi-directif. L'entretien semi-directif s'articule autour de différents thèmes, définis au préalable par l'enquêteur. Il présente l'avantage de garder la discussion relativement ouverte et de laisser à la personne interrogée un espace suffisamment large pour donner son point de vue librement.

L'objectif était d'étudier divers contextes de recherche, pour tenter de comprendre ce qui influe sur les pratiques de gestion et de partage des données. Aussi l'enquête n'était-elle pas restreinte à un nombre fixe de disciplines ciblées, mais se voulait ouverte à l'ensemble des domaines scientifiques.

1.2. Méthode d'échantillonnage

1.2.1. L'Université de Strasbourg pour terrain d'enquête

Le choix du terrain d'étude s'est porté sur l'Université de Strasbourg, sur la base de critères de représentativité disciplinaire et d'« excellence de la recherche ». L'Université de Strasbourg est un établissement d'enseignement supérieur et de recherche, couvrant l'ensemble des champs disciplinaires et comptant 73 unités de recherche pour 2 755 enseignants-chercheurs[241]. Elle bénéficie notamment de l'IdEx, un des financements du programme Investissements d'avenir, visant à *« [réunir], selon une logique de territoire, des établissements d'enseignement supérieur et de recherche déjà reconnus pour leur excellence scientifique et pédagogique »*[242]. L'ambition du Ministère de l'Enseignement supérieur, de la

241 Source : http://www.unistra.fr/index.php?id=27943 (page consultée le 3 mars 2019)
242 Source : http://www.enseignementsup-recherche.gouv.fr/cid51351/initiatives-d-excellence.html

Recherche et de l'Innovation est de *« doter la France d'initiatives d'excellence capables de rivaliser avec les meilleures universités du monde »*[242]. En matière de gestion et d'ouverture des données de la recherche, elle a pour modèle de renommée internationale le Centre de Données astronomiques de Strasbourg[243].

1.2.2. Populations sondées

Échantillon initial

Un premier échantillon a été constitué en novembre 2017, avec pour critère de sélection l'« excellence scientifique », mot d'ordre récurrent dans les discours politiques sur la recherche. L'hypothèse était qu'en étant définis comme « excellents » par les financeurs, ces chercheurs seraient d'autant plus incités à gérer et ouvrir leurs données (à l'instar des projets financés par le programme H2020, auxquels est demandée la rédaction d'un plan de gestion de données ainsi que la diffusion des données, si possible en libre accès[244]).

Ont donc été sélectionnés les laboratoires de l'Université de Strasbourg ayant obtenu une notation supérieure ou égale à A dans le rapport d'évaluation 2011-2012 de l'AERES[245]. Au total 26 unités de recherche ont été identifiées. Dans chacune d'elle, les projets de recherche ayant cours au moment de l'échantillonnage et bénéficiant d'un financement public (type H2020 ou ANR[246]) ont été ciblés, soit 29 projets au total. Leurs coordinateurs (chercheurs et enseignants-chercheurs) ont été sollicités pour participer à l'enquête ; 12 d'entre eux ont répondu favorablement. Au total, 12 entretiens ont donc pu être menés entre novembre 2017 et octobre 2018.

---

243 Voir supra, troisième partie, 2.2.2, *Portée géographique des services*, p.118

244 Voir supra, deuxième partie, 3.2.3, p.74
Horizon 2020 est le programme de financement 2014-2020 de la Commission européenne (http://www.horizon2020.gouv.fr/).

245 Rapport d'évaluation des unités de recherche 2011-2012 : https://www.hceres.fr/LISTE-ALPHABETIQUE-DES-ETABLISSEMENTS-ET-ORGANISMES-EVALUES/UNIVERSITE-DE-STRASBOURG-UNISTRA.
Il s'agissait de la dernière évaluation rendue publique dans sa totalité, lorsque l'échantillon a été constitué en octobre 2017.
L'AERES étant l'Agence d'Evaluation de la Recherche et de l'Enseignement Supérieur, aujourd'hui renommée Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (Hcéres).

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

L'échantillon des 26 laboratoires contenait peu d'unités de recherche en sciences humaines et sociales (3 sur 26). Afin de compenser ce déséquilibre, trois laboratoires susceptibles de manipuler des volumes importants de données ont été ajoutés à l'échantillon initial : une unité mixte de recherche en économie ; une unité mixte de recherche en géographie ; une unité mixte de recherche en sciences sociales. Dans chacune d'elles, les projets de recherche bénéficiant d'un financement ANR ou H2020 ont été identifiés, soit 5 projets au total. Un seul coordinateur n'a pas souhaité participer à l'enquête ; 4 entretiens ont donc été menés.

Cet échantillon initial comptait donc 16 chercheurs issus de 13 laboratoires différents (tableau 9).

Consolidation de l'échantillon

Un second échantillon a été sondé entre mars et mai 2019, dans le but de consolider le premier ensemble d'entretiens. Les chercheurs de chacune des 13 unités de recherche, dans lesquelles les 16 premiers entretiens avaient été menés, ont été contactés, laboratoire par laboratoire, jusqu'à saturation des informations recueillies. *« La saturation est le phénomène par lequel, passé un certain nombre d'entretiens [...], le chercheur ou l'équipe a l'impression de ne plus rien apprendre de nouveau, du moins en ce qui concerne l'objet sociologique de l'enquête »* (Bertaux 1980, p.205).

Au total, 228 chercheurs ont été contactés parmi 11 des 13 laboratoires initiaux. Leurs réponses ont conduit à réaliser 41 entretiens supplémentaires (tableau 10).

Grand domaine scientifique	Laboratoire	Discipline du laboratoire	Nombre de chercheurs contactés	Nombre d'entretiens menés
Vie & Santé	Laboratoire 1	Biologie	1	1
	Laboratoire 2	Chimie	1	1
	Laboratoire 3	Chimie et biologie	1	1
	Laboratoire 4	Chimie et biologie	1	1
	Laboratoire 5	Écologie	1	1
	Laboratoire 6	Neurosciences	1	1
	Laboratoire 7	Neurosciences	1	1
Sciences & Technologies	Laboratoire 8	Astronomie	2	1
	Laboratoire 9	Sciences de l'ingénieur	3	2
	Laboratoire 10	Sciences de la Terre	1	1
Sciences humaines et sociales	Laboratoire 11	Géographie	2	2
	Laboratoire 12	Sciences sociales	3	2
	Laboratoire 13	Théologie	1	1
Vie & Santé	Laboratoire 14	Biologie, Chimie, et Physique	1	0
	Laboratoire 15	Biologie	3	0
	Laboratoire 16	Biologie	1	0
	Laboratoire 17	Biologie	1	0
	Laboratoire 18	Ecologie	1	0
Sciences & Technologies	Laboratoire 19	Science de la Terre	1	0
	Laboratoire 20	Mathématiques	2	0
	Laboratoire 21	Chimie et Physique	3	0
Sciences humaines et sociales	Laboratoire 22	Economie	1	0
	Laboratoire 23	Langues	1	0
		Total	**34**	**16**
		Taux de participation	**47 %**	

Tableau 9 : Composition de l'échantillon initial

Grand domaine scientifique	Laboratoire	Discipline du laboratoire	Nombre de chercheurs contactés	Nombre d'entretiens menés	Taux de participation (%)
Vie & Santé	Laboratoire 3	Chimie et biologie	30	3	10 %
	Laboratoire 4	Chimie et biologie	22	5	23 %
	Laboratoire 5	Écologie	19	3	16 %
	Laboratoire 6	Neurosciences	20	4	20 %
	Laboratoire 7	Neurosciences	9	0	0 %
Sciences & Technologies	Laboratoire 8	Astronomie	15	6	40 %
	Laboratoire 9	Sciences de l'ingénieur	33	5	15 %
	Laboratoire 10	Sciences de la Terre	23	3	13 %
Sciences humaines et sociales	Laboratoire 11	Géographie	21	3	14 %
	Laboratoire 12	Sciences sociales	19	7	37 %
	Laboratoire 13	Théologie	17	2	12 %
			228	**41**	**18 %**

Tableau 10 : Composition du second échantillon

1.3. Conduite des entretiens

Un mail personnalisé, présentant le sujet de l'enquête et proposant un entretien, a été envoyé aux différents chercheurs. Lorsque l'un d'eux acceptait, un rendez-vous était convenu. L'entretien était alors mené sur son lieu de travail ou par téléphone.

Avant de commencer l'entretien, était remis au chercheur un formulaire de consentement éclairé, que nous lisions et signions ensemble (annexe 9). Ce formulaire résumait brièvement le sujet de la recherche et exposait les termes de notre engagement respectif. En signant ce document, le chercheur acceptait de participer à l'enquête, tout en prenant acte de son droit de retrait. Il acceptait également que l'entretien soit enregistré. Pour ma part, je m'engageais à garder son identité confidentielle.

D'une durée d'une heure en moyenne, les entretiens reposaient sur un ensemble de thèmes préalablement définis dans un guide d'entretien (annexes 10 et 11). Des questions ouvertes

étaient posées au coordinateur scientifique. Il s'agissait en premier lieu de questions relatives au contexte du projet de recherche. Cela permettait au chercheur de se replacer d'emblée dans sa condition d'expert et d'évoquer la pratique personnelle et quotidienne de cette expertise. Cette étape contribuait également à instaurer une confiance réciproque entre enquêteur et enquêté, de manière à rendre la parole exprimée la plus sincère et la moins formelle possible. Par la suite étaient posées des questions relatives à la manière de documenter, stocker et partager les données.

Les entretiens étaient enregistrés à l'aide d'un dictaphone puis retranscrits. Tous les chercheurs rencontrés, à l'exception d'un seul, ont autorisé l'enregistrement de la discussion[247].

1.4. Méthode d'analyse des entretiens

Une fois retranscrits, les entretiens ont été relus attentivement et annotés. La forme semi-directive a parfois permis au chercheur d'élargir la discussion à des thèmes connexes, qui n'avaient pas été recensés dans le guide d'entretien et qui se sont révélés très enrichissants pour l'analyse. Les chercheurs ont par exemple souvent mentionné le terme de « collaboration », faisant référence à des formes de travail en commun, plus ou moins tacites. Or il s'est avéré, en creusant cette question au fil des entretiens, que l'échange de données se structurait fréquemment autour ce type de rapports.

Dans une seconde phase, les annotations des différents entretiens ont été rassemblées dans des tableaux thématiques et comparées entre elles (deux exemples sont présentés en annexe 13). Le but était d'observer d'éventuelles récurrences ou variations et d'identifier ce qui pouvait les expliquer (contexte de recherche, mode d'acquisition des données, enjeux sous-jacents...).

247 Une sélection de retranscriptions d'entretiens figure en annexe 12.

2. Résultats et discussion

2.1. Résultats de participation à l'enquête

Au total, 262 demandes d'entretien ont été envoyées. Parmi les chercheurs contactés, 86 d'entre eux étaient affiliés à un laboratoire de sciences et technologies, 111 à un laboratoire de vie et santé et 65 à un laboratoire de sciences humaines et sociales[248].

Sur ces 262 demandes, 57 chercheurs ont répondu favorablement à la demande d'entretien, ce qui correspond à un taux de participation de 22% (tableau 11).

57 entretiens ont donc pu être menés, dont :

- 18 avec des chercheurs affiliés à un laboratoire de sciences et technologies ;
- 22 avec des chercheurs affiliés à un laboratoire de vie et santé ;
- et 17 avec des chercheurs affiliés à un laboratoire de sciences humaines et sociales.

248 Les catégories disciplinaires utilisées ici ont été proposées par le Conseil européen de la recherche (https://cat.opidor.fr/index.php/Nomenclature_ERC).

Grand domaine scientifique	Laboratoire	Chercheur	Discipline	Date de l'entretien
Vie & Santé	Laboratoire 1	Chercheur 1	Génétique	juin 2018
	Laboratoire 2	Chercheur 2	Chimie thérapeutique	janvier 2018
	Laboratoire 3	Chercheur 3	Biologie	avril 2018
		Chercheur 4	Biologie	mai 2019
		Chercheur 5	Biologie	mai 2019
		Chercheur 6	Chimie et biologie	mai 2019
	Laboratoire 4	Chercheur 7	Biologie	mars 2019
		Chercheur 8	Biologie	mars 2019
		Chercheur 9	Biologie	mars 2019
		Chercheur 10	Biologie	avril 2019
		Chercheur 11	Chimie thérapeutique	novembre 2017
		Chercheur 12	Médecine	mars 2019
	Laboratoire 5	Chercheur 13	Écologie	mai 2019
		Chercheur 14	Écologie	mai 2019
		Chercheur 15	Écologie	juin 2019
		Chercheur 16	Éthologie	juin 2018
	Laboratoire 6	Chercheur 17	Bioinformatique	avril 2019
		Chercheur 18	Neurosciences	janvier 2018
		Chercheur 19	Neurosciences	avril 2019
		Chercheur 20	Neurosciences	avril 2019
		Chercheur 21	Primatologie	avril 2019
	Laboratoire 7	Chercheur 22	Neurosciences	mars 2018
Sciences & Technologies	Laboratoire 8	Chercheur 23	Astronomie	avril 2019
		Chercheur 24	Astronomie	avril 2019
		Chercheur 25	Astronomie	avril 2019
		Chercheur 26	Astronomie	avril 2019
		Chercheur 27	Astronomie	avril 2019
		Chercheur 28	Astronomie	avril 2019
		Chercheur 29	Astronomie	juin 2018
	Laboratoire 9	Chercheur 30	Bioinformatique	avril 2019
		Chercheur 31	Géographie	avril 2019
		Chercheur 32	Informatique	avril 2019
		Chercheur 33	Informatique	avril 2019
		Chercheur 34	Informatique	mai 2019
		Chercheur 35	Sciences de l'ingénieur	novembre 2017
		Chercheur 36	Sciences de l'ingénieur	septembre 2018
	Laboratoire 10	Chercheur 37	Géologie	avril 2018
		Chercheur 38	Géologie	mars 2019
		Chercheur 39	Géologie	mars 2019
		Chercheur 40	Géologie	mars 2019
Sciences humaines et sociales	Laboratoire 11	Chercheur 41	Ecologie	mars 2019
		Chercheur 42	Géo-archéologie	avril 2019
		Chercheur 43	Géographie	septembre 2018
		Chercheur 44	Géographie	octobre 2018
		Chercheur 45	Géographie	avril 2019
	Laboratoire 12	Chercheur 46	Anthropologie	mars 2019
		Chercheur 47	Démographie	mars 2019
		Chercheur 48	Démographie	mars 2019
		Chercheur 49	Droit	mars 2019
		Chercheur 50	Histoire	avril 2019
		Chercheur 51	Sciences politiques	octobre 2018
		Chercheur 52	Sociologie	septembre 2018
		Chercheur 53	Sociologie	mars 2019
		Chercheur 54	Sociologie	mars 2019
	Laboratoire 13	Chercheur 55	Théologie	juillet 2018
		Chercheur 56	Théologie	avril 2019
		Chercheur 57	Théologie	avril 2019

Tableau 11 : Liste des chercheurs interrogés

Les résultats présentés ici ne se veulent pas représentatifs d'une discipline. Ils tentent seulement de dégager des tendances à partir de pratiques individuelles décrites par les chercheurs interrogés. Par ailleurs, le panel présente les deux biais suivants. Le premier est de ne pas être parfaitement équilibré en termes de domaines scientifiques. Les chercheurs du domaine vie et santé sont surreprésentés (22 des 57 chercheurs rencontrés) par rapport à ceux du domaine des sciences et technologies (18 chercheurs) et ceux du domaine des sciences humaines et sociales (17 chercheurs). Le second biais du panel est qu'il se limite très probablement à des chercheurs auxquels le terme de « données » est éloquent (bien que plusieurs synonymes aient été listés dans les mails de demande d'entretien). Plusieurs réponses négatives penchent dans ce sens. Dix chercheurs ont en effet décliné l'entretien, au motif qu'ils n'utilisaient pas ou très peu de données dans leurs recherches. On peut également supposer qu'une partie des réponses négatives et des non réponses provient de chercheurs qui s'intéressent peu au devenir des données qu'ils génèrent. En termes de participation, on trouve le taux de réponse le plus élevé en astronomie (laboratoire 8). Le laboratoire en question est adossé au Centre de Données astronomiques de Strasbourg (CDS), qui depuis 1972 collecte et distribue des données astronomiques au niveau international[249]. La question des données et de leur partage est donc inscrite dans la culture du laboratoire. Ses membres en sont familiers. Un des taux de participation les plus bas concerne la théologie (laboratoire 13), où le terme de « données » semble ne pas être utilisé (cela sera évoqué plus loin, 2.4, p.178).

2.2. Terminologie : le sens du terme « donnée » dans le discours des chercheurs

Une des questions posées au cours des entretiens était : « Utilisez-vous le terme de « donnée » dans le cadre vos recherches ? ». La plupart des chercheurs (52 sur 57) ont répondu par l'affirmative.

Dans le domaine des sciences, techniques et médecine, les chercheurs sont unanimes sur l'usage du terme. Le mot « donnée » y est en effet plus courant que dans les sciences humaines et sociales (Cabrera 2014, p.52; Schöpfel 2018b, p.10), dans la mesure où son

249 Voir supra, troisième partie, 2.2.2, *Portée géographique des services*, p.118

acception renvoie originellement au caractère quantitatif de ce qui a été mesuré, c'est-à-dire à la donnée chiffrée. Cette association a d'ailleurs été faite par plusieurs chercheurs.

> *« Quand je dis « données » ou « data », dans ma tête c'est numérique – ce qui n'est pas forcément sensé, ça n'est pas très réfléchi ce que je dis. Si je faisais de l'analyse bibliométrique, là je parlerais de « données », parce que ce sont des chiffres. »* (chercheur 53)

En sciences humaines et sociales, dans des disciplines comme le droit, la théologie ou l'anthropologie, le terme de « donnée » est moins courant. Cinq chercheurs déclarent ne pas l'utiliser (chercheurs 46, 49, 50, 53 et 57). On peut se demander s'il existe un ou des terme(s) équivalent(s) dans ces disciplines.

Certains chercheurs parlent de « matériau » ou de « sources », d'autres d'« informations ». Or il semble que ces deux catégories ne soient pas superposables : une source n'est pas une information ; mais contient des informations. En théologie par exemple, un chercheur travaillait à partir de commentaires bibliques (ses sources). Dans ces textes, il relevait les occurrences de champs lexicaux (les informations). Dans ce cas, quel est l'équivalent du terme « donnée » : la source ou l'information ? Les chercheurs ne sont pas tous d'accord à ce sujet. On peut comparer leurs définitions avec celles qu'en donnent les chercheurs dans les autres domaines. En sciences, techniques, et médecine, les scientifiques associent souvent la donnée à l'information qu'ils extraient d'une observation ou d'une expérimentation (le nombre de cellules cancéreuses sur une coupe de tissu malade par exemple). En sociologie, lorsque le chercheur réalise une enquête qualitative, il n'associe pas l'entretien ou sa retranscription à la donnée. Les données sont plutôt les informations qu'il va extraire de ces entretiens : des données informatives ainsi que les perceptions et représentations des enquêtés (Pinson et Pala 2007). De manière générale, la donnée serait donc plutôt l'information que le chercheur extrait de l'« objet » qu'il étudie (un commentaire biblique, une protéine, un astre...). Dans les disciplines où les chercheurs n'utilisent pas le mot « donnée », il ne semble toutefois pas y avoir de terme générique pour caractériser ce qu'ils extraient des « sources » ou « matériaux ». La terminologie reste donc celle des objets manipulés.

Cette particularité du droit, de l'anthropologie ou de la théologie tient probablement à ce que Joachim Schöpfel, reprenant les termes d'un chercheur, qualifie d'« artisanat ». *« 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 »* (Schöpfel 2018b, p.11). Dans le même ordre d'idées, le chercheur 52 qualifiait sa méthodologie d'analyse d'entretiens de « bricolage » :

> *« Le bricolage n'est pas non scientifique. J'ai été un temps enseignant-chercheur au Royaume-Uni, où là pour le coup il y a une normalisation terrible de la recherche en sciences sociales et où, quand on a des entretiens et qu'on n'utilise pas le logiciel d'analyse des entretiens, on passe pour un guignol. Moi je trouve que c'est vraiment une forme de normalisation et de standardisation de la recherche qui pose problème, à mon sens. Alors, dans certains cas, quand on fait de l'analyse de lexique par exemple, ça peut être pertinent. Mais dans les types d'analyse que, nous, on proposait, c'est-à-dire des études locales assez denses qui cherchent à comprendre des processus, ça n'avait aucun sens de mobiliser du logiciel. »*

Les termes d'« artisanat » et de « bricolage » renvoient à une méthodologie individuelle, unique car propre à chaque chercheur. Celle-ci n'aboutit pas au recueil de données standardisées et homogènes. C'est ce que véhicule la comparaison sous-jacente entre artisanat et industrie : l'objet artisanal a un caractère unique car il procède d'ajustements, qui sont fonction du matériau travaillé ; dans l'industrie, des protocoles normalisés et des gestes répétés permettent de produire des objets chaque fois identiques. C'est cette nuance qui distingue la théologie ou l'anthropologie des autres disciplines scientifiques, dans lesquelles le terme de « donnée » est courant.

Les chercheurs ont également été interrogés sur le sens qu'ils donnaient au terme de « données ». Le tableau 12 liste quelques unes des définitions qui ont été formulées. Les chercheurs ont souvent eu des difficultés à définir ce qu'était pour eux une donnée. Parfois, ils retournaient la question et demandaient : « Vous, qu'est-ce que vous entendez par donnée ? », comme si le terme pouvait recouper différentes choses. Ces hésitations reflètent les

conclusions de Sabina Leonelli, de Christine L. Borgman et de Joachim Schöpfel, présentées dans la première partie (2.2, p.39).

Deux points, caractéristiques des données, ont été relevés de manière récurrente par les chercheurs.

Première caractéristique : la donnée est une information élémentaire, sur laquelle se fonde le raisonnement scientifique. On retrouve cet aspect dans la conception pyramidale « données, informations, connaissances » de Chaim Zins et de Marcia J. Bates[250]. L'image de la brique évoquée par le chercheur 48, à partir de laquelle on construit une connaissance, renvoie également à la définition proposée par Sabina Leonelli : la donnée est la preuve sur laquelle le chercheur s'appuie pour justifier une théorie[251].

Deuxième caractéristique mise en évidence : une donnée seule n'a pas de valeur scientifique. C'est en agrégeant et en comparant plusieurs données qu'une connaissance pourra être élaborée. *« La valeur vient du fait que ces données sont mises dans un fichier Excel et comparées semaine après semaine, qu'on voit une évolution et qu'on peut soumettre cette évolution à un test statistique et voir si on a un résultat ou non »* (chercheur 22). Malingre et al. (2019, p.7) l'expliquaient en d'autres termes : *« En soi, une donnée seule n'a aucune signification, elle ne prendra sens qu'avec le croisement, l'articulation avec d'autres données, ce qui donnera lieu à une information porteuse de sens »*.

250 Voir supra, première partie, 2.1.2, p.35

251 Voir supra, première partie, 2.2.2, p.41

	Discipline	Définition de « donnée de recherche »	Exemples de données utilisées
Chercheur 48	Démographie	« C'est la brique de base de l'information. Une donnée c'est quelque chose qu'on a là. Ça peut être une réponse à une question, qui est encodée. Elle est là et je vais pouvoir la manipuler.C'est-à-dire c'est le fait de pouvoir agréger cette information avec d'autres, la mettre à l'épreuve d'algorithmes... Pour moi, la donnée c'est de la matière première. C'est quelque chose qui va être dans un format qui va me permettre de faire de la recherche. Il me faut de la donnée pour pouvoir produire de la recherche, c'est-à-dire mettre en avant des connaissances. »	Données de grandes enquêtes quantitatives
Chercheur 52	Sociologie	« Le produit brut des enquêtes de terrain. Mais produit pas tant brut que ça non plus, parce que le type de données qu'on récolte est lié aux choix qu'on a fait au départ dans l'enquête, aux hypothèses qu'on a formulées, aux choix d'aller voir tel ou tel acteur plutôt que tel ou tel autre... Donc les données sont déjà passées dans une moulinette au départ. »	Données d'entretiens qualitatifs
Chercheur 55	Théologie	« Informations sur des sources »	Informations bibliographiques sur des recueils de chant du XVIème siècle Informations sur la localisation des recueils (archives, bibliothèques...) Contenus des recueils (textes, mélodies)
Chercheur 57	Théologie	« Pour moi, « données » signifie des informations caractéristiques du monde de l'électronique. Quand je travaille sur un manuscrit, ce n'est pas une donnée, c'est une source. Je préfère garder mon terme traditionnel. »	Commentaires bibliques, traités théologiques, sermons, lettres…
Chercheur 45	Géographie	« C'est quelque chose de factuel, soit en lien avec des pratiques observables (par exemple, quand on utilise le GPS, on est sur des faits mesurables), soit en lien avec les représentations que les individus peuvent avoir et qui sont appréhendées par toute une série d'échelles psychométriques (on considère ces représentations comme des variables aussi factuelles que les choses qu'on peut mesurer de façon extérieure à l'individu). »	Enquêtes quantitatives Données GPS, traçant les déplacements d'un individu Bases de données administratives

Chercheur 16	Ethologie	« Une information qui a été récoltée sur un participant »	Données d'enquête par questionnaire Données de mesures issues de capteurs GPS, accélérométrie et de proximité
Chercheur 43	Géographie	« Pour moi, une donnée reste quelque chose de brut. Une image satellite, par exemple, c'est brut. »	Données terrestres issues d'instruments de télédétection
Chercheur 29	Astronomie	« Généralement, quand on parle de base de données, ça fait plutôt référence à une liste homogène de mêmes choses (une base de données de clients, une base de données de prise de la température...). »	Données de simulations numériques Images satellites du ciel Mesures d'ondes électromagnétiques
Chercheur 15	Écologie	« On en parle tout le temps de « données », mais on ne se pose jamais la question. Les données c'est ce sur quoi on travaille. Les données c'est toutes les variables qu'on va collecter sur le terrain ou sur nos animaux. C'est ce qu'on récolte quand on fait une manip. »	Masse et taille des oiseaux Taille de ponte Données météorologiques Concentrations en métaux lourds de prélèvements de sols
Chercheur 18	Neurosciences	« Le produit de l'expérimentation, qu'il soit qualitatif ou quantitatif »	Données de mesure du comportement des rats Données d'histologie
Chercheur 22	Neurosciences	« Toute information qui peut conduire à un résultat » « ce qu'on interprète, ce qu'on analyse, ce qu'on manipule »	Données de mesure du comportement des souris Données d'histologie Données de séquençage
Chercheur 11	Chimie	« Les résultats scientifiques qu'on obtient des différentes expériences faites dans le laboratoire »	Formule chimique du matériau élaboré Données de mesure du matériau
Chercheur 1	Génétique	« Quelque chose d'informatique » « Quelque chose qui a une composante chiffrée »	Données de spectrométrie Données de séquençage

Tableau 12 : Définitions de « donnée de recherche » proposées par les chercheurs interrogés

2.3. Valeur des données dans un écosystème centré sur la publication scientifique

2.3.1. Publier pour gagner en « crédibilité »

Malgré les possibilités qu'ouvrent les technologies numériques de l'internet et du web, la publication reste le modèle prédominant en matière de communication scientifique. La science, telle que représentée par ce modèle, est une science qui se lit et s'écrit. Une science dans laquelle le chercheur se considère comme un « producteur de faits » et cherche à convaincre ses pairs que ses énoncés sont des faits. C'est ce qu'observent les sociologues Bruno Latour et Steeve Wooglar au milieu des années 1970, pendant les deux années qu'ils passent en immersion dans un laboratoire de neuroendocrinologie de San Diego en Californie (Latour et Woolgar 1979).

Or, si les chercheurs estiment être des producteurs de faits, pourquoi ne donnent-ils pas davantage de visibilité aux données de recherche, qui sont la preuve de leur raisonnement scientifique ? Probablement, en partie, parce qu'ils sont devenus tributaires d'un système, où la publication conditionne leur réussite professionnelle et concentre leur attention.

2.3.1.1. Le cycle de crédibilité

Bruno Latour (2001) associe la publication à un investissement réalisé par le chercheur, dans la perspective de gagner en « crédibilité ». La crédibilité scientifique recouvre les bourses, les postes, les titres honorifiques, mais aussi la confiance accordée par les pairs et la bonne réputation auprès des organismes de financement. Elle est conçue pour être réinvestie, comme le montre le cycle de crédibilité imaginé par Latour (figure 11) : la publication d'articles et d'ouvrages permet d'obtenir de nouvelles subventions, qui seront investies dans l'achat de matériel et le recrutement de personnel (ingénieurs, doctorants, post-doctorants…) ; ces derniers permettront de générer de nouvelles données, grâce auxquelles pourront être formulées de nouvelles théories, qui elles-mêmes seront publiées dans un article par exemple ; et ainsi de suite. Ce cycle de crédibilité rend compte des mécanismes permettant à un

scientifique de continuer à faire des recherches. Il explique à la fois la logique d'ensemble du développement scientifique et les comportements des chercheurs.

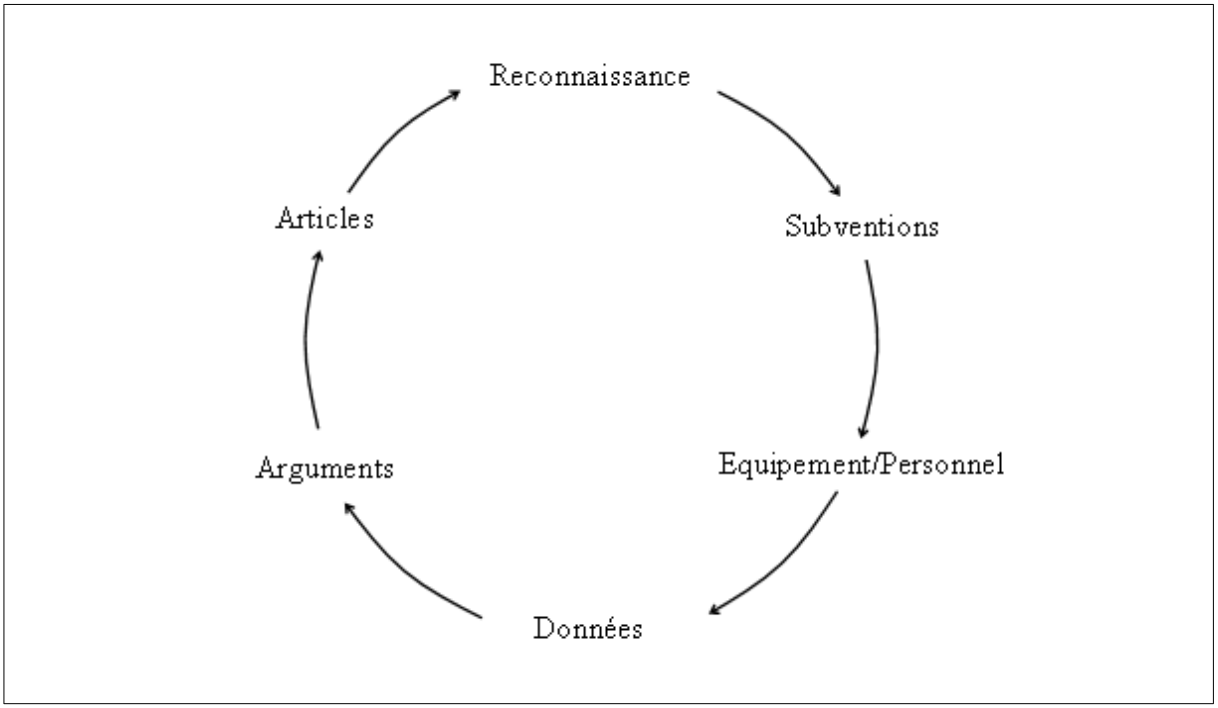


Figure 11 : Cycle de crédibilité selon B. Latour[252]

2.3.1.2. Un mécanisme plus ou moins perceptible selon les domaines de recherche

L'analyse des entretiens a confirmé l'influence de ce modèle sur l'activité scientifique des chercheurs interrogés, avec de sensibles différences en fonction de leur discipline d'appartenance.

Le cycle de crédibilité est d'autant plus à l'œuvre en sciences, technologies et médecine (STM) que toute nouvelle recherche nécessite un apport financier. Dans des disciplines comme la chimie et la biologie, l'acquisition de données dépend en effet souvent de matériel coûteux.

Lorsque la recherche passe par l'étude de modèles animaux (souris, rats…), elle nécessite des frais d'animalerie sur une durée relativement longue – l'étude étant soumise au rythme

252 Source : Latour 2001, p.34

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

Obtenir des financements constitue donc une des principales préoccupations des équipes de recherche. Il revient notamment au chef d'équipe de consacrer du temps à répondre à des appels à projets, afin d'obtenir de nouveaux financements.

> *« De temps en temps, je mets un peu la main à la pâte, mais je n'ai pas beaucoup, beaucoup le temps. Je vais un petit peu au microscope, je fais ce genre de choses. Sinon, je coordonne les différentes parties. Et puis, là, j'ai repris les rédactions de demandes de financement. »* (chercheur 22)

Or, selon le cycle de crédibilité, il n'y a pas de financement sans crédit-reconnaissance et pas de reconnaissance sans publication. La publication d'articles dans des revues prestigieuses constitue donc un enjeu clef pour obtenir un financement et faire en sorte que le travail de recherche se poursuive. Comme nous l'expliquait le chercheur 18, *« si [une équipe de recherche] tombe sur un résultat majeur, [elle va] chercher à le publier dans un journal de forte réputation et, par ce biais-là, [elle va] crédibiliser de futures demandes de financement »*.

En sciences humaines et sociales, la logique du cycle de crédibilité est moins visible, car les projets de recherche ne bénéficient pas systématiquement d'un financement spécifique. D'une part, les subventions dévolues aux sciences humaines et sociales sont moins nombreuses que celles destinées aux STM. D'autre part, le coût d'accès aux sources et outils de collecte de données reste généralement abordable (comparé à celui des STM) et peut être pris en charge sur le budget des laboratoires. Les projets de sciences humaines et sociales mobilisent

généralement des sommes inférieures à celles des projets de sciences et technologies ou de vie et santé. Pour des projets de durée égale (3 ans), les subventions allouées par l'ANR au projet du chercheur 52 en sciences sociales étaient par exemple 2,5 fois moins élevées que celles accordées au projet du chercheur 2 en biochimie (198 501€ contre 493 797€). La logique de crédibilité existe pourtant bel et bien en sciences humaines et sociales. Elle se manifeste au travers de l'évaluation du Haut Conseil de l'Évaluation de la Recherche et de l'Enseignement Supérieur (Hcéres), dont dépendent les subventions accordées par l'État aux laboratoires et universités.

2.3.1.3. Faible valeur d'échange des données

Fecher et al. (2015) qualifient le système d'échange de l'information scientifique d'« économie de la réputation » (*reputation economy*). La production scientifique n'est partagée que si elle permet au chercheur qui la transmet d'accroître sa réputation auprès des pairs. La réputation d'un chercheur étant principalement liée à ses publications (articles de revues, monographies, actes de colloques…), celui-ci ne trouvera que peu d'intérêt à diffuser d'autres produits de ses recherches – les données scientifiques en particulier.

Au cours des entretiens, les chercheurs ont dit avoir peu de temps et de moyens à consacrer à la gestion des données. Un chercheur évoquait par exemple le problème de l'obsolescence des supports de sauvegarde à propos d'anciennes données stockées sur CD Rom : *« Le problème c'est que les transférer sur de nouveaux supports, ça prend du temps et l'un de nos principaux facteurs limitants ça reste quand même le temps »* (chercheur 6). La diminution des ressources humaines et financières est certes une réalité dans la recherche publique. Mais derrière l'argument du manque de temps il faut voir une logique de priorité. Quand les chercheurs disent qu'ils n'ont pas les moyens de s'occuper plus avant des données, c'est que cette tâche n'est pas une priorité pour eux. Leur priorité va naturellement aux activités qui leur permettent d'obtenir une bonne évaluation et de concrétiser de nouvelles recherches – en l'occurrence la publication. Comme l'expliquait un chercheur : *« La seule mesure qu'on a pour passer les concours et devenir directeur de recherche, c'est malheureusement les publications. C'est le seul outil qu'on opère pour pouvoir juger si oui ou non on est un bon chercheur et si on peut avoir une promotion. Donc il faut qu'on publie. Et on est en*

concurrence avec d'autres labos. Le premier qui a trouvé quelque chose publie bien ; celui qui trouve le même résultat mais deux ans après, il ne publie pas bien ou il ne publie pas. Donc il faut à un moment être le premier à avoir ce résultat » (chercheur 9).

Or, selon Fecher et al. (2017), dans ce système de communication de l'information scientifique, donnant la priorité à la publication, les données de recherche ne possèdent qu'une faible valeur d'échange. D'où un impact sur leur diffusion et leur gestion. Il est en effet rare que les fichiers de données soient conservés sur le long terme, notamment lorsque surviennent des problèmes d'encombrement.

> *« On stocke [le matériel histologique] jusqu'à ce que le travail soit complètement valorisé. Après, on a le même problème que tout espace qui n'est pas plastique, à savoir qu'au bout d'un moment il faut qu'on gère de l'encombrement. Et la meilleure façon de gérer l'encombrement, c'est de jeter. Donc, tant que ça n'est pas publié ou que le matériel garde un intérêt, on conserve à la fois les données générées et le matériel brut. Quand c'est publié, en règle générale, on s'en débarrasse au bout de quelques années. »* (chercheur 18)

Seules les données sous-jacentes à une publication font l'objet d'une conservation plus assidue, avec souvent triple ou quadruple support de sauvegarde. Elles sont conservées pour leur valeur de preuve : en cas de revendication des faits énoncés dans la publication, l'auteur pourra justifier son raisonnement en utilisant ces données comme preuve à l'appui.

Dans le système de communication scientifique actuel, les données sont subordonnées aux publications. Leur rôle est de justifier le raisonnement scientifique, dont il est fait démonstration dans la publication. Quel que soit leur domaine, les chercheurs sont contraints de publier s'ils souhaitent obtenir une bonne évaluation et ainsi continuer d'être financés pour ce qu'ils font. En termes d'échanges, la publication a donc davantage de valeur que les données. Quand ils publient des articles et ouvrages, les chercheurs reçoivent en retour (certes de manière indirecte) de la crédibilité. Partager les données, en revanche, attire peu la reconnaissance des pairs – du moins dans l'évaluation institutionnelle de la recherche, où seuls le nombre et le prestige des publications comptent (Schöpfel et al. 2017b).

2.3.2. La valeur indirecte des données : une valeur d'usage

Il semblerait que les données possèdent malgré tout une valeur. Au cours des entretiens, les chercheurs ont tous témoigné d'un attachement indéniable aux données qu'ils avaient collectées. A la question « Quelle valeur ont les données que vous avez générées ? », un chercheur a répondu : *« Ce sont nos enfants ! »* (chercheur 3). Cette image de l'affiliation des données à leur producteur sous-entend un lien fort, lié vraisemblablement aux efforts investis par le chercheur pour les acquérir, les mettre en forme, les analyser et les conserver.

2.3.2.1. Différentes perceptions de leur valeur

Cette valeur reste cependant difficile à saisir. Les chercheurs ne sont pas unanimes à ce sujet : pour certains, les données valent de l'argent, du temps, pour d'autres des connaissances…

En biologie par exemple, les chercheurs considèrent certaines données comme précieuses, du fait de leur coût d'acquisition. L'un d'eux expliquait ainsi : *« pour les données très sensibles comme celles-là (sensibles parce que c'est 1500€ le séquençage), on a un triple stockage »* (chercheur 3).

On retrouve cette même considération en neurosciences. Un des chercheurs disait : *« la valeur qu'on peut donner aux données, c'est toute la masse d'argent qui a été investie pour les obtenir. Par masse d'argent, je pense aux animaux, aux réactifs, aux salaires... »* (chercheur 22).

En géochimie, pour un des chercheurs rencontrés, c'est la question du volume de travail qui détermine la valeur de certaines données : *« Typiquement, quand on fait une mesure isotopique, on ne va pas la mettre sur internet le lendemain. Parce qu'obtenir un rapport isotopique, ça représente des mois de travail »* (chercheur 37). Le calcul d'une mesure isotopique demande un travail d'acquisition trop important pour que celui qui l'a réalisé se permette de le mettre aussitôt à la disposition de tous. Plus la donnée est complexe à acquérir, moins le chercheur sera enclin à la rendre librement consultable.

La diversité des propos tenus témoigne de la relativité des perceptions. La valeur de la donnée dépend de l'utilisation qui en est faite. On qualifiera cette valeur de « valeur d'usage ». Un

bien est défini comme valeur d'usage lorsqu'il est considéré dans un rapport d'utilité. Il peut avoir une valeur d'usage pour nous-mêmes ou pour autrui. Autrement dit, c'est l'utilisateur, par la consommation du bien, qui lui confère une valeur d'usage. Ce concept a été étudié par l'économiste Anne Mayère dans le domaine de l'information (Mayère 1990, chapitre III). L'objectif de cette chercheuse est de sortir l'information du statut économique de marchandise, en définissant sa valeur d'usage selon un critère d'utilité – l'information s'intègrant toujours dans un processus déterminé par les événements ou les décisions de l'utilisateur.

2.3.2.2. Une valeur pour soi

Dans le cas des données scientifiques, l'attachement des chercheurs aux données qu'ils ont collectées est une valeur pour soi (non pour leurs pairs ou pour toute autre personne). Cette valeur réside dans le fait d'utiliser les données pour publier un nouvel article. C'est ce que précisait un chercheur en géographie : *« Il y a un aspect un peu « précieux » de la donnée, qui est qu'on ne la montre pas avant de l'avoir soi-même exploitée. C'est tout l'enjeu de la compétition scientifique »* (chercheur 45). Les données sont une matière première essentielle à l'article. Car c'est sur elles que se fonde le raisonnement scientifique. Tant que l'article (ou la communication ou le chapitre d'ouvrage) n'est pas paru, les données conservent leur valeur d'usage. Si le chercheur perd ses données, il perd aussi la possibilité de publier et d'obtenir en échange la reconnaissance de ses pairs. C'est dans l'utilisation des données (pour la publication) que le chercheur trouve le retour sur investissement du travail de collecte. L'enjeu est donc d'autant plus grand que la donnée a eu un coût d'acquisition élevé.

2.3.2.3. Variation de la valeur de la donnée

<u>En fonction de son degré de convoitise</u>

Par ailleurs, plus une donnée est convoitée par autrui, plus sa valeur d'usage augmente. En sciences, techniques et médecine notamment, les connaissances évoluant très vite, il est fréquent que plusieurs équipes (à l'échelle internationale) travaillent sur le même sujet. Cette situation génère de la compétition : il faut être le premier à publier (étant donné l'exigence d'originalité de la communauté et des éditeurs scientifiques). Tout l'enjeu pour ces équipes

est de garder leurs données confidentielles, afin d'éviter qu'une équipe « adverse » ne s'en empare et ne publie avant elles. Il y a donc un réflexe spontané de protection des données : celles-ci doivent rester confidentielles, tant que l'article n'a pas été publié.

> *« En colloque, la politique est de ne montrer que des choses déjà publiées ou déjà sous forme d'article qu'on va soumettre. On a pris le parti pris de ne pas présenter des choses non publiées. Parce que, dans les colloques, il y a forcément nos compétiteurs. Il suffit qu'ils notent la bonne idée et qu'ils refassent la manip en deux temps, trois mouvements. Ils vont publier avant nous une histoire moins complète et, nous, on se sera fait squeezer. »* (chercheur 1)

Le même mécanisme est à l'œuvre quand il s'agit de breveter une connaissance ou une technologie. Les données, susceptibles d'intéresser l'industrie, sont gardées confidentielles jusqu'au dépôt du brevet. Cette précaution permet de préserver le caractère inédit de la technologie[253]. Dans ce cas, les données ont donc une valeur à la fois scientifique, industrielle et économique.

<u>En fonction du degré d'avancement de son producteur</u>

La valeur d'usage des données peut varier également au fil de la carrière du chercheur. Elle est fonction du degré d'avancement professionnel de ce dernier. Pour les doctorants et jeunes chercheurs, les données constituent un capital symbolique qu'ils mobiliseront pour se faire connaître de la communauté scientifique et obtenir un poste de chercheur.

> *« Je connais beaucoup de collègues allemands qui n'ont pas de position fixe. En Allemagne, les positions fixes sont très rares. […] Je connais des gens qui ont 45 ans et qui n'ont pas encore de position fixe. Donc il y a un enjeu économique derrière, mais pour eux, pour leur propre vie. C'est-à-dire qu'il faut qu'ils publient un maximum. Du coup, ils sont un peu moins partageurs sur certaines données. On sent qu'il y a un enjeu personnel pour eux. Alors que quand on a une position permanente, on n'a plus forcément cet enjeu personnel. »* (chercheur 16)

253 Voir infra, quatrième partie, 2.5.2.2, p.197

Cet aspect a été mis en évidence dans les études de Van den Eynden et al. (2016) et de Tenopir et al. (2015a). Ces dernières montrent que les jeunes chercheurs sont moins disposés à partager leurs données (en comparaison de leurs collègues plus avancés dans leur carrière), car ils craignent de perdre des opportunités de publications futures en rendant leurs données disponibles.

2.4. Traitement des données (collecte, analyse, interprétation) : des méthodes et outils variés et évolutifs

Cette partie esquisse les grandes tendances qui, pour chaque discipline, se sont dégagées des entretiens en termes de collecte et d'analyse de données.

En droit, histoire et théologie

Les juristes, historiens et théologiens travaillent essentiellement à partir de « sources », qu'ils trouvent dans des bibliothèques physiques ou numériques, ainsi que dans des centres d'archives publics ou privés. L'analyse de ces documents consiste le plus souvent en une approche herméneutique[254]. Les chercheurs disent ne pas employer le terme de « données » dans leurs travaux. Deux d'entre eux faisaient néanmoins exception (chercheurs 55 et 56). Chercheurs en théologie, ils travaillaient tous deux à partir de bases de données. L'usage du terme « donnée » leur venait donc relativement spontanément.

En sciences sociales

En sciences sociales, nous avons rencontré des anthropologues, des sociologues et des démographes. Les anthropologues et les sociologues sont assez solitaires dans la collecte et le traitement des données. La nature qualitative de l'analyse leur demande en effet d'avoir une certaine intimité, du moins une certaine proximité avec leur terrain. En démographie, la collecte de données est plus « collective ». Travaillant à partir d'enquêtes quantitatives, il est

254 Voir supra, première partie, 1.3.1, p.28

fréquent que plusieurs chercheurs se rassemblent autour d'une grande enquête à mener. Dans cette discipline, il existe également des pratiques de partage et de réutilisation des données. En France, elles sont structurées autour du réseau Quetelet, un service donnant accès à des banques de données quantitatives, issues de la statistique publique par exemple.

En écologie

Les écologues récoltent des données sur le terrain à propos d'espèces végétales ou animales. Il leur arrive de consulter des bases de données de référence, de taxonomie par exemple (chercheur 41). Parfois ils ont recours à des plateformes de séquençage d'ADN (chercheurs 15 et 41). Plus rarement, ils répondent à des appels à projet de l'Institut Polaire Français pour avoir accès à des bases scientifiques en Arctique et Antarctique, d'où ils peuvent étudier des populations animales (chercheur 13).

En chimie et biologie

Dans ces domaines, l'acquisition des données se fait principalement à la « paillasse » dans le laboratoire. Les chercheurs font appel à des instruments relativement sophistiqués et coûteux. Lorsqu'ils n'ont pas la chance de disposer d'un instrument, ils font appel à d'autres laboratoires de recherche ou bien sous-traitent l'acquisition des données à une plateforme technologique dédiée.

En astronomie

En astronomie, ont été rencontrés des théoriciens et des observateurs. Les observateurs utilisent les données générées par les satellites et les télescopes internationaux : ils les analysent et les interprètent. Les théoriciens, quant à eux, travaillent plutôt à l'élaboration de modèles numériques visant à recréer un phénomène astrophysique. Ils utilisent en entrée de leurs modèles les données d'observation délivrées par les satellites et les télescopes. En sortie, ils obtiennent des données de simulation, qu'ils comparent aux données d'observation pour vérifier la justesse de leur modèle.

En sciences de la Terre

Le même constat a été fait pour les sciences de la Terre, avec d'un côté des chercheurs qui analysent des données de terrain et de l'autre des chercheurs qui modélisent des phénomènes.

En sciences de l'ingénieur

En sciences de l'ingénieur, ont principalement été rencontrés des chercheurs en informatique, travaillant soit dans le domaine de la bioinformatique, soit dans celui de l'imagerie médicale, soit dans celui de la télédétection. Tous avaient pour point commun d'être confrontés à de grands volumes de données, que leur discipline leur demande de traiter (à la place de l'humain qui n'en est plus capable). Leur laboratoire fournit les dispositifs de stockage et de calcul qui leur sont nécessaires. Ils ont recours à des centres de calcul uniquement lorsqu'ils ont besoin de puissances très importantes.

En géographie

Les géographes sont eux aussi confrontés à d'importants volumes de données. Pour l'acquisition des données, ils utilisent des instruments de type capteurs, souvent très coûteux à l'achat. Lorsqu'ils ont besoin d'un équipement spécifique, ils établissent des collaborations avec des collègues disposant de l'instrument.

2.5. Gestion des données (stockage, curation, conservation...)

Dans cette partie sont envisagés différents facteurs pouvant expliquer la variation des modes de gestion de données, décrits par les chercheurs au cours des entretiens.

2.5.1. Premier facteur de variation : les ressources consacrées à la gestion des données dans le laboratoire ou le projet de recherche

2.5.1.1. Inégalité des laboratoires

La gestion des données varie en fonction des ressources à disposition dans le laboratoire ou le projet de recherche.

L'enquête semble montrer que les équipes de recherche les mieux dotées en moyens techniques et humains (présence d'ingénieurs en charge de la gestion des données, mise à disposition de supports de stockage partagés et répliqués...) sont celles où la sécurité, l'organisation et la conservation des données sont le mieux assurées. Dans la majorité des laboratoires, cependant, les scientifiques font état d'une diminution des ressources allouées à la recherche publique. Les unités de recherche disposent de ressources humaines de plus en plus restreintes (de moins en moins d'ingénieurs, de techniciens et de personnels administratifs) et de moyens informatiques souvent insuffisants pour leurs besoins en matière de recherche. Ces restrictions ont un impact sur la gestion des données.

- Les fichiers de données ne sont ni classés ni documentés à hauteur de ce que souhaiteraient les chercheurs, faute de temps et en l'absence de ressources humaines dédiées. *« On n'a pas tout le back-office qui devrait être là, pour pouvoir documenter ces informations et les mettre assez facilement à la disposition de la communauté. Ça c'est quelque chose qui est rare. En tant qu'enseignant-chercheur, on nous demande de plus en plus de choses de ce côté-là, qu'on n'a pas franchement le temps de faire. Il y a un sérieux problème de temps et de personnel dans ce domaine. »* (chercheur 45)
- Les fichiers de données ne sont, par ailleurs, pas forcément répliqués sur plusieurs supports de sauvegarde, entraînant des risques de perte (fausse manipulation, vol, incendie…).
- Enfin, lorsque les espaces de stockage sont saturés, à défaut de pouvoir les augmenter, les chercheurs sont contraints d'effectuer un tri parmi les données et d'en supprimer certaines. Dans le laboratoire 1 (chercheur 1), les équipes de recherche disposent d'un quota de plusieurs gigaoctets pour stocker leurs données gratuitement. Au-delà, le

stockage est payant. Ce coût supplémentaire ne joue donc pas en la faveur des données, qui se verront supprimées en cas d'aléas budgétaires au sein de l'équipe.

Les chercheurs sont conscients de ces lacunes en termes de stockage, de sécurité et d'organisation des données. Ils savent généralement ce qu'il faudrait faire pour en assurer une meilleure gestion.

> *« Il faudrait donner des noms aux fichiers, les ranger proprement dans des répertoires, se souvenir qu'il s'agit de tel projet, écrire à chaque fois un petit fichier qui va avec, avec des métadonnées pour se souvenir d'où ça vient et ce que c'est... »* (chercheur 33)

Néanmoins, comme les préoccupations des chercheurs sont principalement tournées vers la publication, ces problèmes restent souvent au second plan et les besoins sont rarement clairement exprimés. Les chercheurs tentent de trouver des solutions par eux-mêmes, avec les moyens dont ils disposent dans leur laboratoire ou sporadiquement grâce aux subventions des projets de recherche. Les paragraphes suivants rapportent quelques unes des solutions mises en place par les chercheurs rencontrés pour stocker, sécuriser et organiser leurs données.

<u>Pratiques de stockage des données</u>

Pour la majorité des chercheurs rencontrés, les données sont peu volumineuses. Les espaces de stockage à disposition (PC professionnel, serveur partagé du laboratoire ou de l'équipe, cloud universitaire) ne suffisent néanmoins pas pour conserver les données de toute une carrière. Les chercheurs procèdent donc à des actions de « nettoyage » régulières. Ils considèrent qu'acquérir de l'espace de stockage supplémentaire n'en vaudrait pas la peine (en termes de coûts notamment). Les données datant d'il y a plus de dix ans sont considérées comme suffisamment anciennes pour pouvoir être supprimées. Il y a, selon eux, peu de chance pour que ces données soient réutilisées. Comme le soulignait un chercheur en biologie : *« On n'essaie pas de résoudre ces problèmes d'encombrement avec un stockage externe. En fait, ça ne nous servirait à rien. On ne va pas stocker des données d'étudiants, qui sont passés il y a dix ans, sur des résultats qui ont été analysés, interprétés et souvent publiés »* (chercheur 9).

La question du stockage est beaucoup plus présente dans les équipes, où les données acquises sont volumineuses (une image de télédétection, par exemple, a une taille d'environ 1 Go). Dans la plupart des cas observés (chercheurs 1, 3, 6, 17, 20, 23, 24, 25, 26, 27, 28, 29, 33, 34, 43 et 56), les chercheurs résolvent ce problème à l'échelle du laboratoire (voire plus rarement à l'échelle de l'université, en ayant recours aux services des mésocentres ou *datacenters*). Les laboratoires consacrent une part de leur budget à l'acquisition de serveurs de stockage (voire de calcul). Souvent on constate que les équipes en question font partie de grands laboratoires (divisés en départements et composés d'un grand nombre d'équipes), souvent nés de la fusion entre plusieurs unités de recherche. C'est le cas par exemple du laboratoire du chercheur 1. Spécialisé dans la recherche biomédicale, ce laboratoire se compose de 51 équipes scientifiques réparties en quatre départements, associant biologie, biochimie, physique et médecine. Il est issu de la fusion en 1994 de deux laboratoires strasbourgeois de biologie. Ce type de regroupement semble être une façon de mutualiser les moyens. Car la possession de serveurs est relativement coûteuse, comprenant non seulement l'achat de machines, la mise en place d'un environnement adapté (salle dédiée, système de climatisation…), mais aussi des personnels dédiés.

Pratiques de sécurisation des données

Les chercheurs sont soucieux d'éviter toute perte des données dont ils ont encore l'intention de se servir. Cela les conduit à répliquer ces données sur différents supports (serveur du laboratoire, disque dur, cloud…). Comme le résumait l'un d'entre eux : *« Les paranos comme moi ont un disque au labo, un disque à la maison et un disque sur eux. Comme ça, on ne risque jamais rien ! »* (chercheur 3).

On observe une pratique différenciée en fonction du type de données. Les chercheurs vont en effet privilégier certains types de données, selon des considérations de volume et de réplicabilité. Plus il est difficile et chronophage de reproduire les données, plus les chercheurs vont préserver ces données, en multipliant autant que possible le nombre de copies. En revanche, lorsque les données sont volumineuses, le nombre de sauvegardes se limite à deux copies maximum. C'est ce qu'expliquait un chercheur en astronomie (chercheur 29) :

> *« Ici, à l'observatoire, on a un stockage local. On a des serveurs sur lesquels on peut mettre les données. Moi je stocke tout ici à l'observatoire. Les données de simulation ne sont pas répliquées plusieurs fois. Je les ai juste sur une voire deux machines maximum, parce que ça peut prendre très vite beaucoup de place. En revanche, les codes de simulation, qui me permettent de faire les simulations, ceux-là je les ai dupliqués plusieurs fois (sur mon PC, sur un portable, sur un autre serveur de l'observatoire et sur Seafile, le serveur de l'Université). [...] Les données sont moins importantes que le code. Au pire, on peut refaire tourner les simulations pour ré-obtenir les données. »*

Les supports de sauvegarde diffèrent en fonction des laboratoires, de leur budget et de leur politique interne. Comme le soulignait l'un des chercheurs (chercheur 43) : *« stocker les données demande un investissement humain et financier, que tous les laboratoires ne peuvent pas faire »*. Dans les unités de recherche, les personnels disposent a minima d'un ordinateur professionnel, sur lequel ils peuvent stocker des données. Dans certains laboratoires, principalement en STM, il existe des serveurs de stockage partagés, où les chercheurs peuvent sauvegarder et dupliquer leurs fichiers de données. Les laboratoires offrant des solutions de stockage sécurisé avec une double sauvegarde sur sites géographiques distants sont relativement rares (laboratoires 1, 8 et 9). A l'Université de Strasbourg, les chercheurs bénéficient par ailleurs d'un espace de stockage sécurisé de 100 Go sur le cloud institutionnel Seafile.

Ces ressources informatiques sont considérées comme insuffisantes par la majorité des participants à l'enquête, pour garantir une sauvegarde sécurisée des données. Cette lacune conduit certains chercheurs à recourir à des solutions personnelles. Ils dupliquent leurs données sur des supports financés à leurs frais, comme ce chercheur en neurosciences :

> *« Ce système de sauvegarde ne me convient pas tellement, parce que le problème c'est qu'on a effectivement plusieurs copies, mais toutes physiquement au même endroit. Donc, moi ce que je fais, c'est que je fais un back-up régulier de mon ordinateur sur un disque dur que j'ai à la maison. »* (chercheur 22)

Pratiques d'organisation des données

Plusieurs chercheurs ont évoqué les faiblesses de leur système d'organisation et de documentation des fichiers de données. L'organisation et la documentation des données est quelque chose qui, selon eux, demande du temps et des compétences. Six chercheurs (chercheur 1 en génétique, chercheur 6 en chimie et biologie, chercheur 33 en informatique, chercheur 37 en géologie, chercheur 45 en géographie et chercheur 48 en démographie) pointent clairement du doigt le manque de ressources humaines. Une telle organisation demanderait les compétences d'un ingénieur dédié. Or il est très rare qu'un ingénieur soit officiellement chargé de la gestion des données. Dans notre enquête, seul le laboratoire de géographie faisait exception, avec un ingénieur d'études dédié à temps partiel à la gestion des données.

En sciences, technologies et médecine, cela est d'autant plus compliqué que plusieurs personnes travaillent sur les mêmes données. Quatre chercheurs (chercheurs 3, 33, 37 et 38) admettent qu'idéalement il faudrait instaurer des règles communes d'organisation des dossiers et de nommage des fichiers. Par ailleurs, l'équipe de recherche est souvent composée de deux à trois personnels permanents seulement (chercheurs et ingénieurs/techniciens) et de plusieurs doctorants et post-doctorants, qui ne resteront que le temps de leur contrat. Cela conduit le chercheur titulaire à gérer seul les données acquises, traitées et analysées par l'ensemble de l'équipe et à veiller, de surcroît, à ce qu'aucun doctorant ou post-doctorant ne parte avec les données qu'il a générées.

> *« Chaque doctorant va travailler sur un projet. Mais, quand le doctorant part, il ne faut pas que les données partent avec lui. Donc il faut mettre en place une politique de conservation et d'archivage des données. »* (chercheur 37)

Cet enjeu est d'autant plus présent que les doctorants et post-doctorants sont généralement les principaux opérateurs de la collecte de données.

Davantage de ressources humaines permettraient d'assurer une gestion plus systématique et formalisée des données, notamment en ce qui concerne la documentation, l'organisation et le nommage des fichiers, le suivi des espaces de stockage et la sécurité des données.

2.5.1.2. Nouvelles logiques de financement de la recherche

La gestion des données est impactée par un contexte de précarisation de la recherche publique. Un des facteurs de précarisation est l'instauration d'une politique de financement sur projets. Cette dernière a pris une dimension institutionnelle à partir de 2005 avec la création de l'Agence Nationale de la Recherche (ANR). Placée sous la tutelle du Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, l'ANR a vocation à mettre en œuvre le financement de la recherche sur projets. Pour cela, elle propose des appels à projets compétitifs et sélectionne les candidats sur la base de l'évaluation par les pairs.

Jusqu'alors, le principal mode de financement de la recherche était constitué de budgets alloués aux laboratoires par leurs tutelles. Ces budgets sont dits « pérennes », dans la mesure où ils sont reconduits en fonction des évaluations scientifiques auxquelles sont soumis les laboratoires tous les cinq ans (Hcéres). Ils n'ont certes jamais été les seules sources de financement de la recherche : celui-ci est aussi fondé sur des contrats passés avec l'industrie, les associations ou les collectivités locales. Mais la création de l'ANR s'est accompagnée d'un rétrécissement des financements pérennes de la recherche, avec en particulier la suppression de postes statutaires.

La logique des appels à projets a notamment été dénoncée par le collectif P.é.c.r.e.s. (2011). Premièrement, elle entre en conflit avec la temporalité longue de la recherche. *« Il faut du temps pour élaborer intellectuellement une recherche, puis pour la mettre en pratique, puis pour en analyser les résultats ; il faut souvent beaucoup de temps pour qu'un résultat de recherche débouche finalement sur une innovation* ». Les financements pérennes sont considérés comme les seuls capables de s'accorder avec le rythme naturel de la recherche. Deuxièmement, la réponse aux appels à projets est une activité souvent chronophage, qui empiète sur le temps de la recherche. Le montage des dossiers est un processus long et complexe. Il demande d'anticiper les différentes phases du projet (protocole de recherche, modalités d'exécution, résultats attendus, organisation des équipes, modes de valorisation des résultats…), alors que la recherche est souvent faite de tâtonnements et d'infléchissements. Il requiert, par ailleurs, des compétences non scientifiques, d'ordre technique et budgétaire. Enfin, la remise d'un dossier de candidature ne donne pas la garantie que le projet sera retenu,

les budgets des agences de financement ne bénéficiant qu'à quelques uns. C'est donc parfois du temps perdu.

La gestion des données pâtit de ce nouveau mode financement. Elle est, de fait, réduite à l'échelle temps des projets de recherche. Le financement de moyens techniques et humains étant ponctuel, les données ne peuvent être conservées et valorisées sur le long terme. Dès lors que l'ingénieur recruté sur le projet a terminé son contrat, la gestion des données se poursuit rarement (les postes d'ingénieurs titulaires sont, de fait, de moins en moins nombreux dans les laboratoires).

2.5.2. Deuxième facteur de variation : le degré de sensibilité des données

Le mode de gestion des données varie également en fonction de leur degré de sensibilité. Certains types de données – que l'on regroupera sous le terme de « données sensibles » – sont soumis à des règles de gestion strictes, définies au niveau national voire européen.

Parmi ces données :

- Les données à caractère personnel, définies par la Commission nationale de l'informatique et des libertés (CNIL) comme *« toute information se rapportant à une personne physique identifiée ou identifiable »*[255]. En France, les données personnelles sont protégées par la loi.
- Les données sous-jacentes à un projet de recherche appliquée, censé aboutir à un transfert de technologie. Présentant un intérêt économique, ces données font l'objet de convoitise (espionnage industriel, concurrence scientifique). Elles sont donc gardées confidentielles par les équipes qui les produisent.
- Potentiellement, les données relevant du secret défense. Mais les chercheurs n'ont pas mentionné traiter ce type de données au cours des entretiens.

255 Source : https://www.cnil.fr/fr/glossaire

2.5.2.1. Gestion des données personnelles

Au cours des entretiens, 12 chercheurs ont dit utiliser des données personnelles dans le cadre de leurs recherches : deux démographes, un sociologue, un politiste, un anthropologue, un éthologue, un géographe, un informaticien, quatre biologistes et un médecin.

Quand les chercheurs utilisent des données personnelles, ils mettent en place une gestion plus structurée qu'avec les autres types de données qu'ils sont amenés à traiter. En cela, ils se conforment à des normes extérieures. En France, le traitement des données personnelles est réglementé par la loi Informatique et Libertés[256], mise en conformité avec le Règlement général sur la protection des données (RGPD)[257] entre 2018 et 2019. La CNIL[258] se fait le relai de cette législation sur les données personnelles, en fournissant des recommandations aux chercheurs qui collectent ou utilisent des données personnelles. Tout chercheur qui s'adresse à elle doit respecter certaines règles de gestion des données.

Les sous-parties suivantes montrent en quoi les différentes étapes du traitement des données sont impactées par ces réglementations.

<u>Accès et collecte de données personnelles</u>

Les chercheurs soit réutilisent des données collectées par des tiers, soit collectent eux-mêmes les données dont ils ont besoin.

256 *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000886460 (consulté le 20 septembre 2019).

257 *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. 119*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679 (consulté le 20 septembre 2019).

258 La Commission Nationale de l'Informatique et des Libertés (CNIL) est une autorité administrative indépendante. Créée en 1978, elle est composée d'un collège pluraliste de 17 commissaires, provenant d'horizons divers : 4 parlementaires, 2 membres du Conseil économique et social, 6 représentants des hautes juridictions et 5 personnalités qualifiées désignées par le Président de l'Assemblée nationale (1), le Président du Sénat (1) et le Conseil des ministres (3). Le mandat de ses membres est de 5 ans.

Réutilisation de données personnelles

Un des domaines réutilisant très fréquemment des données personnelles est la démographie. Dans cette discipline, les scientifiques s'intéressent aux caractéristiques et dynamiques des populations. Ils travaillent à partir de vastes enquêtes quantitatives (échantillons de 5 000 à 50 000 personnes). Produire des enquêtes de ce genre peut être très coûteux (jusqu'à plusieurs millions d'euros en fonction de l'échelle de l'enquête, locale ou nationale). Les budgets de recherche permettent rarement de réaliser de tels sondages. Il est donc plus fréquent que les chercheurs réutilisent des enquêtes produites par des instituts de statistique publique spécialisés (comme l'Insee ou l'Ined) voire par des fédérations de recherche (les 12 centres de recherche associés au Céreq par exemple).

Les données personnelles sont d'une très grande richesse pour la recherche. C'est pourquoi ce secteur d'activité bénéficie de dérogations légales, lui permettant de traiter plus facilement des données à caractère personnel[259]. Leur accès reste néanmoins soumis à des procédures très réglementées, visant à préserver l'intégrité des personnes. C'est un des aspects propres à la gestion des données personnelles dans les laboratoires de recherche.

Les instituts producteurs de grandes enquêtes, comme l'Ined ou l'Inserm, proposent aux chercheurs des données pseudonymisées, c'est-à-dire des données traitées *« de telle façon qu'[elles] ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable »*[260]. La pseudonymisation consiste par exemple à remplacer le nom d'un individu par un code ou par un faux nom. Une fois l'enquête réalisée et une fois la base de données nettoyée et codée, est créé un premier groupe d'exploitation. L'institut producteur de l'enquête sollicite plusieurs chercheurs spécialisés dans des thématiques particulières. Ceux-ci sont chargés d'analyser et de publier les premiers résultats de l'enquête. L'enquête sera ensuite disponible via PROGEDO et son réseau Quetelet. Tout chercheur ou doctorant ou étudiant en master 2 (dans le cadre de son mémoire) peut faire une demande de données. Il doit pour cela justifier d'un projet de

259 Article 4 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

260 *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016*, op. cit., article 4

recherche précis. Si l'accès lui est accordé, il doit signer une charte, par laquelle il s'engage à utiliser les données à des fins exclusives de recherche et d'enseignement, selon le projet de réutilisation qu'il aura décrit dans sa demande d'accès.

Dans certains cas, notamment lorsqu'il s'agit de données sensibles au sens de l'article 8 de la LOI n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles, le niveau de sécurité doit être un cran supérieur. Les données sensibles sont des *« données à caractère personnel qui révèlent la prétendue origine raciale ou l'origine ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale d'une personne physique »* ; elles englobent également les données génétiques, les données biométriques, les données concernant la santé, ainsi que les données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique[261]. Pour qu'un chercheur puisse analyser les résultats d'une enquête comportant des données sensibles, il doit bénéficier dans son laboratoire de dispositifs de sécurité suffisants pour préserver la confidentialité des données. Pour avoir accès aux données d'Eurostat[262], par exemple, qui est la direction de la Commission européenne chargée de l'information statistique, les chercheurs doivent obtenir au préalable une double certification.

- La première certification concerne le laboratoire. Elle porte sur l'environnement de travail. Comme le résumait un chercheur démographe : *« il faut avoir une certification Eurostat, pour dire que le chercheur va travailler dans un espace sécurisé, que les données mises à disposition seront sécurisées sur un ordinateur avec un mot de passe »* (chercheur 48). Cette certification atteste donc de la sécurité d'accès aux données. Elle est délivrée à l'échelle du laboratoire, à toute unité de recherche qui en fait la demande.

- La seconde certification concerne le projet de recherche, dans le cadre duquel les données seront utilisées. Cette certification est délivrée individuellement à chaque chercheur qui en fait la demande, après examen de son projet de recherche (est examiné l'usage que le chercheur souhaite faire des données dans le cadre de ce projet).

261 Article 8 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

262 Eurostat est une direction générale de la Commission européenne chargée de l'information statistique à l'échelle communautaire (https://ec.europa.eu/eurostat/).

Collecte de données personnelles

Le chercheur peut aussi être lui-même à l'origine de la collecte de données personnelles, par exemple lorsqu'il réalise sa propre enquête. Dans ce cas, lui incombe des tâches supplémentaires liées à la protection des personnes sondées. Parmi celles-ci, effectuer une déclaration auprès de la CNIL : *« On doit faire une déclaration à la CNIL, en décrivant ce que l'on fait et en précisant qu'on le fait à des fins de recherche, et non dans un but commercial par exemple »* (chercheur 51). La loi précise en effet que les données personnelles ne peuvent être collectées que pour une finalité précise définie en amont et portée à la connaissance des personnes concernées[263]. Très souvent, le chercheur devra également recueillir l'accord préalable des personnes, par le biais d'un formulaire de consentement.

Analyse des données personnelles

Le contrôle de l'accès n'est pas la seule réglementation régissant l'utilisation de données personnelles à des fins de recherche.

Toujours selon la loi Informatique et Libertés, les données personnelles doivent être *« traitées de façon à garantir une sécurité appropriée [...], y compris la protection contre le traitement non autorisé ou illicite et contre la perte, la destruction ou les dégâts d'origine accidentelle, ou l'accès par des personnes non autorisées, à l'aide de mesures techniques ou organisationnelles appropriées »*[264]. La CNIL recommande donc vivement leur anonymisation ou, a minima, leur pseudonymisation. Un chercheur en sciences politiques expliquait ainsi : « *Là où on est susceptible de nous surveiller, c'est qu'il ne faut pas qu'on ait de fichiers nominatifs. Par exemple, si on a des entretiens, il ne faudrait pas que dans nos ordinateurs on ait le nom des gens qui ont dit ceci. Il faudrait qu'on anonymise tout de suite »* (chercheur 51). Ces règles de gestion rendent parfois malaisé le travail d'analyse des chercheurs. C'est le cas de la pseudonymisation des entretiens. Pseudonymiser des entretiens dès leur retranscription rend l'analyse plus compliquée : lorsque les noms des personnes sont codés, il est plus difficile de s'y retrouver. C'est ce qu'expliquait le chercheur 51.

263 Articles 4 et 5 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

264 Article 4 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

En géographie, un chercheur travaillant sur des données d'enquêtes et des données GPS racontait qu'il devait suivre des protocoles de sécurisation bien définis, à savoir : *« Toutes les données sont stockées sur des disques durs qui sont cryptés et qui ne sont pas accessibles par internet. Elles sont résidentes sur certains ordinateurs et pas sur tous les ordinateurs du laboratoire. Les analyses se font directement sur ces ordinateurs. Ce qui sort, ce sont uniquement les données traitées et analysées »* (chercheur 45).

Des contraintes s'appliquent également pour le partage des fichiers de données entre collègues travaillant sur le même projet. C'était le cas d'un chercheur en éthologie, étudiant le réseau social des personnes âgées, à partir des données d'une cohorte de l'Inserm. *« On ne peut pas partager les fichiers par internet, par exemple. Normalement, tout est sur CD ou sur clé, lesquels sont enfermés dans un placard à code, auquel seul B.C., le responsable de la cohorte, a accès. Quand on va chercher les données, B.C. nous donne le fichier. On a donc ensuite le fichier sur notre ordinateur, mais on ne peut pas le partager par mail »* (chercheur 16).

<u>Conservation des données personnelles</u>

Les données personnelles ne peuvent par ailleurs pas être conservées indéfiniment. Leur conservation est limitée à *« une durée n'excédant pas celle nécessaire au regard des finalités pour lesquelles elles sont traitées »*[265].

Un géographe, collectant des données GPS liées aux déplacements des participants à une enquête, relatait qu'il devait supprimer ces données une fois l'analyse réalisée.

> *« On les détruit une fois qu'on a calculé nos indicateurs synthétiques, dans un souci d'anonymat. Parce qu'avec des données GPS, on arrive à identifier assez facilement le lieu de résidence des personnes et puis on a aussi bien sûr l'ensemble des lieux d'activité de ces personnes. »* (chercheur 45)

Dans un billet de blog consacré aux conséquences du RGPD pour la recherche, Lionel Maurel explique toutefois que le principe de limitation de la durée de conservation fait l'objet d'une dérogation pour les traitements réalisés à des fins de recherche.

265 Article 4 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

> *« Les données peuvent être conservées au-delà de la durée qui a été nécessaire pour atteindre la finalité de recherche (par exemple, au-delà de la durée d'un projet de recherche déterminé) du moment qu'elles sont ensuite conservées uniquement pour être utilisées à des fins de recherche. »* (Maurel 2018a)

Cet aspect est notamment utilisé par les unités de service responsables de cohortes de population. Pour préserver le « capital » que représentent ces données collectées sur plusieurs années, des recherche doivent continuellement être menées dessus. Dans sa déclaration à la CNIL, le responsable d'une cohorte de l'Inserm (avec lequel collaborait un des chercheurs rencontrés) avait par exemple demandé à ce que les données soient stockées pendant 10 ans, au motif que de nouvelles recherches seraient planifiées au cours de cette période. *« On peut garder les données pendant cinq à dix ans, si d'autres travaux sont ré-effectués dessus. C'est une période qui est renouvelable, c'est-à-dire que, si un nouveau travail est effectué sur ces données, on peut prolonger à nouveau la conservation »* (chercheur 16). Les données de la cohorte peuvent donc être conservées tant que des recherches continuent à être menées. D'où l'intérêt de toujours renouveler les projets de recherche, afin de pouvoir conserver ces données.

<u>Diffusion des données personnelles</u>

En ce qui concerne la publication des données personnelles, le décret du 1er août 2018 précise que *« ces données ne peuvent pas être diffusées sans avoir été préalablement anonymisées »*[266]. Pour que ces données soient rendues publiques, il doit être impossible de pouvoir remonter à l'individu. Ce qui peut s'avérer difficile à obtenir. Les chercheurs disent plutôt publier des données sous la forme d'*« indicateurs synthétiques »* (chercheur 45) ou de *« données agrégées »* (chercheurs 16 et 47).

Ils rencontrent des difficultés, quand une revue leur demande de mettre à disposition les données brutes liées une publication. Les données brutes étant des données personnelles, ils ne peuvent accéder à la demande de la revue.

266 *Décret n°2005-1309 du 20 octobre 2005 pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=20034927FA69108CF64868188CDA7239.tplgfr31s_3?cidTexte=JORFTEXT000000241445&dateTexte=20180812 (consulté le 20 septembre 2019). Article 100-1

> *« [Les éditeurs] deviennent de plus en plus gourmands et exigeants, dans le sens où il faut de la transparence – ce qui est tout à fait logique et normal, pour voir quelle est la qualité de la recherche qui a été menée. Mais après, il y a quand même le côté protection des données des individus. Pour un des derniers articles qu'on a publié, on s'est battu longuement avec PLOS One, pour leur expliquer pourquoi on ne mettait pas à disposition un certain nombre d'informations, pour leur expliquer aussi pourquoi on détruisait les données GPS et pourquoi on ne conservait que des indicateurs synthétiques. »* (chercheur 45)

Particularité des données de santé

Les données de santé sont une catégorie de données personnelles à caractère sensible. Elles sont définies dans le Règlement général sur la protection des données (RGPD)[267] comme *« l'ensemble des données se rapportant à l'état de santé d'une personne concernée qui révèlent des informations sur l'état de santé physique ou mentale passé, présent ou futur de la personne concernée »* (raison 35). Leur traitement à des fins de recherche est encadré par le Code de la santé publique[268] et la loi « Informatique et Libertés »[269]. Les chercheurs amenés à collecter ou réutiliser des données de santé doivent entamer des démarches auprès de la CNIL et, selon les cas, auprès de l'Institut National des Données de Santé (INDS)[270] et d'un Comité de Protection des Personnes (CPP)[271]. L'avis préalable d'un comité de protection des personnes est notamment requis pour toute recherche impliquant la personne humaine[272].

267 *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016*, op. cit.

268 *Code de la santé publique*. https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006072665 (consulté le 20 septembre 2019). Articles L1121-1 à L1126-12

269 Articles 64 à 77 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

270 https://www.indsante.fr/

271 Les comités de protection des personnes (CPP) sont chargés d'émettre un avis préalable sur les conditions de validité de toute recherche impliquant la personne humaine. Ils se prononcent sur les conditions dans lesquelles le promoteur de la recherche assure la protection des personnes et notamment des participants, sur le bien-fondé et la pertinence du projet de recherche et sur sa qualité méthodologique. Les CPP sont agréés par la Ministre de la Santé pour une durée de 6 ans. Leurs membres sont nommés par le directeur général de l'Agence Régionale de Santé pour une période de 3 ans renouvelable et exercent leurs fonctions bénévolement. Source : https://www.ars.sante.fr/comite-de-protection-des-personnes-1

272 Les recherches impliquant la personne humaine sont définies par le Code de la santé publique comme étant des « recherches organisées et pratiquées sur l'être humain en vue du développement des connaissances biologiques ou médicales » (article L1121-1).

Parmi les chercheurs rencontrés, sept travaillaient à partir de données de santé (chercheurs 5, 7, 9, 10, 12, 33 et 45). Leurs projets de recherche étaient réalisés en partenariat avec des médecins hospitaliers. La collecte des données était réalisée par les médecins auprès de leur patientèle. Un chercheur relate les procédures, relativement nombreuses, préalables à la collecte de ces données :

> *« On passe devant différents conseils pour avoir des autorisations. On demande aussi des autorisations auprès d'un comité d'éthique de l'Université de Strasbourg (il ne s'agit pas du nouveau comité d'éthique de la recherche mais d'un comité d'éthique médical). La législation est assez importante dans ce domaine. »* (chercheur 45)

Si les chercheurs déterminent avec le médecin la nature des données dont ils ont besoin, ils ne prennent en revanche pas part à la collecte elle-même. Celle-ci est réalisée par le praticien dans le cadre d'une consultation traditionnelle avec le patient. Les données sont ensuite conservées à l'hôpital, duquel elles ne sont pas censées sortir, du moins sans avoir été anonymisées au préalable.

Les chercheurs semblent pouvoir facilement avoir accès au matériel biologique (échantillons de sang, coupes histologiques…). L'hôpital les leur transfèrent et ils peuvent ensuite les analyser dans leurs laboratoires. Pour ce qui est des données biographiques du patient (âge, lieu d'habitation, stade de la maladie…), les conditions d'accès sont beaucoup plus réglementées. Les médecins n'ont pas le droit de faire sortir les données personnelles de l'hôpital, sans l'autorisation de comités d'éthique. Or les chercheurs ont parfois besoin de ces données pour aller plus loin dans leurs analyses, en les corrélant par exemple avec les données qu'ils extraient du matériel biologique. Pour obtenir les fichiers de données et réaliser l'analyse dans leur laboratoire, les chercheurs doivent donc solliciter l'expertise d'un comité d'éthique, qui émettra un avis sur la manière d'anonymiser les données avant leur transfert.

L'appropriation des normes par les chercheurs

Les normes décrites ci-dessus semblent plus ou moins intégrées par les équipes. Elles sont souvent perçues comme contraignantes : *« on a des protocoles très contraignants, mais on les respecte »* (chercheur 45). Dans le domaine de la santé, deux chercheurs (chercheurs 12 et 33)

considèrent le passage par les comités d'éthique comme chronophage et empiétant sur le temps de la recherche. L'un le qualifie de cauchemardesque : *« Quand on a besoin [des données biographiques du patient], pour qu'ils nous les transfèrent, c'est un cauchemar. Parce qu'il faut absolument qu'on ne puisse pas retrouver la personne. Et rien que d'avoir l'âge, le sexe et l'adresse (enfin, la ville), c'est déjà presque trop. Parce qu'on peut faire des croisements ensuite avec d'autres bases de données, aller sur Internet chercher d'autres infos... Enfin, bref, c'est compliqué. Donc il y a vraiment des protocoles, c'est super strict. Pour anonymiser les données, ce sont des règles nationales. Chaque pays a ses règles pour dire : 'voilà, il faut faire ça, supprimer ça, ça, ça et ça des données ; sinon vous n'avez pas le droit de les sortir de l'hôpital' »* (chercheur 33). Ce chercheur et ses collègues s'arrangent donc pour ne pas sortir les données de l'hôpital et ainsi éviter le passage par les commissions de contrôle. Ils envoient à leurs collègues médecins les données qu'ils souhaitent croiser avec les informations personnelles des patients. L'analyse est alors réalisée par des biostatisticiens au sein de l'hôpital, sans que les données personnelles n'aient à en sortir.

Les chercheurs composent donc avec les normes de gestion que leur fixe le cadre législatif, choisissant de s'y soumettre ou non selon ce qui leur paraît le plus avantageux en termes de temps et de moyens. Deux disciplines semblent néanmoins se distinguer : l'anthropologie et la sociologie. Les chercheurs interrogés dans ces domaines (chercheurs 46, 52, 53 et 54) gèrent les données personnelles de la même manière que les autres données qu'ils collectent, c'est-à-dire librement, selon des principes qu'ils se sont fixés pour eux-mêmes. Aucun d'entre eux n'a d'ailleurs employé le terme de « donnée personnelle » au cours des entretiens. En sociologie, par exemple, les chercheurs peuvent être amenés à collecter des données personnelles (au cours d'entretiens par exemple). Pourtant, à la différence des démographes, ils ne font pas de déclaration à la CNIL et n'instaurent pas de règles particulières visant à assurer la sécurité des données (c'est en tout cas ce qui a été constaté au cours de l'enquête). Cette non-appropriation des normes légales n'est pas le fruit de chercheurs isolés mais reflète des habitudes ancrées dans la culture des disciplines. Comment expliquer ces différences de pratiques entre sociologie et anthropologie, d'une part, et démographie et sciences politiques, d'autre part ? Les démographes et les politistes traitent les données de manière quantitative, tandis que les anthropologues et les sociologues adoptent une approche plus compréhensive, fondée sur l'observation de cas particuliers (le recueil de données se fait par le biais

d'entretiens, de notes d'observation…). Or, pour le traitement quantitatif des données, les chercheurs bénéficient de l'appui d'ingénieurs d'étude ou de recherche, qui les aident à constituer des bases de données puis à générer des statistiques ou des représentations graphiques. Ces ingénieurs ont également une connaissance précise des exigences légales en termes de données personnelles. Ils guident donc les chercheurs dans leurs démarches auprès de la CNIL. *« Les ingénieurs d'études, ici, sont très sensibilisés. Ce sont eux qui sont au contact de la CNIL, qui gèrent les bases de données, etc. Ils sont très au fait de ces questions et ils nous disent ce qu'il faut faire (déclarer à la CNIL, etc.). Pour les entretiens, en revanche, c'est vrai qu'on n'est pas… »* (chercheur 51). Le cas de ce chercheur est révélateur : il est amené à la fois à constituer des bases de données quantitatives et à réaliser des entretiens qualitatifs. S'il bénéficie du soutien des ingénieurs pour la partie « bases de données » de ses recherches, en revanche, il est davantage livré à lui-même pour la partie « entretiens ». Or on constate que les sociologues et les anthropologues sont en effet plus solitaires dans la gestion et le traitement de leurs données. Ils ne bénéficient pas d'intermédiaires pour les aiguiller dans leurs démarches auprès de la CNIL et ne prennent pas forcément le temps de se renseigner sur les procédures de signalement. La présence de relais semble donc avoir une influence positive sur l'utilisation des normes de gestion des données personnelles. Cette présence est d'autant plus décisive que l'application de la loi est peu contraignante : il n'y a pas de relevé systématique des projets de recherche collectant des données personnelles ; c'est aux chercheurs de signaler spontanément à la CNIL qu'ils ont l'intention d'utiliser des données personnelles. Cette explication reste néanmoins une hypothèse. Elle ne justifie peut-être que partiellement les différences de pratiques entre démographie/sciences politiques et anthropologie/sociologie.

2.5.2.2. Gestion des données issues de la recherche appliquée

Les modes de gestion varient également entre recherche fondamentale et recherche appliquée. Ce sont les chercheurs eux-mêmes qui, au cours de l'entretien, ont défini leurs travaux comme relevant de la recherche appliquée (chercheurs 2, 3 et 35). Le chercheur 2 développait une molécule destinée à un usage pharmaceutique (conception d'un médicament). Les deux autres

chercheurs travaillaient à mettre au point une technologie, pour l'un (chercheur 35), d'imagerie médicale, pour l'autre (chercheur 3), d'imagerie moléculaire.

L'OCDE définit la recherche appliquée comme *« [consistant] en des travaux originaux entrepris en vue d'acquérir des connaissances nouvelles. Cependant, elle est surtout dirigée vers un but ou un objectif pratique déterminé. […] Les connaissances ou les informations tirées de la recherche appliquée sont généralement susceptibles d'être brevetées mais peuvent également être conservées secrètes »*[273]. La recherche fondamentale est, quant à elle, décrite comme *« [consistant] en des travaux expérimentaux ou théoriques entrepris principalement en vue d'acquérir de nouvelles connaissances sur les fondements des phénomènes et des faits observables, sans qu'il y ait une application ou une utilisation particulière en vue »*.

Dans la recherche appliquée, l'importance de publier dans des revues à haut facteur d'impact existe bel et bien[274]. Mais elle se double d'un souci d'assurer le transfert de la technologie qui a été créée (Rebouillat, à paraître). Le gain de « crédibilité » passe donc par la publication d'articles (figure 12, cycle 1) mais pas seulement : la crédibilité s'acquiert également par le transfert de technologie, qui est l'enjeu même des recherches appliquées (figure 12, cycle 2). Pour ce type de projets, le cycle de crédibilité serait donc double. La réussite du transfert de technologie apporte en effet des bénéfices financiers aux tutelles, puisque dans le cadre d'une invention les tutelles sont propriétaires du brevet déposé. Grâce à ces bénéfices, les chercheurs s'attirent la confiance des tutelles et indirectement des agences de financement qui, par la suite, leur attribueront d'autant plus volontiers de nouvelles subventions. Dans le cycle 2, la crédibilité provient donc des bénéfices financiers issus du transfert de technologie.

273 ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (1975). *La mesure des activités scientifiques en techniques : méthode-type proposée pour les enquêtes sur la recherche et le développement expérimental. « Manuel de Frascati »*. https://hal.archives-ouvertes.fr/hal-01511852

274 Cf. supra, 2.3.1, p.170

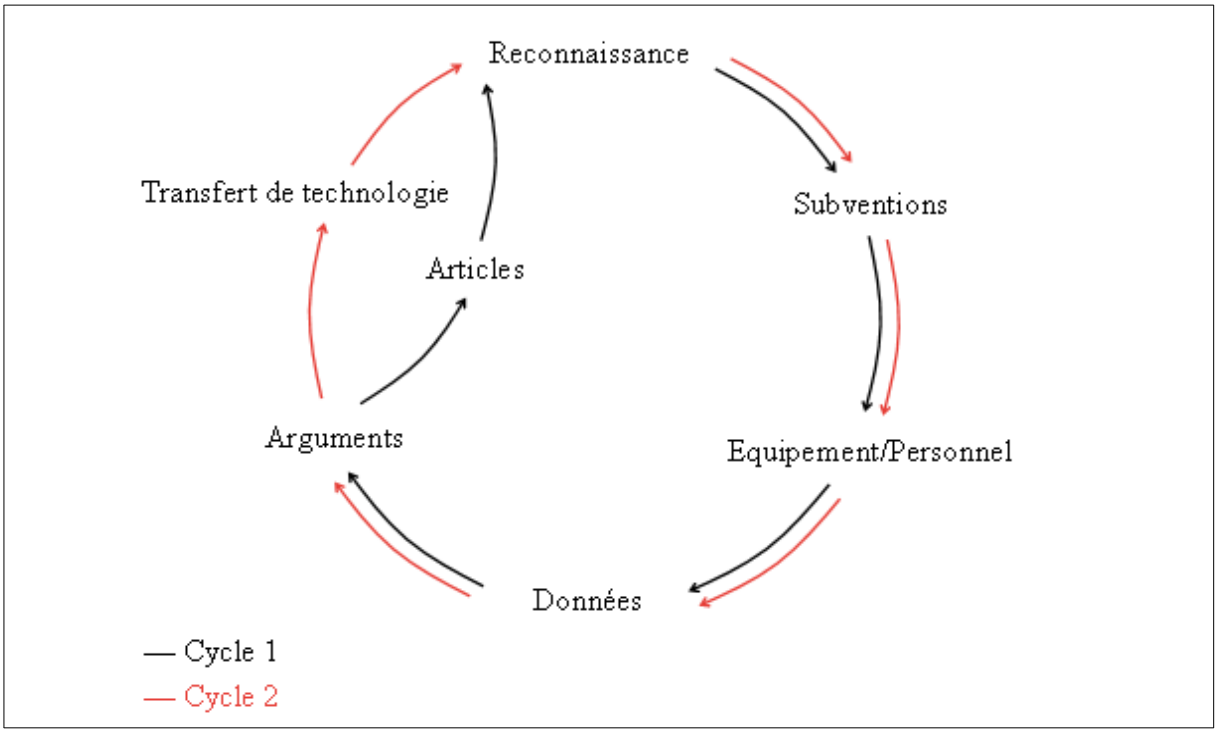


Figure 12 : Cycles de crédibilité dans la recherche appliquée[275]

Dans un projet de recherche appliquée, obtenir de la crédibilité et, par la suite, de nouveaux financements suppose de mener à bien le transfert de technologie. Or on constate que celui-ci dépend de la manière dont sont gérées les données de recherche. Le chercheur 2 explique qu'*« à partir du moment où on veut valoriser des résultats, où on va trouver un partenaire industriel qui s'intéresse à ces résultats, il va falloir que l'industriel soit convaincu des données que nous avons générées – ce qui suppose de mettre en place un certain nombre de pratiques »*. Pour garantir la reprise d'une invention par l'industrie, il est nécessaire que les données aient été (1) validées, (2) conservées et (3) protégées tout au long du projet de recherche.

(1) Valider les données suppose de s'être assuré d'un résultat avant d'entamer l'étape suivante de la recherche. Par ailleurs, la tenue rigoureuse des cahiers de laboratoire permet de garantir la traçabilité des données (c'est quelque chose à laquelle les chercheurs 3 et 35 veillaient

275 Cette figure a été conçue à partir du cycle de crédibilité proposé par Bruno Latour (voir supra, figure 11, p.171).

particulièrement). *« On documente le plus possible. […] Quand on fait une acquisition, tout est marqué dans le fichier d'acquisition, en termes de conditions de mesure »* (chercheur 35).

(2) Conserver les données consiste, dans le cas des chercheurs rencontrés, à stocker l'ensemble des données générées et à les répliquer sur des serveurs sécurisés et géographiquement distants.

(3) Enfin, protéger les données vise à préserver leur caractère inédit, c'est-à-dire à garantir l'exclusivité de la découverte et de son exploitation. C'est pourquoi les données, lorsqu'elles ont besoin d'être échangées entre les partenaires du projet, sont transmises via un réseau de messagerie non commercial voire sous forme de fichiers cryptés. *« Si on ne protège pas les données, on ne sera pas en mesure de les transférer à un industriel. Parce que l'industriel, ça ne l'intéresse pas si c'est public. Je pense à des entreprises qui veulent exploiter, faire de la licence sur ce qu'on fait ; eux, ils veulent des droits de licence pour être sûrs d'être en mesure d'exclure toute personne qui voudrait faire la même chose »* (chercheur 35). Ce constat du chercheur 35 rappelle que le transfert de technologie s'inscrit dans une logique de propriété intellectuelle, qui sous-tend exclusivité de diffusion et intérêts marchands.

Ces pratiques, qui divergent de la recherche fondamentale telle qu'elle a été décrite par les autres chercheurs interrogés, témoignent d'une préoccupation plus grande pour les données scientifiques. Car, dans un contexte d'innovation, ces dernières possèdent une valeur économique indirecte. De leur qualité et de leur confidentialité dépend la confiance accordée par le partenaire industriel. Le chercheur 2 comparait les méthodes de gestion de données utilisées dans les projets de recherche appliquée à celles d'une entreprise : *« Ça se rationalise. C'est de la gestion de projet, qui se professionnalise, qui n'est plus académique. On utilise les mêmes méthodes de management de projet que dans une société »*. Adopter les normes utilisées et approuvées par l'industrie contribue donc au transfert de technologie (de chaque côté les acteurs parlent un langage commun).

2.6. Ouverture des données

2.6.1. Le discours plutôt favorable des chercheurs

Les questions de science ouverte n'étaient pas totalement étrangères aux chercheurs rencontrés. Nous n'avons pas eu de mal à nous faire comprendre en utilisant l'expression « ouverture des données ». La question du libre accès est en effet un sujet relativement familier des chercheurs de l'Université de Strasbourg. Le discours politique incitatif de la direction de l'Université y a probablement contribué, notamment depuis la création de l'archive ouverte UnivOAK[276] en 2016. Mais la politique de l'établissement n'est pas la seule source de familiarité des chercheurs avec le thème de l'ouverture. En STM, suite à des cas de fraude, les équipes de recherche ont été sensibilisées par leurs tutelles aux questions d'intégrité scientifique. C'est dans ce cadre (dans un but de contrôle) que plusieurs revues scientifiques exigent désormais le dépôt des données à côté de la publication. En SHS, c'est plutôt par le biais des humanités numériques qu'est véhiculée l'idée d'une science ouverte, avec par exemple la mise en ligne de base de données thématiques (Dacos et Mounier 2015).

Tous les chercheurs interrogés se sont montrés favorables au principe de mise à disposition des données.

Ils justifient leur accord avec l'ouverture des données en utilisant les arguments suivants :

- Leurs recherches sont financées par des fonds publics. Eux-mêmes sont des fonctionnaires de l'État. Il est donc logique que le fruit de leur travail soit rendu public. *« C'est tout à fait justifié et naturel, puisque dans notre cas c'est du financement public. Donc, a priori, il n'y a pas de raison de faire de la rétention d'informations »* (chercheur 29). *« On fait des recherches qui sont financées par le secteur public, donc il est tout à fait logique que les données ou les résultats soient accessibles »* (chercheur 43).
- Rendre les données disponibles, quand il s'agit de données sous-jacentes à une publication, permet au lecteur de reproduire les résultats et de s'assurer ainsi de leur

276 https://univoak.eu/

intégrité. Cet argument intervient dans un contexte de lutte contre la fraude scientifique. Nous l'avons surtout entendu dans la bouche des chercheurs en STM. Les chercheurs ont néanmoins une vision critique de cette politique. Selon eux, exiger le dépôt des données liées à la publication réduira probablement les tentations de fraude, mais ne résoudra pas le problème. Celui qui veut faire mentir ses données pourra toujours le faire. Il lui suffira de truquer les données dans le fichier lié à l'article.

« Ce n'est pas parce que vous donnez la donnée brute que vous êtes obligé de donner la donnée « vraie ». Si vous êtes malhonnête, si vous avez obtenu 0,33 et que ça vous arrange d'avoir 0,22, si vous présentez 0,22 comme une donnée brute, personne ne peut vraiment le vérifier. Au final, je ne suis pas complètement certain que ça permette d'éviter la fraude. Malgré tout, ça peut y contribuer. […] Le fait qu'on nous dise « montrez-le nous » ou « mettez-le à disposition », on peut imaginer que ça évitera la fraude. C'est peut-être un obstacle psychologique supplémentaire pour la personne qui voudrait s'engager dans le chemin de la fraude. La personne, qui a juste à écrire en une phrase « voilà les données, vérifiez bien mon modèle », a peut-être psychologiquement un exercice supplémentaire à faire, en disant « d'une part mon modèle colle bien et en plus de ça je vous donne un jeu de données complètement bidonné ». » (chercheur 40).

- Un seul chercheur (chercheur 16) avait une vision plus stratégique de l'ouverture des données. Il était habitué, dans sa pratique, à rendre spontanément ses données accessibles en ligne. Pour lui, signaler les données augmente la probabilité de s'attirer des collaborations. Car les données donnent un aperçu de l'expertise du chercheur. En connaissant mieux ce qu'un chercheur produit, il est plus facile pour un tiers d'identifier des opportunités de collaboration. *« Ça n'est pas juste partager les données pour partager les données. Il y a toute une mentalité derrière, qui est concomitante. La modélisation, que j'ai faite durant ma thèse sur les primates, a ensuite été réutilisée par d'autres personnes, qui ont utilisé mes formules chez les humains et qui m'ont posé des questions auxquelles j'ai répondu. Du coup, c'est intéressant, parce que vous n'êtes pas forcément co-auteur de ces publications mais vous êtes cités. Et ensuite les personnes vous connaissent. Ça veut dire que, quand*

elles ont besoin d'organiser un symposium, elles vont faire appel à vous. Donc vous n'êtes pas récompensés tout de suite, mais un peu plus tard » (chercheur 16). C'est dans ce sens que la démarche d'ouverture est stratégique : elle est un catalyseur pour la carrière personnelle du chercheur, car elle donne de la visibilité à son travail et ouvre par ce biais des opportunités de recherches et de partenariats. Le « retour sur investissement » du partage des données n'est pas immédiat. Il intervient sur le plus ou moins long terme. Il y a une sorte de pari dans cette démarche : le chercheur gage que ses données lui attireront des propositions de collaboration futures.

Aucun des chercheurs ne s'est donc montré réfractaire à l'idée de partager ses données. Rendre les données librement accessibles selon les principes de la science ouverte a toutefois soulevé chez eux des interrogations.

Des interrogations concernant le délai de mise à disposition des données

Une majorité de chercheurs (29 sur 57) a spontanément fait remarquer qu'ils étaient d'accord pour ouvrir leurs données, mais seulement une fois l'article ou l'ouvrage relatif publié. C'est l'exploitation des données par celui qui les a collectées qui prime. On en revient ici à la question de la valeur d'usage des données pour le chercheur qui les a générées.

Des interrogations concernant la validité des données

Un chercheur faisait remarquer que les données scientifiques étaient souvent des données *« interprétées »*, c'est-à-dire qu'elles étaient le fruit d'une interprétation subjective, donc qu'elles étaient contestables. Or selon lui, *« c'est très délicat de mettre à disposition des données interprétées, parce qu'à valider ça n'est pas simple. […] Alors que, quand on publie une donnée interprétée, on passe par un comité de review. Les reviewers disent « oui, je suis d'accord » ou « non, je ne suis pas d'accord ». Notre donnée a été publiée, donc elle peut être réutilisée »* (chercheur 37). La donnée publiée aurait donc davantage de crédibilité que la donnée non publiée, car elle est validée par les relecteurs. Une personne tierce peut la réutiliser avec d'autant plus de confiance. Il y a donc chez les chercheurs le souhait d'une

validation des données par les pairs avant leur réutilisation par la communauté. A ce jour, c'est la publication via les comités de relecture qui fait office de structure de validation.

Une remarque similaire nous a été faite par un chercheur en astronomie. Dans ce domaine, les données d'observation du ciel sont de plus en plus réutilisées par des personnes autres que celles qui les ont générées. Or, du fait des conditions d'acquisition, les données possèdent toujours des biais implicites, qui ne sont pas forcément connus du chercheur qui les réutilise.

> *« En pratique, on a la bonne ou la mauvaise habitude d'aller toujours au plus profond des données, au plus près du bruit des données. Et c'est là que ça devient problématique. Quand on a observé nos données nous-mêmes, qu'on a fait l'analyse nous-mêmes, on sait quoi croire dans les données ou pas. Si on voit un signal très, très faible, on peut se dire : « C'est peut-être parce que notre travail de calibration n'est pas parfait ». On a introduit un biais. Qui est très faible. [...] Alors que si on n'a jamais travaillé sur des données, il y a toujours cette impression que quand on nous donne à disposition une grande base de données, tout a été fait correctement, tout a été bien calibré. Parce qu'il y a une paresse intellectuelle qui fait qu'on se dit : « Ok, quelqu'un a déjà fait tout le travail de test et de vérification ». Sauf qu'on attaque tous des données avec des points de vue différents. Donc il y aura effectivement tout un tas de tests de calibration qui auront été faits pour vérifier que tout fonctionne à peu près correctement. Mais, nous, la question à laquelle on va répondre, c'est peut-être quelque chose à laquelle ils n'avaient absolument pas pensé. […] C'est plus une philosophie générale. C'est-à-dire que si on est allé au télescope (je suis un grand défenseur d'envoyer les étudiants au télescope), si on a obtenu nos propres données, je pense qu'on voit ce qui peut aller de travers. […] On se rend compte de tout ce qui peut impacter les images et les spectres qu'on va observer. Tout d'un coup, au milieu de la nuit, les nuages vont arriver. On termine quand même la pose qu'on avait commencée, mais elle sera peut-être de moins bonne qualité. Du coup, on a des données hétérogènes dans notre jeu de données. »* (chercheur 28)

La question de la validité des données, brutes ou interprétées, constitue donc un aspect important à prendre en compte, aux yeux des chercheurs, quand on envisage de les rendre librement accessibles.

Des interrogations concernant la réutilisation des données

La réutilisation des données est la troisième question à avoir été soulevée par les enquêtés, quoique de manière moins récurrente que les précédentes. Plusieurs chercheurs s'interrogent sur l'intérêt que peuvent avoir leurs données pour d'autres personnes. Il s'agit essentiellement de chercheurs travaillant à partir d'enquêtes qualitatives (en sociologie notamment). Dans ce type de recherches, le questionnaire est contextuel et spécifique à la question de recherche que se pose le chercheur. Un sociologue nous disait : *« les données qui sont récoltées dans le cadre d'un entretien sont liées à la problématisation qui a été faite au préalable, aux hypothèses qui ont été construites. Donc je suis toujours un peu perplexe quant à la capacité de réutiliser des données produites par et pour un projet de recherche donné dans le cadre d'un autre projet »* (chercheur 52). Ce chercheur ne voit pas quelle réutilisation pourrait être faite de ses entretiens, si ce n'est exactement la même recherche. Or en sciences humaines et sociales c'est quelque chose qui n'a pas réellement de sens. Chaque chercheur s'efforce plutôt d'étudier un terrain qui lui est propre.

La réutilisation des données soulève également une seconde interrogation : celle de la confidentialité de certaines d'entre elles. Là encore en font partie les données d'entretiens. Il peut aussi s'agir de coupes histologiques issues de la biopsie d'un patient malade (chercheur 33), dans le domaine de la santé. Dans les deux cas, les données contiennent des informations personnelles, dont le chercheur est tenu de préserver la confidentialité. Selon les recommandations de la CNIL, ces données doivent notamment être anonymisées après leur collecte. Selon les chercheurs interrogés, réutiliser de telles données (anonymisées) est relativement difficile pour quelqu'un qui ne les a pas collectées. Car leur anonymisation occulte un certain nombre d'informations contextuelles que les chercheurs jugent essentielles à toute personne qui souhaiterait réanalyser ces données. *« A quoi ça sert de mettre en libre accès un entretien, si on ne connaît pas la personne qui est interrogée ? »* (chercheur 52).

Rendre les données directement accessibles en ligne reste cependant une exception. Parmi les chercheurs rencontrés, certains avaient déjà déposé des données en ligne – notamment des données de séquençage ou des données d'enquêtes quantitatives – mais ces dépôts restent à l'état de pratiques ponctuelles, motivées par l'injonction des éditeurs au moment de la publication d'un article.

2.6.2. Nature du partage dans la culture scientifique : un échange censé contribuer à la renommée du chercheur

Les politiques d'ouverture des données s'introduisent dans un univers – la communauté scientifique – qui a ses propres règles[277]. Le partage des données y est régi par des principes différents de ceux de l'ouverture systématique et sans barrières. Comme le formulait Joachim Schöpfel dans les résultats d'une enquête menée auprès des laboratoires de sciences humaines et sociales de l'Université de Lille, il n'y a pas une science « *open* » et une science « *closed* » (Schöpfel 2018b, p.22). Les chercheurs partagent leurs données. Tout d'abord, dans les publications. Les données sont en effet rendues publiques dans les communications scientifiques, de manière plus ou moins sélective, en fonction de la longueur du document (les données pourront ainsi être plus complètes et mieux décrites dans un ouvrage que dans un article). Ce mode de communication des données, on l'a vu, est source de crédibilité pour le chercheur. Ce dernier obtient en échange de la publication la reconnaissance de ses pairs et, par là, des perspectives de financement et d'avancement.

Il existe également d'autres formes de partage des données dans la communauté scientifique. En particulier celle qui intervient dans le cadre de ce que les chercheurs appellent une « collaboration ». Ce terme est en effet souvent revenu au cours des entretiens. Les chercheurs l'emploient pour désigner un accord :

- Entre plusieurs chercheurs ;

277 Voir supra, 2.3, p.170

- Ou bien entre un/des chercheur(s) et un/des organisme(s) public(s) ou privé(s) (l'Office National des Forêts[278], l'Eurométropole de Strasbourg[279] ou l'Association pour la Recherche sur la Sclérose Latérale Amyotrophique[280] par exemple).

C'est un accord qui peut être tacite ou bien faire l'objet d'un contrat écrit (dans le cadre d'un projet de recherche multi-partenaire par exemple). La forme tacite est essentiellement utilisée pour les collaborations entre chercheurs. Elle est fondée sur une relation de confiance préexistante. Dans les deux cas, la connaissance des partenaires est une dimension importante (on sait avec qui on travaille).

Il existe différents types de collaborations :

- Les collaborations fondées sur l'association de compétences complémentaires. Par exemple, une équipe de chimistes et une équipe de biologistes vont travailler ensemble sur un même sujet de recherche (chercheur 11). Il existe deux modes de collaborations pluridisciplinaires.
    - Dans le premier mode de collaboration, le projet est segmenté en différents axes et chaque partenaire travaille sur des données différentes. L'échange de données n'est alors pas nécessaire.

      Comme un chercheur l'expliquait pour sa propre situation (chercheur 18), *« il n'y a pas d'échanges de données brutes dans le cadre de ce projet, dans la mesure où chacun avait sa propre ligne de travail et où ces lignes de travail étaient indépendantes pour la presque totalité de ce que nous faisions. […] Il y a des échanges de bons procédés, des échanges de données analysées, des discussions aussi sur l'interprétation qu'il convient d'apporter à ce qu'on observe ; mais ce n'est pas dans notre habitude d'échanger des données brutes. »*

      Les données analysées (données finales) sont mises en commun au moment de la publication. C'est un cas de figure qui semble être assez courant dans les projets de recherche pluridisciplinaires. Un chercheur expliquait par ailleurs

278 https://www.onf.fr/

279 https://www.strasbourg.eu/

280 L'ARSLA (https://www.arsla.org/) est une association de patients, dont le but est de soutenir les personnes atteintes de sclérose latérale amyotrophique.

que *« ça n'[avait] pas toujours de sens de faire tourner la donnée brute. Parce que moi, si un chimiste m'envoie un spectre [inaudible], je ne comprendrai pas ce qu'il y a dessus »* (chercheur 3). Chaque partenaire possède son propre domaine de compétences et n'a pas l'expertise pour comprendre les données (a minima brutes) de l'autre partenaire. L'opacité des données pour chaque partenaire respectif rend donc l'échange de données brutes peu utile.

- Le second mode de collaboration a pour finalité l'analyse des données. L'échange de données y est donc nécessaire. Ce type de collaboration consiste pour un chercheur à s'associer les compétences d'un autre chercheur, chargé d'analyser les données. C'est le cas de disciplines où, en raison de l'augmentation des volumes de données, il est de plus en plus nécessaire de s'adjoindre les compétences de chercheurs ou d'ingénieurs en informatique.

  En biologie par exemple, les volumes très importants de données issues du séquençage d'ADN demandent des compétences nouvelles en bioinformatique. Un des biologistes rencontrés (chercheur 1) avait notamment mis en place une collaboration avec un bioinformaticien, n'ayant pas lui-même les compétences pour analyser ses données. *« Dans ces cellules de Sertoli, il a appliqué et il sait très bien appliquer de nombreuses approches de bioinformatique. C'est pile poil ce qu'il me faut pour comprendre mes approches holistiques. Je vais pouvoir lui donner mes résultats bruts de sites de fixation des RAR et mes mesures quantitatives d'expression des gènes. Et lui va pouvoir me mouliner tout ça ».*

- Il existe également des collaborations fondées sur l'échange réciproque de données. On peut citer l'exemple d'un projet de recherche en sciences de la Terre, associant des équipes de recherche et des organismes publics, dont l'Office National des Forêts (ONF). La contribution de l'ONF consistait en la fourniture de *« données chiffrées sur les quantités de bois exportées par parcelle »* (chercheur 37). En échange, l'équipe de recherche en géochimie leur remettait les résultats de leurs analyses : *« Ils ont récupéré nos données de sol, comme ça, ils n'ont pas eu besoin de les mesurer. Ça*

leur a permis d'économiser ça. Donc ce sont de petites choses comme ça : ils nous donnent des données ; nous, on leur donne des données ».

- Enfin, ont pu être identifiées des collaborations centrées sur la production de données. Ces collaborations concernent essentiellement les domaines de recherches nécessitant des équipements d'envergure. Dans le cadre de l'enquête, nous avons notamment pu étudier le cas de l'astronomie. Dans ce domaine, l'observation du ciel est aujourd'hui réalisée à partir d'instruments (imageurs, spectroscopes…) extrêmement coûteux à mettre en place. Les pays sont conduits à s'associer pour financer leur construction. Aussi chaque instrument fédère-t-il des collaborations internationales, permettant aux chercheurs qui y participent d'avoir directement accès aux données générées. Ici l'échange se fait à un niveau supra-institutionnel : les États paient la construction de l'instrument ; en échange, leurs chercheurs peuvent bénéficier des données acquises.

La collaboration est un cadre qui permet au chercheur de partager ses données, tout en obtenant quelque chose en échange. La collaboration crée un lien de réciprocité, où chaque partenaire donne et reçoit. Elle garantit une situation d'équilibre entre les collaborateurs.

De même qu'un article n'est pas « offert » à la communauté (puisque le chercheur en retire du crédit), les données sont rarement communiquées sans attente d'un retour. Lorsqu'un chercheur utilise les données d'un autre chercheur, généralement il ajoute le nom de ce dernier à la liste des co-auteurs de l'article élaboré à partir des données.

La culture scientifique semble donc fondée sur un principe d'échanges réciproques. Le partage de données ne s'inscrit pas dans une logique de don désintéressé. Le chercheur qui communique ses données en attend un bénéfice personnel. Tout honorable qu'il paraisse aux yeux des chercheurs, le principe d'ouverture ne sera donc considéré avec sérieux par les communautés scientifiques que s'il offre une rétribution, directe ou indirecte, en échange.

3. Conclusion

L'enquête révèle une très grande variété de pratiques de gestion et de partage de données, alors même qu'elle n'inclue pas tous les domaines scientifiques (les mathématiques, la linguistique et la physique par exemple ne font pas partie de l'échantillon). De manière générale, elle montre des pratiques sur mesure, qui sont fonction des cultures épistémiques (Knorr-Cetina 1981), du type de données, du cadre de leur collecte et des normes qui s'y appliquent, ainsi que des ressources humaines et techniques à disposition. En cela, elle confirme l'influence du cadre épistémique (première hypothèse de recherche) et du cadre institutionnel (deuxième hypothèse de recherche) sur les modes de gestion et de partage des données.

Dans quelques rares communautés étudiées (une partie de l'astronomie, de la démographie et de la santé publique), la gestion et le partage des données sont institutionnalisés, c'est-à-dire qu'ils sont réalisés par des personnels spécifiques dans des structures dédiées. Dans la communauté des sciences omiques, le partage des données tend également à se systématiser et se standardiser, mais ce sont les équipes de recherche qui doivent par elles-mêmes déposer les données dans des entrepôts. A la différence de l'astronomie, la gestion et le partage ne sont pas pris en charge par des infrastructures. C'est par le biais des revues scientifiques que s'instaurent progressivement des pratiques d'ouverture, celles-ci conditionnant la publication d'un article au dépôt préalable des données.

Un des points mis en évidence est que, dans la logique scientifique, le partage des données répond à un besoin. Il s'intègre dans la stratégie du chercheur, de son équipe de recherche ou de sa communauté. Les données s'inscrivent dans un cercle de crédibilité (Latour 2001). Elles ont d'abord une valeur d'usage, puisqu'elles vont servir à publier de nouveaux articles et ouvrages. Elles ont aussi une valeur d'échange dans le cadre de collaborations : elles sont transmises au collaborateur en échange d'une expertise ou d'une publication dans laquelle le chercheur sera co-auteur. La stratégie du chercheur consiste à estimer, dans une situation donnée, quelle valeur est la plus grande : la valeur d'usage ou la valeur d'échange. L'objectif du chercheur est avant tout de préserver ses intérêts professionnels. Ces résultats confirment donc partiellement la troisième hypothèse de recherche, selon laquelle le partage des données a lieu dans le cadre d'échanges privés et réciproques. Si la dimension de réciprocité a bien été

établie, en revanche on ne peut pas dire que les échanges de données soient exclusivement privés. Les chercheurs communiquent leurs données aussi bien à des personnes physiques (leurs collègues) qu'à des personnes morales (associations, revues scientifiques…), que ce soit pour une utilisation publique ou privée. Ce qui varie, c'est le degré de contractualisation de l'échange : moins le chercheur connaît personnellement son ou ses interlocuteur(s), plus l'échange – la communication des données et sa contre-partie – sera formalisé (par une convention de recherche ou un contrat d'édition par exemple).

Cinquième partie

-

Adéquation entre les services de données et les pratiques des chercheurs

Les autorités publiques prônent la gestion et l'ouverture des données de la recherche. En appui à ce discours, des services se créent afin d'aider les chercheurs à structurer et diffuser leurs données. Dans certaines communautés scientifiques, des pratiques d'ouverture existent déjà, répondant à un besoin en termes d'exploitation des données. Pour d'autres communautés, l'ouverture n'a pas d'utilité à court terme ; aussi n'est-elle pas ancrée dans les habitudes. Par conséquent, quels services de données utilisent les chercheurs ? Ont-ils recours aux services créés dans un contexte d'ouverture ? Telle sera la réflexion menée dans ce dernier chapitre, avec pour perspective de vérifier la quatrième hypothèse de recherche.

1. Enquête sur l'utilisation de services de données par les chercheurs

1.1. Méthodologie

Pour tenter de répondre à ces questions, un des volets de l'enquête sur les pratiques des chercheurs[281] a été consacré à l'utilisation de services de données.

L'approche choisie est une approche qualitative, visant à comprendre pourquoi et comment les chercheurs ont recours à des services de données. L'objectif n'est pas de quantifier l'usage qu'ils en font au moyen d'un échantillon représentatif et de statistiques d'utilisation.

Le catalogue Cat OPIDoR, commandité par la Bibliothèque Scientifique Numérique, a servi de base d'étude pour cette enquête. Il a été présenté aux chercheurs au cours des entretiens, qui l'ont parcouru, tout en émettant des observations.

L'objectif était de demander aux chercheurs :

- s'ils utilisaient des services de données ;
- s'ils connaissaient Cat OPIDoR et les services qui y étaient répertoriés ;
- s'ils utilisaient d'autres services[282].

Ces questions se sont ajoutées à celles posées au cours de l'enquête sur les pratiques de recherche. Elles ont été posées aux 41 chercheurs du second panel de l'enquête (de visu ou par téléphone, entre mars et mai 2019), ainsi qu'à 5 des 16 chercheurs du premier panel, qui ont accepté un second échange (par mail ou par téléphone en février 2019). Au total, 46 chercheurs ont donc été interrogés sur l'utilisation de services de données (tableau 13).

281 Cf. supra, quatrième partie, 1, p.156

282 Voir le guide d'entretien du second panel de chercheurs interrogés (annexe 11)

Grand domaine scientifique	Laboratoires	Discipline du laboratoire	Chercheurs	Discipline	Mode de réponse
Vie & Santé	Laboratoire 1	Biologie	Chercheur 1	Génétique	Mail
	Laboratoire 3	Chimie et biologie	Chercheur 4	Biologie	Entretien de visu
			Chercheur 5	Biologie	Entretien de visu
			Chercheur 6	Chimie et biologie	Entretien téléphonique
	Laboratoire 4	Chimie et biologie	Chercheur 7	Biologie	Entretien de visu
			Chercheur 8	Biologie	Entretien téléphonique
			Chercheur 9	Biologie	Entretien de visu
			Chercheur 10	Biologie	Entretien de visu
			Chercheur 12	Médecine	Entretien de visu
	Laboratoire 5	Écologie	Chercheur 13	Écologie	Entretien de visu
			Chercheur 14	Écologie	Entretien téléphonique
			Chercheur 15	Écologie	Entretien de visu
	Laboratoire 6	Neurosciences	Chercheur 17	Bioinformatique	Entretien de visu
			Chercheur 19	Neurosciences	Entretien de visu
			Chercheur 20	Neurosciences	Entretien de visu
			Chercheur 21	Primatologie	Entretien téléphonique
	Laboratoire 7	Neurosciences	Chercheur 22	Neurosciences	Mail
Sciences & Technologies	Laboratoire 8	Astronomie	Chercheur 23	Astronomie	Entretien de visu
			Chercheur 24	Astronomie	Entretien de visu
			Chercheur 25	Astronomie	Entretien de visu
			Chercheur 26	Astronomie	Entretien de visu
			Chercheur 27	Astronomie	Entretien de visu
			Chercheur 28	Astronomie	Entretien de visu
			Chercheur 29	Astrophysique	Entretien téléphonique
	Laboratoire 9	Sciences de l'ingénieur	Chercheur 30	Bioinformatique	Entretien de visu
			Chercheur 31	Géographie	Entretien de visu
			Chercheur 32	Informatique	Entretien téléphonique
			Chercheur 33	Informatique	Entretien de visu
			Chercheur 34	Informatique	Entretien de visu
	Laboratoire 10	Sciences de la Terre	Chercheur 38	Géologie	Entretien téléphonique
			Chercheur 39	Géologie	Entretien de visu
			Chercheur 40	Géologie	Entretien téléphonique
Sciences humaines et sociales	Laboratoire 11	Géographie	Chercheur 41	Ecologie	Entretien de visu
			Chercheur 42	Géo-archéologie	Entretien de visu
			Chercheur 43	Géographie	Mail
			Chercheur 45	Géographie	Entretien téléphonique
	Laboratoire 12	Sciences sociales	Chercheur 46	Anthropologie	Entretien de visu
			Chercheur 47	Démographie	Entretien de visu
			Chercheur 48	Démographie	Entretien de visu
			Chercheur 49	Droit	Entretien de visu
			Chercheur 50	Histoire	Entretien de visu
			Chercheur 51	Sciences politiques	Mail
			Chercheur 53	Sociologie	Entretien de visu
			Chercheur 54	Sociologie	Entretien de visu
	Laboratoire 13	Théologie	Chercheur 56	Théologie	Entretien de visu
			Chercheur 57	Théologie	Entretien de visu

Tableau 13 : Liste des chercheurs interrogés sur l'utilisation de services de données

1.2. Avis des chercheurs sur Cat OPIDoR

Méconnaissance de l'existence du répertoire

Aucun des chercheurs interrogés ne connaissait l'existence de Cat OPIDoR[283]. Le catalogue a néanmoins été perçu comme intéressant et potentiellement utile.

Remarques sur son organisation et son ergonomie

Plusieurs remarques précises ont également été faites sur l'ergonomie et l'organisation de Cat OPIDoR, soulignant son manque d'intuitivité[284].

Plusieurs chercheurs auraient apprécié :

- que figure sur la page d'accueil une phrase introductive expliquant le rôle et le contenu de Cat OPIDoR ;
- que la classification disciplinaire soit davantage mise en valeur sur la page d'accueil (elle n'est pas visible au premier coup d'œil ; on aperçoit d'abord la typologie des services, mais celle-ci parle peu aux chercheurs) ;
- qu'elle soit plus développée (certains chercheurs n'ont pas remarqué qu'on pouvait déplier chaque domaine en sous-domaines – un clic supplémentaire est en effet nécessaire pour faire apparaître les sous-domaines) ;
- que la dénomination des disciplines soit plus claire. *« Moi je ne sais pas où je suis. Je peux être dans « Vie ». Il y a des trucs qui s'y apparentent : « Biologie de l'évolution des populations et environnementale ». Mais c'est vrai qu'il n'y a pas forcément « Écologie » ou « Environnement ». »* (chercheur 41).
- que les résultats soient présentés de manière plus uniforme. Ils prennent tantôt la forme d'un tableau, tantôt la forme d'une liste. Les résultats sous formes de listes ne délivrent que deux types d'informations (l'acronyme du service et la discipline qu'il cible) ; ils ne résument pas à quoi est destiné le service. Les résultats sous forme de

283 Aucun n'a non plus fait mention du répertoire Re3data (https://www.re3data.org/), bien que la question n'ait pas été directement posée.

284 Les annexes 4, 5 et 6 présentent des copies écran du catalogue, permettant de mieux se figurer les retours des chercheurs.

tableau fournissent davantage d'informations mais n'indiquent pas systématiquement la discipline du service.

Problèmes de terminologie

Plusieurs chercheurs ont pensé que Cat OPIDoR était un entrepôt ou un portail de données. Probablement parce que le terme de « répertoire de services » ne leur était pas parlant. Celui-ci a été choisi par les concepteurs de Cat OPIDoR, pour désigner sous un même terme un ensemble d'infrastructures disparates. Il n'était pas utilisé auparavant.

De même, au cours d'un entretien, un chercheur démographe a associé le terme d'« entrepôt de données » à une méthodologie utilisée dans son domaine. *« Je forme mes étudiants à la méthode de l'entrepôt de données. Souvent ils vont être confrontés à ça : pour une étude qu'ils auront à mener, les données viendront de différentes bases de données, pouvant se croiser et se compléter. C'est le principe de l'ELT* (Extraction, Load, Transformation). *Vous avez les données sur plusieurs bases de données. Vous les extrayez. Vous les chargez dans une base commune. Vous les transformez pour les uniformiser et pour pouvoir travailler dessus »* (chercheur 47). Cette confusion révèle que le vocabulaire utilisé par la sphère des professionnels de l'IST n'est pas forcément le même que celui utilisé dans tel ou tel discipline scientifique.

Le terme de « services » étant vaste, il a été compris de différentes manières par les chercheurs. Certains ne considéraient par exemple pas les banques de données comme des services.

Un périmètre trop restreint

Trois répondants (en neurosciences, chimie/biologie et écologie) ont pointé comme inconvénient le périmètre national du catalogue, précisant qu'un recensement international serait plus utile dans le cadre de leurs recherches.

- *« Ça va dépendre énormément de la discipline. Je me demande si un recensement national n'est pas plus adapté à des disciplines de sciences humaines, où il y a un contexte un peu plus culturel parfois. C'est vrai que nous, pour le coup, quand on*

travaille sur le vivant, les frontières n'ont plus trop d'importance. Elles sont d'ailleurs parfois un peu limitantes pour la recherche » (chercheur 41). Selon ce chercheur, les banques de données ont un périmètre plus biogéographique qu'étatique.

- *« Je ne connais aucun des outils listés et je n'arrive pas à savoir si c'est parce que c'est trop pointu (calculs et analyses que seul un petit nombre de bio-informaticiens pourraient réaliser), ou si c'est simplement parce que les sites équivalents, dont je pourrais avoir entendu parler, sont internationaux (quel est l'intérêt d'une plateforme française à l'heure actuelle ?) »* (chercheur 22).
- *« Par exemple, chez nous, la Protein Data Bank est connue. C'est un monument tellement elle est connue. Et elle n'est pas chez vous. C'est international mais bon, tous les laboratoires dans mon domaine l'utilisent. Il n'existe pas d'équivalent national. Et l'équivalent national n'a aucun sens, parce que par définition ça doit être international. […] Un site qui répertorie les bases de données au niveau national, c'est peu utile. Il faut raisonner au minimum à l'européen et encore c'est à l'international. Et au niveau international, si vous faites un répertoire de tout ce qui existe dans le monde, humainement ça n'est pas possible. Il faut que ça soit automatisé »* (chercheur 6).

<u>Méconnaissance des services répertoriés dans Cat OPIDoR</u>

Par ailleurs, rares étaient les chercheurs connaissant ou utilisant un des services répertoriés dans Cat OPIDoR. Ils en connaissaient ou utilisaient au mieux un ou deux :

- *« A part le synchrotron SOLEIL, le reste ne me dit rien »* (chercheur 6 en chimie et biologie) ;
- Les chercheurs démographes et politistes connaissaient et, pour certains, utilisaient les services de PROGEDO (chercheurs 47, 48 et 51) ;
- Les astronomes connaissaient SIMBAD et VizieR, du Centre de Données astronomiques de Strasbourg (chercheurs 23 à 29);
- En écologie, un chercheur connaissait Recolnat (chercheur 41) ;

- Le pôle HPC Unistra était connu et utilisé de chercheurs en informatique (chercheur 34), en théologie (chercheur 56), en géographie (chercheur 43) et en biologie et chimie (chercheur 6).

Selon les chercheurs, c'est en raison de l'hyperspécialisation des thématiques de recherche que de nombreux services répertoriés dans Cat OPIDoR ne leur sont pas connus. Ils ne connaissent que les services qui ont trait à leur domaine et qui leur sont utiles. Un démographe expliquait ainsi : *« Je ne connais pas personnellement cette cohorte [ELIPSS[285]], parce que je ne les connais pas toutes. Je me concentre plus sur les enquêtes de ma thématique. Il y a beaucoup d'enquêtes. Il y a eu un développement. Donc il y a un développement des possibilités de traitement de ce type de données »* (chercheur 47 en démographie).

Il existe par ailleurs beaucoup plus de services que n'en répertorie Cat OPIDoR. *« Chaque domaine est tellement spécifique, ça bouge tellement vite que, pour le moment, comme ça, à chaud, j'ai du mal à imaginer qu'un seul site, sans faire une collecte automatisée, puisse être vraiment utile ou vraiment à jour. […] Dans notre domaine, par exemple, ça va être tellement spécifique ce qu'on utilise. Les gens qui sont dans le laboratoire juste à côté vont utiliser d'autres trucs spécifiques, parce qu'ils travaillent sur d'autres sujets et utilisent d'autres techniques. Même si pour vous ça va être la même chose, pour nous ce sont des choses très différentes »*, expliquait un chercheur en chimie et biologie (chercheur 6).

1.3. Un recours peu fréquent aux services de données

De manière générale, les chercheurs rencontrés ont peu recours à des services de données. Ils collectent, analysent, gèrent et partagent leurs données avec les moyens à disposition dans leur laboratoire.

- Parfois leurs recherches ne requièrent ni outil ni expertise externes. C'est le cas de 3 des 46 chercheurs rencontrés, en droit (chercheur 49), en histoire (chercheur 50) et en théologie (chercheur 57). Ces chercheurs manipulent essentiellement des textes

285 Étude Longitudinale par Internet pour les Sciences Sociales (https://www.elipss.fr/)

(juridiques, historiques…). Leur démarche consiste à lire, analyser et interpréter ces textes selon une approche herméneutique, ne requérant aucun service de données.

- Un autre facteur explicatif est la méconnaissance des services existants. Plusieurs chercheurs ont déclaré se sentir dépassés par le flot d'informations. Deux d'entre eux (le chercheur 39 en sciences de la Terre et le chercheur 45 en géographie) ont explicitement rendu compte de difficultés à repérer quels services existaient. Il est probable que la diversité et le manque de visibilité des services rendent ce repérage malaisé. A cette difficulté s'ajoute un manque de temps global, non favorable à une recherche d'informations approfondie. Cat OPIDoR a été jugé utile sur ce point. Les chercheurs ont supposé qu'il pourrait leur permettre de rechercher un service dans le cadre d'un besoin précis. *« C'est vraiment bien, parce que les données sont très éparses. On peut passer beaucoup de temps à trouver des données. Donc ces grands sites-là facilitent notre travail »* (chercheur 45).
- Parfois encore, les chercheurs font sciemment le choix de ne pas utiliser de services, considérant que le travail d'analyse requiert une nécessaire proximité avec l'origine des données. Par « origine » on entend le terrain d'étude. Plusieurs chercheurs considèrent en effet qu'ils doivent collecter les données par eux-mêmes. En sciences sociales, les anthropologues et les sociologues disent collecter et traiter les données de manière solitaire. La nature qualitative de leur analyse nécessite, selon eux, de se confronter personnellement au terrain. Ils n'utilisent donc ni service ni logiciel. *« Je travaille avec moi-même. C'est ce que je vous disais dans le mail : je n'utilise pas vraiment de programmes de traitement, ni pour les entretiens, ni pour les images. Un entretien par exemple, je le relis, je le transcrits et je fais des commentaires à côté. C'est voulu. C'est un choix de ne pas utiliser de programme d'analyse d'entretiens. Parce que c'est une connaissance d'un matériel qui est très intime. Quand c'est nous-mêmes qui le transcrivons et qui revenons dessus (« là je n'ai pas compris, là j'ai posé la mauvaise question »), ça permet un travail réflexif qu'on n'a pas, si c'est quelqu'un d'autre qui le transcrit. Moi je considère que mon travail de recherche passe par le traitement de mes données. Je ne le conçois pas autrement. Le traitement c'est la transcription, la lecture, le coloriage... »* (chercheur 46).

1.4. Deux ordres de services

Deux ordres de services semblent se distinguer :

- Un premier ordre de services, né sous l'influence du mouvement d'ouverture des données ;
- Un second ordre de services, conçus pour répondre aux besoins des communautés de recherche.

Le premier ensemble de services est apparu récemment en France (après 2010). Il entend répondre aux politiques d'Open Science et propose essentiellement des services pour la gestion et l'ouverture des données.

Le second ensemble de services, que nous appellerons « services disciplinaires », rassemble des dispositifs nés au sein des communautés scientifiques pour répondre à un besoin particulier. Ce type de services existe depuis plus longtemps que les services pour la gestion et l'ouverture des données. Il s'agit de services qui ont une utilité pratique et immédiate pour le chercheur qui y a recours, lui permettant de mener à bien son projet de recherche.

Dans Cat OPIDoR, on trouve les deux ordres de services. Il a été plus facile pour l'équipe projet d'identifier les services d'ouverture des données car le recensement était commandité par des professionnels de l'information scientifique et technique. En revanche, il a été plus difficile de répertorier les services disciplinaires, car ceux-ci font partie de communautés scientifiques très spécifiques, dont il faut connaître le langage et les réseaux.

L'enquête semble montrer que les chercheurs se tournent avant tout vers les services du second ordre (Rebouillat et Chartron 2019). Les parties suivantes (2 et 3) tentent d'expliquer ces différences d'utilisation.

2. Utilisation par les chercheurs des services nés sous l'influence du mouvement d'ouverture des données

Le discours véhiculé par le mouvement d'ouverture de la recherche a, depuis les années 2000, donné naissance à divers services dédiés aux données scientifiques. Qu'il s'agisse d'entrepôts de données, d'équipes d'accompagnement ou de sites d'information, ces services ont pour point commun d'utiliser un vocabulaire caractéristique de ce mouvement, à savoir celui de l'ouverture et de la réutilisation. A la différence des services disciplinaires, qui se sont développés au sein des communautés scientifiques, ces services sont généralement fournis par des acteurs tiers. Ils revendiquent leur légitimité dans leur rôle de médiation entre politiques d'ouverture et communautés de recherche.

Dans Cat OPIDoR, ces services correspondent notamment :

- Aux 17 services développés par la sphère des professionnels de l'information scientifique et technique[286] ;
- Ainsi qu'aux entrepôts de données SEANOE, Ortolang, DATA BRGM, CoCoON, Kinsources.net et aux annuaires de données Portail Epidémiologie France, IO Data Science et ECOSCOPE.

2.1. Une méconnaissance de ces services

Il semble que ces services nés du mouvement d'ouverture des données soient peu utilisés des scientifiques. Aucun des chercheurs interrogés dans l'enquête ne connaissait ou n'avait déjà eu recours à un des 25 services mentionnés ci-dessus. La question de l'ouverture des données est certes un sujet qui parle aux chercheurs mais qui, à quelques exceptions près, ne fait pas partie de leurs préoccupations quotidiennes. Le fait que l'évaluation de la recherche et l'attribution de financements reposent essentiellement sur la publication d'articles et d'ouvrages ne leur permet pas de consacrer plus de temps et d'intérêt aux données[287].

286 Voir supra, troisième partie, 2.4, p.134

287 Voir supra, quatrième partie, 2.3, p.170

Cela ne signifie cependant pas que les chercheurs n'ont pas de besoins en termes de données. Les résultats de l'enquête révèlent en effet des besoins pour le stockage et l'organisation des données[288]. Ces besoins s'expriment toutefois en arrière-plan. Les préoccupations des chercheurs étant principalement tournées vers la publication, les problèmes liés aux données restent en effet au second plan et les besoins sont rarement clairement exprimés. Les chercheurs tentent de trouver des solutions par eux-mêmes, avec les moyens dont ils disposent dans leur laboratoire. Derrière cette question de priorité se dessine aussi une autre réalité : celle d'un manque de ressources humaines et financières (Schöpfel 2018b).

2.2. Un rapport privilégié à la publication

En 2017, l'éditeur Springer Nature (Stuart et al. 2018) a mené une des plus vastes enquêtes internationales sur les pratiques de partage des données de recherche (7719 répondants)[289]. L'objectif était de comprendre quelles étaient les pratiques des chercheurs en matière de partage de données, au moment où ceux-ci publient un article scientifique. L'étude a montré que 63% des répondants partageaient les données sous-jacentes à une publication dans des entrepôts de données et/ou comme matériel supplémentaire à l'article. Les chercheurs en biologie étaient les plus nombreux à partager ces données (75% d'entre eux), suivis par les sciences de la Terre (68%), les sciences médicales (61%), les sciences physiques (59%) et les « autres sciences » (46%), rassemblant sciences humaines, sciences sociales, informatique et mathématiques.

Dans notre enquête, la proportion de chercheurs partageant les données sous-jacentes aux publications est légèrement supérieure : 70% (le tableau 14 détaille la répartition par laboratoire). En SHS, rares étaient les chercheurs à avoir déjà fourni, à la demande de l'éditeur, les données liées à une publication (5 chercheurs sur 14). Les sciences humaines et sociales étant sous-représentées dans notre panel (3 laboratoires sur 13), ces résultats ne sont toutefois pas forcément significatifs. A l'inverse, dans les autres domaines, une majorité de

288 Voir supra, quatrième partie, 2.5.1.1, p.181

289 Un questionnaire a été envoyé à chacune des 249 000 personnes inscrites à nature.com, biomedcentral.com et springer.com

chercheurs a déjà été confrontée au moins une fois à ce type de demande (23 chercheurs sur 32).

Grand domaine scientifique	Laboratoire	Discipline du laboratoire	Nombre de chercheurs rencontrés	Nombre de chercheurs ayant déjà partagé des données sous-jacentes à une publication	Proportion
Vie & Santé	Laboratoire 1	Biologie	1	1	100 %
	Laboratoire 3	Chimie et biologie	3	3	100 %
	Laboratoire 4	Chimie et biologie	5	5	100 %
	Laboratoire 5	Écologie	3	3	100 %
	Laboratoire 6	Neurosciences	4	4	100 %
	Laboratoire 7	Neurosciences	1	1	100 %
Sciences & Technologies	Laboratoire 8	Astronomie	7	2	29 %
	Laboratoire 9	Sciences de l'ingénieur	5	2	40 %
	Laboratoire 10	Sciences de la Terre	3	2	67 %
Sciences humaines et sociales	Laboratoire 11	Géographie	4	3	75 %
	Laboratoire 12	Sciences sociales	8	2	25 %
	Laboratoire 13	Théologie	2	0	0 %
			46	**28**	**70 %**

Tableau 14 : Proportion de chercheurs par laboratoire à avoir déjà partagé des données sous-jacentes à une publication

Les données sont majoritairement partagées sous forme de fichiers liés à l'article, dans la section appelée « matériel supplémentaire ». Dans le panel, seuls les biologistes et les astronomes ont recours à des banques de données (ou entrepôts de données). Pour les astronomes, ce sont les observatoires (là où sont situés les télescopes), qui se chargent de rendre les données disponibles (chaque observatoire dispose de sa propre archive de données). On peut citer notamment le télescope Canada France Hawaii. Les chercheurs ne sont donc pas directement à l'initiative du partage. En biologie, ce sont essentiellement les données omiques qui sont concernées par le dépôt dans des banques de données. Lorsque les biologistes soumettent un article dans une revue prestigieuse, ils sont contraints de mettre à disposition les séquences ADN. L'article ne sera pas publié, tant que celles-ci n'auront pas été déposées dans une banque de données dédiée (GenBank, en l'occurrence).

Il convient néanmoins de nuancer ces résultats. Car, si 70% des chercheurs interrogés disent mettre à disposition les données sous-jacentes aux publications, il ne s'agit pas de pratiques systématiques. A l'exception des biologistes, cela est même relativement rare. Les chercheurs rendent leurs données publiques, lorsque cela leur est demandé – en particulier par les éditeurs et les relecteurs. Dans ce cas, ils suivent un protocole bien défini par la revue (dépôt dans une banque de données particulière ou dépôt comme matériel supplémentaire à l'article). Cette tendance fait écho aux résultats de l'enquête réalisée par Nicholas et al. : plusieurs des jeunes chercheurs interrogés déclarent en effet ne mettre leurs données à disposition que si la revue dans laquelle ils prévoient de publier le leur demande (Nicholas et al. 2017, p.214).

Selon les chercheurs, les revues sont de plus en plus nombreuses à exiger la mise à disposition des données sous-jacentes. La demande des éditeurs s'inscrit dans une logique d'intégrité scientifique. La fourniture des données joue un rôle d'administration de la preuve. Elle serait gage de qualité et permettrait de limiter les cas de fraude.

Or, même s'ils émettent des réserves quant à l'efficacité de ce système[290], les chercheurs respectent cette clause d'ouverture des données. L'éditeur PLOS faisait le même constat, ayant instauré une politique de mise à disposition des données pour ses propres revues (Byrne 2017). Un des chercheurs rencontrés explique qu'il n'est pas dans son intérêt de refuser la requête de l'éditeur. Pour lui, refuser c'est peut-être perdre l'opportunité d'être publié. *« Pour moi, l'idée c'est que ma publication soit acceptée. Donc, si l'éditeur me demande de faire ça, je le fais. Quand on veut publier, on rend les choses fluides »* (chercheur 41) – il s'agissait en l'occurrence de données génétiques, dont l'éditeur requérait le dépôt dans GenBank. La publication est primordiale pour la suite de ses recherches (pour l'obtention de nouveaux financements, pour une éventuelle promotion…). En raison de la prédominance des publications dans le système d'évaluation, les éditeurs semblent donc avoir une forte influence sur la communauté scientifique.

290 Cf. supra, quatrième partie, 2.6.1, p.201

3. Utilisation des services disciplinaires par les chercheurs

Aux services influencés par le mouvement d'ouverture préexiste une autre catégorie de services, qu'on nommera « services disciplinaires ». Cette catégorie rassemble des services très spécifiques, propres à chaque communauté de recherche. Les plus anciens datent des années 1970, comme le Centre de Données astronomiques de Strasbourg (CDS) créé en 1972 et le Grand Accélérateur National d'Ions Lourds (GANIL) créé en 1975. L'évolution constante des technologies pour la collecte, l'analyse et l'échange de données conduit néanmoins au renouvellement régulier de ces services. Les deux sous-parties suivantes présentent des exemples de services disciplinaires : des services ayant pour fonction l'acquisition ou la collecte de données (3.1.1) et des services ayant pour fonction la réutilisation de données existantes (3.1.2). Cette répartition a valeur de proposition. Elle n'est peut-être pas exacte, ni même exhaustive. Il a été difficile de comparer ces services entre eux, car ils sont disparates et spécifiques (à tel point qu'ils ne sont pas connus d'une discipline à l'autre).

Les résultats de l'enquête indiquent que les chercheurs interrogés, lorsqu'ils ont recours à des services de données, se tournent essentiellement vers des services disciplinaires, et non vers les services issus des politiques d'ouverture. Ils acquièrent la connaissance de ces services tout au long de leur carrière grâce à la communication entre pairs.

> *« D'une façon générale, l'information de « où se trouve quoi », c'est quelque chose qui moi m'a beaucoup manqué quand je suis entré au CNRS. Maintenant, petit à petit, avec les différentes personnes qu'on côtoie et avec le bouche à oreille, on finit par trouver ce qui nous intéresse. Mais, en particulier pour les nouveaux entrants, c'est quelque chose qui manque. »* (chercheur 40)

3.1. Exemples de services disciplinaires

3.1.1. Des services d'acquisition de données

Il existe différentes sortes de services permettant l'acquisition de données de recherche. Les chercheurs interrogés dans l'enquête ont mentionné : des plateformes technologiques ; des observatoires astronomiques ; des stations de recherche ; des cohortes de population.

3.1.1.1. Plateformes technologiques

Les plateformes technologiques sont des infrastructures hébergeant un voire plusieurs instruments ou logiciels, gérés par des personnels dédiés. Elles peuvent être indépendantes ou rattachées à une unité de recherche, être destinées aux membres d'un laboratoire ou ouvertes à une communauté plus large.

Elles sont récurrentes dans des disciplines comme la biologie, la chimie et les sciences de l'ingénieur. Il peut s'agir de plateformes pour l'analyse de fragments ADN ou ARN. A l'Université de Strasbourg, l'Institut de Génétique et de Biologie Moléculaire et Cellulaire propose par exemple un service de séquençage haut débit et d'analyse bioinformatique des résultats : la plateforme GenomEast[291].

Ces services sont généralement payants. Les chercheurs sous-traitent l'acquisition de données, moyennant paiement. Ils y ont recours, lorsqu'ils n'ont pas l'appareil adéquat dans leur laboratoire. Cela s'explique :

- Par la diversité des instruments
  Il est rare que les laboratoires possèdent toutes les machines, dont leurs équipes ont besoin. En chimie et en biologie notamment, le recours à des appareils est quasi systématique pour analyser les échantillons (molécules chimiques, prélèvements de sang, protéines, fragments d'ADN…). Chaque machine a des fonctions très spécifiques, comme l'expliquait un chercheur biologiste : *« En cytométrie, on peut soit juste analyser des cellules, soit trier des cellules. Ça n'est pas tout à fait le même appareil. Nous, on a l'analyseur. Donc l'analyse on peut la faire. Par contre, si on*

291 http://genomeast.igbmc.fr/

veut trier les cellules sur un cytomètre, on n'a pas le bon appareil. Là on doit aller à la plateforme de l'IGBMC » (chercheur 9). Pour le tri des cellules, ce chercheur a donc recours à une prestation de service, déléguant l'analyse à la plateforme de l'Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC).

- Par le coût des instruments
  Les instruments sont des dispositifs coûteux, si bien qu'un laboratoire est rarement en mesure d'acheter tous les appareils dont ses équipes ont besoin. Celles-ci choisissent entre achat de l'appareil et paiements ponctuels de prestations de services selon des considérations de fréquence d'utilisation et de facilité d'accès à l'appareil.

Un des chercheurs interrogés a évoqué la difficulté à localiser les instruments existants. Il semble que l'information ne soit pas facile à trouver. *« J'ai dû acquérir une machine pour mes mesures et j'ai appris, après avoir acquis la machine, passé pas mal de temps et dépensé de l'argent pour l'acquérir, qu'on pouvait trouver un modèle sensiblement équivalent pas très loin sur le campus de Cronenbourg »* (chercheur 40). Ce chercheur a alors mentionné l'idée d'un catalogue qui recenserait les appareils détenus dans les laboratoires : *« Si j'étais amené à utiliser un catalogue, ce serait plus pour y chercher une compétence, un outil ou un instrument. Si on imagine le même genre de répertoire [que Cat OPIDoR] mais avec différents noms de machines et en cliquant on aurait les labos où on peut les trouver, ce serait super. En tout cas, moi ce serait dans ce sens-là que j'aurais un besoin »* (chercheur 40). Cette suggestion a également été faite par un chercheur en chimie et biologie : *« Parfois les chercheurs auraient besoin de catalogues avec les instruments accessibles, mais ceux-ci devraient être alimentés par les spécialistes du domaine, voire de manière automatisée (si c'est possible), car l'évolution [des technologies] est très rapide »* (chercheur 6).

3.1.1.2. Observatoires astronomiques

En astronomie, l'observation des objets célestes est conditionnée par l'installation de grands télescopes au sol ou dans l'espace. Ces appareils sont si coûteux que leur financement requiert l'association de plusieurs pays. Des observatoires sont ainsi constitués. Ce sont des

organisations scientifiques et techniques, presque toujours intergouvernementales, chargées de concevoir et de gérer des équipements pour l'observation des objets du ciel. L'Observatoire Européen Austral (ESO)[292] est la principale organisation de ce type au niveau européen. Quinze pays européens en sont membres et contribuent en proportion de leur PIB.

Les chercheurs des pays financeurs peuvent utiliser les télescopes pour acquérir des données, en demandant du « temps d'observation ». Ce temps d'observation est généralement attribué à partir d'appels à propositions évaluées et classées par un comité scientifique. Lorsque leur demande est acceptée, les chercheurs se rendent sur place pour acquérir les données dont ils ont besoin. Il arrive également de plus en plus souvent, comme l'ont évoqué les chercheurs interrogés, que l'acquisition des données soit faite « en mode service » : les chercheurs n'ont plus à se déplacer ; les données sont acquises par les équipes de l'observatoire, avant de leur être transmises.

A l'inverse des plateformes technologiques, l'utilisation des télescopes est totalement gratuite pour le chercheur. Les échanges financiers se situent en amont, au niveau des agences spatiales. *« Ça n'est pas complètement gratuit : c'est au niveau gouvernemental que la France contribue chaque année au financement de cet organisme européen. Ça n'est pas au niveau du laboratoire mais au niveau de l'État français – du Ministère des Affaires étrangères »* (chercheur 26).

En astronomie, la mutualisation des dispositifs d'acquisition de données se situe donc à un niveau international. Dans le même registre, on peut citer le domaine de la physique nucléaire et des hautes énergies, bien que cette discipline ne soit pas représentée dans le panel. L'étude des constituants de la matière nécessite des infrastructures telles que le Grand Collisionneur de Hadrons du CERN[293].

292 European Southern Observatory (ESO), https://www.eso.org/public/

293 L'organisation de la communauté des physiciens des hautes énergies autour des grands instruments que sont les accélérateurs de particules a notamment été étudiée par la sociologue des sciences Karin Knorr-Cetina (1999).

3.1.1.3. Stations de recherche

Les stations de recherche sont des infrastructures logistiques, permettant aux chercheurs de réaliser la collecte d'échantillons et l'acquisition de données in situ. Il peut s'agir de flottes navales ou aériennes, de stations en milieux polaires…

Un seul chercheur parmi ceux rencontrés avait recours à ce type de service (chercheur 13). Dans le cadre de ses recherches sur le jeûne des manchots, celui-ci bénéficiait d'un accès à la base Alfred Faure, située dans les Terres australes et antarctiques françaises[294].

Les stations de recherche polaires sont gérées par l'Institut Polaire Français (IPEV). Celui-ci lance chaque année des appels à projets à destination des équipes de recherche françaises (en biologie, géophysique, océanographie…). Les projets retenus bénéficient alors de moyens humains, logistiques, techniques et financiers, permettant aux équipes de se rendre sur place et de réaliser des campagnes d'observation et de collecte de données. Les programmes financés par l'IPEV sont des programmes relativement bien dotés (plusieurs centaines de milliers d'euros par projet, selon le chercheur 13). Car l'enjeu n'est pas seulement scientifique mais aussi géopolitique – ces stations confortent la présence de la France en Antarctique et sa place parmi les leaders scientifiques sur ce continent.

3.1.1.4. Cohortes de populations

Les services d'acquisition de données peuvent aussi prendre la forme de cohortes de population. Celles-ci sont notamment utiles pour des recherches en sciences sociales et en santé.

Une cohorte est un dispositif qui permet de suivre un groupe de personnes de manière longitudinale. Des informations sont recueillies régulièrement sur ce groupe d'individus. Les données ainsi collectées permettent de répondre à des questions de recherche plus ou moins vastes. Une cohorte peut par exemple avoir pour but : d'évaluer les effets de certains facteurs de risque sur la santé des individus ; d'étudier la dynamique des trajectoires individuelles et leurs déterminants ; de suivre l'évolution d'une maladie... En France, l'Inserm est l'un des principaux administrateurs de cohortes épidémiologiques et de santé publique.

294 https://www.institut-polaire.fr/ipev/infrastructures/les-bases/subantarctique/

En démographie, les deux chercheurs rencontrés (chercheurs 47 et 48) ont mentionné l'enquête Elfe (Étude longitudinale française depuis l'enfance)[295]. Cette cohorte porte sur 18 000 enfants nés en 2011. Les enquêtés font l'objet d'études régulières portant sur divers aspects de leur développement (santé, scolarité, alimentation, vie familiale et sociale, environnement…). Les données recueillies visent à mieux comprendre les processus de développement et de socialisation des enfants.

En éthologie, un chercheur (chercheur 16) travaillait à partir de la cohorte RECORD, gérée par l'Institut Pierre Louis d'Épidémiologie et de Santé Publique de l'Inserm. Cette cohorte vise à analyser les disparités de santé existant en Île-de-France, en étudiant l'influence des environnements géographiques de vie. Pour ce faire, des données sont collectées sur les caractéristiques physiques de l'environnement, les services présents ou non à proximité et les interrelations sociales au sein des quartiers. En s'appuyant sur la cohorte RECORD, le projet porté par le chercheur 16 avait pu acquérir des données à partir d'un échantillon déjà constitué, permettant un gain de temps et de ressources non négligeables.

Dans les deux exemples précités, les chercheurs ont eu accès aux cohortes de population grâce à la mise en place d'une collaboration. Les cohortes ne sont pas des services qui se monnaient. Elles sont accessibles sur sollicitation ou sur appel à projets. Dans ces dispositifs, soit les chercheurs participent à l'enquête depuis sa préparation jusqu'à l'analyse des résultats, soit ils interviennent seulement au moment de la phase d'analyse des données (ces dernières ayant été collectées en amont par la structure responsable de la cohorte). Dans le cas des cohortes les mieux dotées, les chercheurs peuvent être amenés à bénéficier du soutien de personnels dédiés pour constituer le panel de participants et gérer les données collectées (nettoyage, documentation, conservation…).

3.1.2. Des services de réutilisation de données

Il existe différentes sortes de services offrant aux chercheurs la possibilité de réutiliser des données existantes – que celles-ci soient d'origine scientifique ou non. Les chercheurs

295 L'enquête Elfe (https://www.elfe-france.fr/) est pilotée par l'Institut National d'Études Démographiques (Ined) et l'Institut National de la Santé et de la Recherche Médicale (Inserm).

interrogés dans l'enquête ont mentionné : des banques de données d'enquêtes en sciences sociales ; des banques de données omiques ; des archives d'observatoires astronomiques.

3.1.2.1. Banques de données d'enquêtes quantitatives en sciences sociales

Les chercheurs en sciences sociales travaillant à partir d'enquêtes quantitatives utilisent les données mises à leur disposition par le réseau Quetelet de PROGEDO.

En démographie, les chercheurs 47 et 48 s'intéressent en particulier aux enquêtes de statistique publique. En tant que chercheurs, ils bénéficient d'un accès sur demande à ces enquêtes. *« Toutes ces grandes enquêtes de statistique publique nous sont accessibles à peu près un an après la fin de la collecte. C'est quand même génial »* (chercheur 48). Le chercheur 48 considère en effet ce dispositif comme avantageux, comparé aux conditions d'accès aux données administratives, qu'il peut être amené à utiliser. Les chercheurs souhaitent en effet parfois analyser les données que collectent les administrations, car celles-ci recèlent des informations pertinentes voire uniques pour leurs recherches. Un des chercheurs rencontrés (chercheur 45), en géographie, réutilisait par exemple les résultats de l'enquête Ménages Déplacements (EMD), pilotée par le CEREMA, dans le cadre de ses recherches sur la sédentarité des populations urbaines. Néanmoins, les données administratives ne sont pas mises à la disposition de la recherche de manière systématique comme le sont les données du recensement et de la statistique publique. Les chercheurs sont donc contraints, lorsqu'ils souhaitent accéder à une base de données, de contacter par eux-mêmes les administrations et d'établir une convention avec elles pour l'exploitation des données. *« Rien n'est prévu pour [les données administratives]. Elles ne sont accessibles que via des conventions, qu'on va faire dans le cadre d'un partenariat de recherche. Certains fichiers comme le Sniram sont prévus pour faire de la recherche. Le Sniram c'est un fichier de la caisse primaire d'assurance maladie, qui va retracer toutes les trajectoires de soins et de prescriptions des individus. Mais, la plupart du temps, les administrations ne livrent aucune donnée (le Conseil départemental par exemple – on a parfois besoin de leurs données pour traiter de questions sociales). Ça n'est que dans le cadre de conventions [qu'on peut avoir accès aux données] »* (chercheur 48).

Par ailleurs, n'étant pas collectées dans un but initial de recherche, les données administratives présentent plusieurs inconvénients. C'est ce qu'explique le chercheur 47 : *« Le principal problème (c'est le problème des bases administratives), enfin les principaux problèmes, ce sont des données qui sont collectées dans un objectif de gestion et pas dans un objectif de produire des données statistiques. Donc il y a des informations qui, par exemple, ne sont pas collectées, parce qu'elles ne sont pas pertinentes par rapport à la gestion, mais qui pourraient nous intéresser. Il y a des données qui sont collectées mais auxquelles on n'a pas accès. On a des fichiers réduits, on ne nous donne pas toutes les informations, parce qu'elles sont trop personnelles. Même si elles sont anonymisées, on pourrait éventuellement remonter à la personne. Et il y a les classiques : des données incomplètes (des indicateurs ne sont pas renseignés), des données en doublon (un défaut d'enregistrement)... Et la vraie difficulté c'est que ça n'est pas une seule personne qui va rentrer toutes ces données administratives. Ce sont plusieurs personnes, souvent dispatchées sur le territoire, qui ont chacune été plus ou moins formées à rentrer les données et qui ont plus ou moins chacune une pratique professionnelle un peu différente. »*. Il n'existe donc pas de services de données pour faciliter la réutilisation des données administratives.

3.1.2.2. Banques de données omiques

Les banques de données omiques sont généralement des dispositifs internationaux. Les plus connues sont GenBank et Gene Expression Omnibus (GEO), maintenues par le NCBI aux États-Unis, et les bases du European Bioinformatics Institute (EMBL-EBI), telles que UniProt, PDBe et Ensembl[296]. Ces bases donnent accès à des données de génomique, de transcriptomique, de protéomique et de métabolomique, regroupées sous le terme générique de « données omiques ». Elles sont librement accessibles en ligne et sont exploitées de manière intensive par les chercheurs du domaine de la bioinformatique. Ceux-ci créent des applications pour la visualisation et le traitement des données omiques (c'est le cas de l'équipe de bioinformaticiens, dont est responsable le chercheur 30). Les banques de données omiques leur servent de base pour l'alimentation de leurs applications, destinées à être utilisées par des biologistes.

296 https://www.ebi.ac.uk/services

3.1.2.3. Archives des télescopes

En astronomie, les données générées par les grands télescopes (décrits plus haut[297]) sont systématiquement ouvertes à la réutilisation. Elles ont été acquises initialement par des chercheurs en vue d'étudier une question de recherche particulière. Une fois cette première exploitation réalisée, elles sont conservées par les archives des télescopes et mises à disposition de la communauté scientifique internationale. Il arrive régulièrement que des chercheurs en astronomie les réutilisent pour répondre à leurs propres problématiques de recherche.

Les données sont disponibles entre un et trois ans après leur acquisition : *« Pour les données qui sont observées de manière générale au télescope, c'est généralement un an. Les grands relevés, ça dépend de la structure des collaborations. Ça nécessite quand même énormément de travail pour que ça soit mis en place. Je pense que, au moins au tout début, c'est plus d'un an. Ce qui se fait aussi de plus en plus, c'est que la première distribution de données est publique au bout de 3 ans, et puis peut-être que la troisième ou la quatrième sera publique beaucoup plus rapidement. Parce qu'il faut le temps d'analyser les données qui nous arrivent brutes. Il faut le temps de les calibrer, de les comprendre, de les analyser et, après, de faire la science dessus. Au tout début, ça prend toujours beaucoup de temps pour caractériser les données qu'on récupère »* (chercheur 28).

3.1.2.4. Points communs des services de réutilisation

Les services de réutilisation présentés ci-dessus ont pour point commun de porter sur des données dont le potentiel d'exploitation est riche ou, du moins, qui ont été perçus comme utiles par les communautés de recherche. La mise en place de services de réutilisation se justifie également par le coût élevé d'acquisition de ces données (une enquête de statistique publique peut coûter plusieurs millions d'euros, par exemple). Plutôt que de les acquérir une seconde fois, les communautés de recherche ont jugé plus utile de les conserver et de les partager. Le chercheur 47 (en démographie) expliquait que les données des enquêtes de population acquéraient avec le temps une valeur historique. *« Quand je travaille sur les*

297 Voir supra, 3.1.1.2, p.230

enquêtes « conditions de vie »[298]*, je compare au moins à la dernière enquête, si ce n'est pas aux deux dernières. Parce que c'est toujours intéressant de voir s'il y a eu une évolution dans le temps. Donc les données ne sont pas obsolètes – tant qu'il y a comparaison possible dans le temps, c'est-à-dire que la question a été posée de la même manière ».*

Émanant directement des besoins des communautés, les services de réutilisation de données sont appréciés des chercheurs interrogés dans l'enquête. Leur utilisation reste néanmoins conditionnée par la qualité des données. *« Le problème, du coup, de la donnée c'est qu'il faut qu'elle ait des caractéristiques et, notamment pour la recherche, il faut qu'elle ait une qualité suffisante. La qualité, pour nous démographes, vient de la collecte, c'est-à-dire en amont (d'où vient la donnée). C'est cette capacité à dire : « cette information est bonne ou n'est pas bonne, parce qu'elle a été collectée comme ça ». Parce que derrière la donnée, en sciences sociales, il y a une question, il y a des logiques sociales, politiques..., qui viennent que tout d'un coup on a cette information. [...] La qualité est liée au processus de production. C'est comme en sciences fondamentales finalement. Si vous avez un instrument de mesure, qui vous donne un tas d'informations, par exemple une balance, il faut que la balance soit bonne. Si elle est déréglée, votre mesure ne vaut rien. [...] Est-ce que cette information est suffisamment fiable pour que je puisse l'agréger et en dire des choses en recherche ? Pour moi, ce serait ça la qualité de la donnée »* (chercheur 48).

Au vu des témoignages des chercheurs, l'astronomie, la démographie et les sciences omiques semblent être les seules disciplines, parmi celles investiguées dans l'enquête, où la réutilisation de données passe par des services. Dans les autres disciplines sondées, les chercheurs disent établir des contrats avec des partenaires pour pouvoir utiliser les données de ces derniers. C'est le cas de l'informatique. Un chercheur dans ce domaine expliquait que ses travaux étaient quasi systématiquement basés sur des projets de recherche appliquée, avec un prestataire, pour lequel il réalisait la mise en forme et l'analyse des données (chercheur 34).

298 Statistiques sur les ressources et conditions de vie : https://www.insee.fr/fr/metadonnees/source/serie/s1220

3.2. Caractéristiques des services disciplinaires

3.2.1. Une majorité de services internationaux

Les chercheurs interrogés n'utilisent pas uniquement des services français. Ils ont également recours à des services étrangers, européens ou multinationaux, comme :

- Le Très Grand Télescope de l'ESO (chercheur 28)
- L'entrepôt Genbank du NCBI (le chercheur 41 en écologie ainsi que les chercheurs des laboratoires 6 et 7 de neurosciences et les laboratoires 1, 3, 4, 9 en chimie et biologie)
- Protein Data Bank (chercheur 6)
- Les bases de données taxonomiques Tropicos et The Plant List (chercheur 41)
- Eurostat (chercheurs 47 et 48 en démographie)

C'est pourquoi plusieurs chercheurs se sont dits plus intéressés par un catalogue qui recenserait des services à l'échelle internationale que par un catalogue comme Cat OPIDoR, se limitant au périmètre national[299].

3.2.2. Des services publics et privés

Les services auxquels ont recours les chercheurs peuvent être de nature publique ou privée. Tous les services disciplinaires listés dans la partie 3.1 sont des services émanant de structures publiques. Plusieurs chercheurs nous ont dit avoir également recours à des services issus du secteur privé. C'est le cas par exemple d'un chercheur en géochimie, qui avait sous-traité une partie de l'analyse des données à un bureau d'études privé : *« Dans le projet ANR, on a aussi un partenaire privé, qui est un cabinet d'étude de sols. Là c'est une prestation : pour 10 000€ on travaille ensemble, mais ils nous rendent une carte pédologique du bassin. C'est une prestation de service. On a prévu un certain budget pour un certain rendu »* (chercheur 37).

Le recours à des services privés s'explique, dans l'un des cas rencontrés (chercheur 41), par des considérations de coût : *« Le séquençage est sous-traité à un prestataire de services, […] hors campus. Il y a des services campus, mais actuellement plus vous êtes une grosse boîte,*

299 Cf. supra, 1.2, *Un périmètre trop restreint*, p.219

plus vous êtes capables de proposer des tarifs bas. Parce qu'ils font ça en routine. […] Les UMR qui font de la prestation de service vers d'autres UMR mettent peut-être en place un tarif campus, mais pour l'instant le tarif campus qu'elles mettent en place n'est pas super avantageux ». Pour ce chercheur, le choix de faire appel à un prestataire privé se justifie donc par les prix concurrentiels que celui-ci propose.

3.2.3. Le recours à des services, mais pas seulement

Les services ne sont pas le seul recours des chercheurs, lorsque ceux-ci souhaitent utiliser des données qu'ils ne peuvent acquérir ou collecter avec les moyens de leur laboratoire. Lorsqu'il s'agit de besoins ponctuels, les chercheurs disent parfois passer par le biais de « collaborations »[300], pour acquérir des données avec l'instrument de tel autre laboratoire ou pour réutiliser les données générées par telle entreprise ou telle administration.

En sciences de la Terre, le chercheur 40 estime acquérir 20% de ses données en externe, en partie par le biais de prestations de services et en partie par le biais de collaborations. *« A une courte majorité, c'est plutôt l'option de collaboration. On envoie les échantillons à d'autres collègues pour faire les caractérisations et ils sont associés à la publication. On a aussi, comme vous le disiez, des plateformes, auxquelles on paie un service analytique. Par exemple moi, là où je vais souvent, c'est à l'Université d'Aix-Marseille, où il y a une plateforme avec un ingénieur dédié et où on paie le service[301]. Dans ce cas, on n'associe pas l'ingénieur à la publication, dans le sens où c'est vraiment un service effectivement »* (chercheur 40).

Le recours aux collaborations s'explique également par l'hyperspécialisation des thématiques de recherche. Certaines équipes mettent au point des technologies très pointues, pour répondre à leurs propres problématiques de recherche. Ces technologies, à l'instar de celle élaborée par l'équipe du chercheur 3 dans le domaine microfluidique, sont relativement rares et attirent donc régulièrement des chercheurs travaillant sur des thématiques similaires. *« Ce sont des instruments qui sont fabriqués par les laboratoires. […] La machine qui est fabriquée ici est une machine unique. Bon, il y en a d'autres dans le monde. Mais celle qui est là a vraiment*

300 Voir supra, quatrième partie, 2.6.2, p.206

301 La plateforme en question est le Centre Pluridisciplinaire de Microscopie électronique et de Micro-analyse (CP2M).

été personnalisée pour nos applications. Il y a très peu d'équipes dans le monde qui ont ce type d'équipement. Parce qu'il faut l'équipement, parce qu'il faut la personne qui sache le fabriquer, le maintenir, parce qu'il faut toutes les compétences autour (il faut notamment être capable de fabriquer les puces qu'on manipule dessus)... Enfin, il faut énormément de choses. Il faut de la chimie, c'est très, très complexe. Donc on ne parle pas ici d'une technologie qu'on commande sur catalogue. Il y a des sociétés qui proposent de vendre la machine, clés en main, pour x centaines de milliers d'euros. Mais ça n'est pas que ça. En fait, la vraie valeur de la technologie, ça n'est pas la machine. La vraie valeur c'est le savoir-faire qui a été développé à côté. Il y a des petits trucs, qu'on n'écrit pas dans les publis, parce que ce sont des détails. Mais ce sont des détails qui font que ça fonctionne ou que ça fonctionne mal. […] Donc tout ça c'est ce qu'on appelle le savoir-faire. C'est le secret de fabrication, on va dire, qui fait que, même si on laissait les clés de la machine à quelqu'un d'autre, qui n'est pas de l'équipe, a priori ça aurait du mal à fonctionner, parce qu'il y a tout ce savoir-faire qui a été développé autour. Or ce savoir-faire, et bien justement, les gens aujourd'hui le recherchent. […] Donc, ce qui va se passer, c'est que, quand il y a des opérations de communication (par exemple un congrès ou ce genre de choses), les congrès sont naturellement choisis en fonction des personnes qu'on veut approcher ou pas, et ces gens-là vont dire « tiens, mais c'est ça qu'il nous faut ». Là, par exemple, j'ai quelqu'un qui est en visite pour un mois. Il vient d'Allemagne et il a passé trois ans de thèse à galérer, parce qu'il voulait essayer de faire ce qu'on fait, mais sans la technologie » (chercheur 3). Les chercheurs, n'ayant ni la technologie ni le savoir-faire, vont donc vouloir s'associer à son équipe, afin de répondre à leur problématique de recherche.

Comme évoqué dans la quatrième partie (2.6.2, p.206), la collaboration suppose que l'accord soit avantageux pour les deux parties.

- En échange du prêt d'un instrument, la personne est généralement ajoutée aux co-auteurs de la publication qui en résulte.
- En échange de la fourniture de données, le chercheur délivre au producteur les résultats de l'analyse, parfois en réponse à une commande prédéfinie à l'avance.

Toutefois, selon le chercheur 9, l'échange n'est pas systématique. Tout dépend du degré d'investissement demandé.

> *[>Chercheur 9]: « Ça peut nous arriver aussi d'utiliser des choses qui sont en dehors de plateformes et qu'on peut trouver en neurochimie ou dans d'autres domaines. Là on envoie un mail aux gens et on demande.*
> *[>Enquêteur]: Est-ce que ça signifie que vous ajoutez la personne aux co-auteurs ?*
> *[>Chercheur 9]: Ça dépend. Si c'est pour utiliser ponctuellement un appareil de chez nous, qui se trouve dans l'institut, non. Si c'est pour utiliser régulièrement quelque chose qu'on n'a pas (par exemple, en neurochimie, un microscope), si la personne participe un peu au projet, nous montre comment utiliser l'appareil, oui. Si c'est juste nous montrer comment marche l'appareil, elle sera dans les remerciements. Si c'est participer au projet, parce que c'est un appareil qui nécessite par exemple de faire de l'imagerie in vivo sur un petit animal, comme c'est quelque chose qu'on n'a pas, on va mettre en place une collaboration. Parce que, là, la personne va aussi nous aider à designer la manip. Comme c'est un appareil qu'on n'a pas, on peut commettre des erreurs sur « quels sont les bons contrôles ? », « comment je dois faire ? », etc. Donc là on va voir la personne qui, elle, connaît bien sa machine et qui va vraiment nous apporter un plus sur ce qu'on doit faire dans notre projet pour que ça soit bien. Donc forcément la personne sera dans les auteurs ».*

3.2.4. Vers une partition entre producteurs et utilisateurs de données ?

Les chercheurs – notamment les jeunes chercheurs – travaillent de moins en moins à partir des données brutes. Leurs recherches ont pour point de départ des données dérivées (extraites par d'autres chercheurs).

Dans le cas de l'astronomie, les données brutes sont des images (acquises dans le cadre de grands relevés du ciel). Et les données dérivées sont les informations extraites de ces images (la taille des étoiles par exemple). Les chercheurs sont donc de moins en moins familiers des images et des instruments permettant de les acquérir. Un chercheur s'interroge sur leurs

capacités à participer aux grands relevés. Si l'on va vers une hyper-spécialisation, qui sera en mesure d'acquérir les données brutes et d'en extraire les données dérivées ?

> *« Le petit bémol c'est qu'on s'oriente peut-être, comme je le disais, vers une communauté qui réfléchit moins aux problématiques de ce que sont des données, de comment on en extrait de l'information. Comme on se spécialise de plus en plus, je ne sais pas comment on va vraiment répondre à cette problématique.*
>
> *On aura toujours besoin de gens dont le boulot sera de réduire les données, peut-être même de manière permanente (ils ont un poste et leur tâche de service c'est de réduire des données ou d'aider à créer des algorithmes pour réduire des données). Mais pour faire ça, il faut savoir ce que c'est des données. Il faut l'avoir déjà fait avant. Typiquement, si on est recruté sur un poste comme ça, c'est parce qu'on a déjà l'expérience.*
>
> *Alors qu'on s'oriente de plus en plus vers des grandes collaborations avec des accès aux données, où on ne regarde plus l'image ou le spectre [acquis par le télescope], comment garde-t-on cette technicité, cette expérience avec des gens qui savent le faire ? C'est presque un travail d'ingénieur. Peut-être que c'est ça la réponse. Peut-être qu'on va s'orienter vers : les chercheurs vont faire leurs recherches, c'est-à-dire cadrer ce que l'on veut retirer des informations ; et après on va embaucher des ingénieurs pour faire l'analyse des données. »* (chercheur 28)

Selon ce chercheur, une des évolutions possibles est la partition entre : des ingénieurs, qui acquièrent les données brutes et en extraient les données dérivées en autonomie ; et des chercheurs, qui *« font la science »*, autrement dit qui interprètent les données dérivées.

3.3. Quels services complémentaires pour les données de la recherche ?

Au cours de l'enquête, plusieurs chercheurs ont évoqué le besoin de services de données complémentaires. En voici quelques exemples :

Des espaces collaboratifs en ligne, pour que les chercheurs d'un même projet de recherche puissent travailler sur des données communes – a fortiori lorsque les équipes sont éloignées géographiquement. Le chercheur 33 travaille avec des médecins de Hanovre en Allemagne. Pour ce projet, ils bénéficient d'un partenariat privilégié avec une société belge, qui met à leur disposition une plateforme logicielle pour l'analyse collaborative d'images histologiques. Cette situation est relativement exceptionnelle. Selon le chercheur, ces outils sont peu répandus dans le milieu scientifique. Premièrement parce qu'ils sont spécifiques à un type de données et à une discipline particulière. Deuxièmement parce que leur acquisition a un coût, que ne peuvent pas toujours endosser les projets de recherche et les laboratoires (d'autant plus si chaque type de données requiert un logiciel spécifique). Cela peut fonctionner si l'outil a été conçu par une autre unité de recherche (en informatique par exemple) dans le cadre d'une collaboration, comme c'est le cas du projet dont fait partie le chercheur 33.

> *« C'est un outil qui est super bien. Il a été développé en Belgique. Il s'appelle Cytomine. Les données (les images) sont chez moi ici sur le serveur. Les annotations sont partagées (via un site web). C'est-à-dire que les médecins à Hanovre peuvent faire des annotations, corriger mes annotations... Moi je fais des annotations et ils les voient en temps réel. Il suffit que je leur crée un compte sur le serveur et ils peuvent avoir accès à tout. Ici par exemple j'ai tous mes projets. Quand je prends un projet, j'ai la liste de toutes les images qui sont dans le projet. Je sais tout de suite combien il y a d'annotations. Je peux aller dessus. Et eux peuvent faire pareil à distance. Ils peuvent venir et regarder mes annotations et les leurs, ils peuvent en rajouter. Donc ça c'est vraiment un outil indispensable pour nous. Sinon, je ne sais pas comment on ferait. S'envoyer par mail des trucs, ça devient... C'est un outil développé vraiment pour ça à l'Université de Liège. Maintenant ils ont créé ce qu'ils appellent une coopérative. Ça n'est pas une entreprise, ça n'est pas une association. C'est un truc belge. C'est entre les deux. Pour nous, c'est gratuit. Parce qu'on fait de la recherche et qu'on travaille avec eux depuis le début. Après, sinon, ils font des prestations de services pour les entreprises. Nous, on héberge chez nous le serveur, on l'administre nous-mêmes, donc c'est gratuit. Ce genre d'outil, c'est vraiment ça qu'il faudrait avoir de plus en plus. Parce que tous les projets qu'on fait ne sont*

> *jamais avec des gens d'ici. J'ai des gens en Italie, en Israël, aux Pays-Bas, à Paris, à Toulouse... S'il fallait s'envoyer des mails, à chaque fois ce serait un cauchemar. »* (chercheur 33)

Des espaces de discussion en ligne, dédiés au partage de savoir-faire entre chercheurs. Ce besoin est remonté d'un chercheur en biologie (chercheur 9) :

> *« Ce que je trouve génial, c'est le partage de toutes les expériences. C'est-à-dire le partage des expériences qui ne donnent pas lieu à des résultats scientifiques, mais qui sont toutes les informations concernant le déroulement d'expériences. Quand j'étais en thèse, parfois on se lançait dans une manip et ça ne marchait pas, on ne savait pas pourquoi – parfois c'est pour des raisons de choses qui ne vont pas être publiées, parce que ça n'est pas tellement intéressant scientifiquement. Maintenant il suffit de taper dans Google et on trouve dans ResearchGate toutes les discussions entre chercheurs : « est-ce que vous avez déjà observé ça » ou « est-ce que vous savez que, quand je stimule mes cellules de telle manière, ça marche moins bien que quand je fais ça », etc. C'est plus du partage d'infos un peu techniques. Ça c'est génial. Il faut le partager au maximum. […] Il pourrait y avoir des plateformes. Avec des sous-branches thématiques, par exemple « cytométrie en flux », « culture de cellules humaines »... Parce que là [dans ResearchGate] les questions se posent un peu dans tous les sens. Il n'y pas de vraie plateforme qui existe, sur laquelle on peut poser sa question ».*

Aux dires de ce chercheur, les scientifiques semblent actuellement utiliser le réseau social ResearchGate[302] pour échanger entre eux sur des questions de méthodologie. Selon le chercheur 9, ResearchGate n'est cependant pas le support idéal, car telle n'est pas sa fonction première. Un outil dédié, centralisant et organisant les discussions par thématiques, serait plus adapté.

302 https://www.researchgate.net/

Des services de proximité pour les laboratoires de sciences humaines et sociales. Selon le chercheur 48, directeur de la Maison Interuniversitaire des Sciences de l'Homme en Alsace (MISHA), les chercheurs en sciences humaines et sociales sont amenés à traiter des données de plus en plus nombreuses et de plus en plus hétérogènes.

> *« En sciences humaines et sociales, la plupart des chercheurs ne savent pas programmer. C'est un problème. [...] Il faut développer une culture du traitement de la donnée. Dans le master Démographie que je dirige, les étudiants ont maintenant des cours d'algorithmique. Parce que j'estime qu'il faut qu'ils sachent parler le langage qui permet de traiter les données. [...] Si tu ne sais pas ça, on arrive à un autre truc : c'est qu'il faut qu'on recrute des gens qui le font pour toi et ça l'Université et le CNRS ne le financent plus. Ils vont peut-être financer un poste mutualisé mais c'est tout. Donc il faut savoir le faire. Nous, notre idée c'est de former les futurs chercheurs. »*

En parallèle de cette formation des étudiants au traitement des données, le chercheur 48 propose de développer des services au sein de la MISHA, avec notamment le recrutement de personnels mutualisés (informaticiens, statisticiens, traducteurs...). Cette offre de service serait destinée à la génération des chercheurs actuels, qui n'ont pas forcément la maîtrise des logiciels de traitement des données (R[303] par exemple). Ces nouveaux personnels auraient une fonction de relais entre les chercheurs et les infrastructures de données (telles Huma-Num et PROGEDO). Selon ce chercheur, *« les services locaux sont une nécessité, car chaque site a sa logique, ses demandes... »*.

Des supports dédiés à la description fine des données. En sciences humaines et sociales, le chercheur 46 a évoqué un projet de création de revue, destinée à valoriser les matériaux de terrain collectés par les chercheurs (cartes, photographies, articles de presse, tracts, notes d'observation, extraits d'entretiens…). Ce projet est né d'un sentiment de « montée en théorie » des revues classiques ; il traduit, par conséquent, le besoin chez les chercheurs de publier une description plus fine de ces matériaux.

303 R Studio : https://www.rstudio.com/

> *« Là je participe à une nouvelle revue qui, j'espère, verra le jour et qui s'appelle Sources[304]. […] L'idée c'est justement d'écrire des articles plus méthodologiques sur les conditions de collecte des données – ou plutôt d'une donnée – et de présenter la donnée brute – enfin, pas « brute » mais in extensu. On a pas mal d'historiens et d'anthropologues dans le comité, qui disent : « j'ai parfois une archive énorme et aucune revue n'accepte de la publier ». Donc c'est un peu cette idée d'avoir dans la revue une réflexion plus méthodologique sur la source et de donner accès à la source qu'on essaie de traiter. Mais il y a déjà un traitement et un choix. […] L'idée c'est [...] de répondre à une demande de plus en plus de chercheurs, qui disent « dans les revues, on nous demande de plus en plus de monter en théorie ». A partir d'une phrase, on écrit un article en anthropologie par exemple. Et on ne veut pas ça. On aime nos données, on aime nos entretiens, on aime en réciter de gros morceaux. Et à chaque fois on se fait couper les entretiens, on se fait couper ces données justement plus brutes, parce qu'elles n'ajoutent pas grand chose à l'économie du texte [...]. Donc on aimerait un lieu où il y ait vraiment une description et une place pour la description et le matériau ethnographique. C'est un peu la partie militante de la revue. Donc il faut aussi des outils techniques pour pouvoir rendre consultables certaines archives, certaines données, qui sont lourdes ou... Voilà. Donc il y a une réflexion qui se fait. »* (chercheur 46)

La revue serait donc conçue spécialement pour présenter des données, issues en l'occurrence de « matériel ethnographique ». L'objectif premier ne serait pas de présenter des données dans le but de voir celles-ci réutilisées par d'autres chercheurs (le chercheur 46 considère que, dans son domaine, c'est rarement le cas). La revue servirait plutôt à présenter une analyse des données. Ce serait donc la méthodologie décrite qui pourrait avoir un intérêt pour les lecteurs. On pourrait penser que cette revue s'apparente à un *data journal*[305]. En réalité, elle s'en distingue, dans la mesure où elle inclut une partie dédiée à l'analyse des données – partie

304 Initié par les UMIFRE (Unités Mixtes des Instituts Français de Recherche à l'Étranger) installées en Afrique, ce projet de revue porterait en l'occurrence sur les matériaux de la recherche sur l'Afrique, avec un enjeu additionnel de conservation et d'accès aux sources.

305 Un *data journal* est une revue scientifique, dans laquelle sont publiés des *data papers* (voir note de bas de page 168, p.94).

qu'on ne trouve pas dans un *data paper*. Ce type d'initiative va donc dans le sens d'une mise en valeur des données, mais dont l'objectif n'est pas forcément la réutilisation des données.

Conclusion

Cette thèse s'est intéressée aux données de la recherche dans leur contexte de production, à partir d'un panel de 57 chercheurs de l'Université de Strasbourg, avec qui ont été menés des entretiens qualitatifs. La particularité du terme de « données de la recherche » est d'avoir été défini par des acteurs extérieurs au cercle de production des données, rassemblant ainsi sous une même dénomination des entités qui, à l'origine, sont appelées de diverses manières dans les disciplines scientifiques. Cette action de catégorisation sert un objectif politique, aujourd'hui répandu à l'échelle internationale : l'ouverture des données généralisée à l'ensemble des communautés de recherche. Cet objectif est devenu l'un des principaux axes des politiques scientifiques française et européenne. Celles-ci brandissent comme exemples d'ouverture les pratiques de partage ayant lieu en astronomie, en génomique et en physique des hautes énergies. Ces modèles sont forcément plus connus, puisqu'ils requièrent le développement et la pérennisation d'infrastructures d'envergure pour assurer le partage des données. Mais en mettant ces exemples sur le devant de la scène et en les prenant comme modèles, les discours politiques occultent les motivations originelles du partage dans ces communautés et relèguent au second plan des formes de partage, certes moins visibles mais présentes dans d'autres communautés.

Vérification des hypothèses de recherche

Le but de cette thèse a été de mettre en lumière ce qu'occultent les politiques d'ouverture. A savoir, **quelles formes de gestion et de partage existent dans les communautés de recherche et par quoi sont-elles motivées ?**

A cette première question de recherche avaient été associées trois hypothèses, selon lesquelles les modes de gestion et de partage des données seraient influencés **(1)** par le cadre épistémique, **(2)** par le cadre institutionnel, **(3)** par le cadre social.

L'**influence du cadre épistémique (1)** a été confirmée. Les modes de gestion et de partage varient en fonction des méthodologies et des dispositifs de recherche propres à chaque « culture épistémique » (Knorr-Cetina 1981). L'hypothèse initiale prenait l'exemple de l'astronomie, où gestion et partage des données s'articulent autour de grands équipements que sont les télescopes. Nous nous étions demandé si cette configuration se répétait dans d'autres

disciplines utilisant de grands équipements. Dans l'enquête s'est présenté le cas de la démographie, où les chercheurs font eux aussi appel à de grands équipements. A la différence de l'astronomie, ces équipements ne sont pas des appareils de mesure comme les télescopes mais des dispositifs d'enquête de populations comme ceux du Céreq ou de l'Ined. Il s'agit de dispositifs coûteux autour desquels la communauté des démographes se fédère. Des groupes de chercheurs – souvent pluridisciplinaires (incluant démographes, économistes, géographes, épidémiologistes…) – se rassemblent pour créer un questionnaire d'enquête à envoyer à un échantillon de la population et pour ensuite réaliser une analyse thématique des données. Les tâches de structuration et de documentation des données (nettoyage des bases de réponses, préparation des variables...) est fait par ce que Denis et Pontille (2012) appellent des « petites mains », en l'occurrence des ingénieurs affiliés à l'organisme producteur de l'enquête. Une fois les premiers résultats publiés (sous forme d'ouvrage collectif), les données sont rendues disponibles à l'ensemble de la communauté scientifique. Il convient de noter que l'ouverture de données, en astronomie comme en démographie, n'est pas immédiate : elle n'intervient qu'après une période propriétaire de un à trois ans, pendant laquelle l'exploitation des données est exclusivement réservée aux chercheurs associés initialement à la collecte.

Les exemples de l'astronomie et de la démographie confirment donc l'influence des dispositifs instrumentaux sur les modes de gestion et de partage des données : le coût des grands équipements conduit les communautés de recherche à collaborer autour de la collecte des données ; la mise en commun des données rend alors nécessaire l'élaboration de règles de gestion standardisées ; le partage des données au reste de la communauté scientifique permet ensuite de rentabiliser les efforts investis dans la collecte et la standardisation des données. Ces modes de gestion et de partage ont à l'origine été instaurés par les communautés scientifiques. Ils sont progressivement institutionnalisés, avec la création d'unités de service dédiées, si bien que les jeunes chercheurs acquièrent dès leur doctorat cette culture des données communes. Ayant appris à faire de la recherche selon ce modèle, ils ne le remettent pas en cause, ne voyant d'ailleurs pas comment ils pourrait travailler autrement.

L'**influence du cadre institutionnel (2)** a été étudiée à travers le concept de normes (Cacaly 1997 ; Schöpfel 2018a). L'enquête a confirmé l'influence des normes institutionnelles sur la

gestion et le partage des données, quoiqu'à des degrés divers. Il est à vrai dire difficile d'appréhender quelles sont ces normes, car celles-ci proviennent d'instances diverses, prennent des formes différentes et touchent des domaines hétéroclites. Comme le résumait Joachim Schöpfel, *« il ne faut pas imaginer cet écosystème établi selon différents domaines et finalités comme un univers cohérent et fonctionnel. Du point de vue normatif, force est de constater des degrés de normalisation très différents, d'une directive européenne via une loi nationale vers des normes industrielles, des déclarations politiques et des simples incitations »* (Schöpfel 2018a).

Au cours des entretiens, les chercheurs ont évoqué quatre familles d'instances ayant produit des normes qui influencent la manière dont ils gèrent et communiquent les données : la loi française (ainsi que les directives et les règlements européens dont elle est parfois la transposition) ; les établissements de recherche ; l'édition scientifique ; et l'industrie. La loi française encadre notamment le traitement des données personnelles. Son application est régie par la CNIL, qui a un rôle de conseil, d'évaluation et de sanction. Les établissements de recherche – l'Université de Strasbourg et le CNRS par exemple – émettent des recommandations concernant notamment la sécurité des données (dans le but de préserver celles-ci des risques de perte et de vol). Ils peuvent également être à l'initiative de comités d'évaluation, comme les comités d'éthique pour les recherches portant la personne humaine (dans le domaine médical notamment). La sphère de l'édition scientifique fixe, quant à elle, des règles pour la publication des articles et ouvrages. En matière de données de recherche, les éditeurs sont de plus en plus nombreux à demander la mise à disposition des données sous-jacentes aux publications. Enfin, le secteur de l'industrie produit des normes de gestion qui sont reprises par la recherche appliquée. La reproduction de ces normes s'avère en effet indispensable pour le transfert de technologie.

De manière générale, ces normes tendent vers une standardisation des pratiques de gestion et de partage. Elles ne sont toutefois pas intégrées de façon uniforme par les chercheurs. Certains les appliquent sans les remettre en cause. D'autres s'y conforment mais en ont une vision critique. D'autres encore ne les prennent pas en compte (la CNIL, par exemple, reste peu sollicitée des chercheurs – parce qu'elle implique des démarches souvent jugées fastidieuses, et que les enjeux restent moindres). Deux paramètres semblent contribuer à l'intégration des normes institutionnelles dans les activités de recherche des scientifiques : la

perspective d'un avantage symbolique ou économique à respecter telle ou telle norme (si l'application d'une norme sert les intérêts du chercheur, celui-ci sera d'autant plus enclin à se l'approprier) ; la présence de relais contribuant à la mise en application des normes (tels les ingénieurs dans les laboratoires de sciences sociales).

L'**influence du cadre social (3)** a été vérifiée, mais ne confirme que partiellement l'hypothèse de départ. Celle-ci postulait que les chercheurs ne partageaient leurs données qu'avec des pairs de leur connaissance, selon un principe de don/contre-don (Mauss 2007). L'enquête a montré que le partage de données s'inscrivait dans des cercles de confiance mais qu'il pouvait aussi s'étendre à des cercles plus éloignés. Dans les deux cas, une relation s'instaure entre le chercheur et le destinataire des données. Lorsque le chercheur partage ses données avec un pair de sa connaissance ou avec un pair qu'il connaît moins mais avec lequel il va établir une « collaboration », la relation donne lieu à un échange, que l'on peut associer au principe de don/contre-don. Les données sont partagées en échange d'une rétribution (symbolique), qui n'est pas toujours définie en amont de façon explicite. L'image du don/contre-don fonctionne, puisque le chercheur donne, sans assurance certaine de ce qui lui sera attribué en retour. Lorsque le partage s'étend au-delà, avec des pairs plus nombreux par exemple ou avec une personne morale (éditeurs scientifiques, infrastructures de recherche, agences de financement…), la relation établie prend la forme d'un échange contractualisé (convention de recherche, contrat d'édition…). Ce que le chercheur obtient en échange des données est ici clairement posé. Les données sont une monnaie d'échange. Le chercheur décide de s'engager en fonction des avantages que lui apportera le contrat. Dans un cas comme dans l'autre, le partage des données n'est donc jamais désintéressé. Il est pensé de manière stratégique par le chercheur.

La question des services d'appui à la gestion et l'ouverture des données a également été abordée dans la thèse. A l'interface entre décideurs politiques et communautés de recherche, ces services ont été étudiés dans le but de comprendre **comment ils contribuaient à l'ouverture des données et dans quelle mesure ils étaient utilisés par les scientifiques**. Telle était la seconde question de recherche.

Le travail de cartographie des services, réalisé pour la Bibliothèque Scientifique Numérique, avait fait émergé l'hypothèse que les chercheurs en faisaient peu usage.

Ont été interrogés sur cet aspect 46 chercheurs de l'enquête initiale sur les pratiques de recherche. Leurs retours semblent confirmer l'hypothèse de départ. L'exemple de Cat OPIDoR montre en effet que les services développés dans un contexte d'ouverture des données correspondent pour une faible partie à ceux qu'utilisent les chercheurs, quelle que soit leur discipline d'appartenance. Trois explications peuvent être avancées à cela. (1) Le panel de chercheurs interrogés n'est pas représentatif des usages de la communauté scientifique. Les services de gestion et d'ouverture des données sont peut-être davantage utilisés dans des établissements comme l'Irstea ou l'Ifremer, qui ont été précurseurs en termes de libre accès (via l'instauration de mandats d'ouverture, le développement d'infrastructures dédiées…). (2) L'offre de services arrive trop tôt pour rencontrer les besoins des chercheurs. L'offre a anticipé la demande. L'enquête révèle en effet que la gestion et l'ouverture des données ne sont pas considérées comme prioritaires par les chercheurs. L'évaluation et la reconnaissance des activités scientifiques étant principalement fondée sur la publication d'articles et d'ouvrages, les chercheurs ne trouvent que peu d'intérêt à diffuser leurs données. Avec le développement des mandats d'ouverture (en provenance des financeurs de la recherche notamment), il est probable que les chercheurs soient davantage amenés à utiliser les services de gestion et d'ouverture dans les années à venir. Ils auront peut-être recours à l'outil en ligne DMP OPIDoR par exemple, pour rédiger les plans de gestion de données que leur demandent les financeurs. (3) La faible utilisation des services d'ouverture des données peut également s'expliquer par le fait qu'ils sont proposés par des acteurs relativement éloignés des communautés de recherche. Les professionnels de l'information scientifique et technique ont notamment investi cette offre de services. Néanmoins, les chercheurs ne pensent pas spontanément à ce corps de métier pour les seconder dans la gestion et le partage de leurs données. Si les pratiques de partage venaient à se généraliser, sous l'impulsion des politiques d'ouverture, les chercheurs se tourneront-ils vers les services développés par les professionnels de l'IST ? Ou bien feront-ils appel à des services plus proches de leurs communautés ? La thèse a mis en évidence l'existence de services « disciplinaires », nés dans les communautés scientifiques pour répondre à des besoins en termes d'acquisition et de réutilisation de données. Ces services sont davantage utilisés par les chercheurs que les

services de gestion et d'ouverture. Divers scénarios sont donc possibles. Ou bien les chercheurs apprendront à connaître et utiliser les services de gestion et d'ouverture développés (entre autres) par les professionnels de l'IST. Ou bien on observera une évolution des services disciplinaires, qui progressivement se conformeront aux principes d'ouverture des données. Ou bien de nouveaux services disciplinaires seront créés pour répondre aux exigences politiques. Ou bien encore les services disciplinaires existants s'adjoindront les compétences des services de gestion et d'ouverture actuels, afin de proposer des dispositifs sur mesure aux communautés de recherche. L'Open Science semble d'ores et déjà avoir un impact sur les services disciplinaires. Ceux-ci s'efforcent, à la demande de ceux qui les financent, de répondre aux principes d'ouverture des données. C'est le cas des grandes infrastructures de recherche. Tout comme les infrastructures de recherche européennes de la feuille de route de l'ESFRI[306], les infrastructures françaises sont elles aussi peu à peu concernées par les politiques de gestion et d'ouverture. Depuis 2016, elles doivent *« être en mesure de mettre à disposition les données produites, soit immédiatement, soit après une période d'embargo correspondant aux pratiques internationales du domaine concerné »*[307]. La feuille de route nationale précise désormais pour chaque infrastructure le volume et l'accessibilité des données que celle-ci génère[308]. La mise en conformité des grandes infrastructures de recherche est donc en passe de devenir garante de leur pérennité, dans un contexte où l'ouverture de la science fait désormais partie des axes stratégiques de la politique scientifique.

L'étude des politiques publiques et des pratiques de recherche laisse entrevoir une dichotomie dans le rôle qui est attribué aux données. Deux visions coexistent :

- Celle des communautés scientifiques ;

306 Voir supra, deuxième partie, 3.3.1.2, p.80

307 MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION (2016). *Stratégie nationale des infrastructures de recherche*. Edition 2016. http://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/74/5/feuille_route_infrastructures_recherche_2016_555745.pdf (consulté le 19 septembre 2019). Page 7

308 Ces informations sont renseignées dans un encart « Données » dans la fiche de description de chaque infrastructure. Voir par exemple la feuille de route des infrastructures de recherche 2018 (Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation 2018, op. cit.).

- Celle des politiques, en partie relayée par des intermédiaires tels les professionnels de l'information scientifique et technique.

Les politiques ont une vision unique de la science des données. Ils véhiculent l'idée selon laquelle l'ubiquité des données numériques a engendré un changement de paradigme dans l'ensemble des domaines scientifiques. La science est entrée dans un « quatrième paradigme » (Hey et al. 2009) : celui de l'exploration de grandes masses de données. Aussi appelé *data-driven science*, ce nouveau paradigme a pour fondement les données scientifiques, alliées à des capacités de traitement massif qui permettent d'en extraire des connaissances. Dans cette vision, la donnée est donc placée au centre de l'activité de recherche. Celle-ci se structure autour de grandes infrastructures d'acquisition, de traitement et de stockage, où des *data managers* assurent la structuration et l'accessibilité des données. L'utilisation des données va même au-delà de la sphère scientifique : elle s'étend au domaine économique, dans une perspective d'innovation et de croissance. Les instances politiques semblent donc avoir une vision appliquée de la recherche, avec pour préoccupation première la valorisation sociale et économique des résultats scientifiques.

Alors que les politiques considèrent les données comme le point de départ de la science, les chercheurs en ont une vision différente, plus fonctionnelle. Les données constituent des preuves à l'appui de leurs raisonnements. Elles sont donc plus secondaires. L'avènement du *big data* n'a pas modifié les pratiques des scientifiques : ceux-ci restent dans les épistémologies qui sont les leurs (observation, expérimentation, théorisation...), selon leurs disciplines et leurs objets de recherche. Il existe certes des domaines de recherche qui, guidés par la performance des outils de collecte, ont saisi l'opportunité de renouveler leurs méthodes de recherche (simulation numérique, visualisation de données, fouille de données...). Ce sont par exemple la biologie, l'informatique, les sciences de la Terre ou encore les humanités numériques. Ces champs ne représentent toutefois qu'une petite partie de la recherche. Parmi les 57 participants à l'enquête par exemple[309], 19 d'entre eux utilisaient ou généraient des données massives (soit 33%)[310]. Le changement de paradigme ne semble donc pas concerner tous les domaines scientifiques.

309 Voir supra, quatrième partie, tableau 11, p.163

310 Chercheurs 1, 6, 17, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 39, 43 et 56.

Au cours des entretiens, plusieurs chercheurs ont évoqué les enjeux liés à l'interprétation de données qu'ils n'ont pas eux-mêmes générées : réutiliser des données suppose de connaître leur origine, le contexte de leur collecte, d'être conscient des biais qui ont pu s'y introduire… La diffusion des données provoque ce que Edwards appelle des « frictions » (Edwards 2010 ; Edwards et al. 2011). D'importantes transformations doivent être opérées pour rendre les données intelligibles à leurs réutilisateurs. La concurrence entre chercheurs ou le caractère sensible des données sont aussi des sources de frictions à leur ouverture. Ce constat contraste avec la vision simplifiée, qui émane parfois des discours politiques sur l'ouverture des données. Les obstacles techniques, économiques, juridiques et culturels y sont occultés au profit d'une argumentation sur les valeurs de l'ouverture. C'est une tendance qu'observait déjà Samuel Goëta en 2016 au sein du mouvement de l'Open Data : *« En présupposant l'existence de données brutes dans les administrations et en prônant la circulation fluide des données non modifiées qui doivent être non seulement de bonne qualité, mais aussi intelligibles par les humains et les machines, les militants de l'Open Data ont rendu le travail d'obtention des données invisible »* (Goëta 2016, p.221).

Ouverture vers de nouvelles pistes de recherche

Des considérations qui précèdent émergent **trois pistes de recherche** : l'une ayant trait à la valeur des données scientifiques **(1)** ; l'autre aux compétences mobilisées pour ouvrir et réutiliser ces données **(2)** ; la troisième porte sur la définition d'un modèle de diffusion centré sur le chercheur **(3)**.

Le travail de collecte des données introduit tout d'abord un questionnement sur leur valeur de ces dernières. **Quel est le coût de la donnée (1) ?** Produire des données a un coût, mais les préserver et les diffuser en a un également. Autrement dit : est-il plus rentable de créer des infrastructures pour préserver et diffuser les données ou de reproduire les données quand on en a besoin ? Trois éléments entrent en ligne de compte, comme l'ont montré les modèles de partage des données en astronomie et en démographie :

- Le coût de production des données (plus il est élevé, plus la conservation et le partage des données sont considérés comme pertinents) ;

- La reproductibilité des données (les données non reproductibles sont d'autant plus précieuses qu'elles sont uniques) ;
- L'utilité des données (il n'y a d'intérêt à garder les données que si celles-ci sont susceptibles d'être réutilisées, bien que les usages soient difficiles à prédire).

La valeur est un point qui reste encore peu abordé, quand on parle d'ouverture des données de la recherche. Sous-peser le coût de l'ouverture et de la conservation des données à l'aune des trois éléments cités ici (travail, rareté, utilité) permettrait d'évaluer s'il est pertinent d'ouvrir les données de tous les chercheurs.

La réutilisation des données, en particulier, est une dimension qui mériterait d'être explorée plus avant. Dans le discours des promoteurs de l'ouverture, elle constitue l'un des principaux objectifs de la mise à disposition des données. Le premier objectif est la réutilisation des données par les chercheurs. Le second est leur combinaison avec des données d'autres secteurs – l'administration et l'industrie –, comme l'ambitionne la Commission européenne dans une communication de 2017, *Construire une économie européenne fondée sur les données*[311]. Effectivement, il semble exister un potentiel de réutilisation des données. Une enquête de Tenopir et al. (2011, p.17) montre par exemple que les chercheurs sont intéressés par les données de leurs pairs : 83 % des répondants disent qu'ils utiliseraient les données d'autres chercheurs si celles-ci leur étaient facilement accessibles. Néanmoins ces usages présumés ne permettent pas d'identifier quels seraient, en pratique, le nombre de réutilisations et leur nature. Il serait donc intéressant de mener un travail de recherche sur la réutilisation des données scientifiques, à l'instar d'Irene Pasquetto (2018), dont la thèse a porté sur l'utilisation secondaire des données biomédicales du consortium DataFace[312]. Ce travail consisterait par exemple à identifier les réutilisateurs potentiels des données, à comprendre leurs besoins et leur culture (vision, pratiques informationnelles...) et à envisager la façon dont ils pourraient s'insérer dans les projets d'ouverture des données. Car l'implication des publics cibles reste un des points faibles des politiques de données ouvertes. Ces dernières partent du postulat qu'il suffit de rendre les données accessibles pour que celles-ci trouvent

311 COMMISSION EUROPÉENNE (2017a). *Communication de la Commission au Conseil, au Parlement européen, au Comité économique et social et au Comité des régions : Créer une économie européenne fondée sur les données*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:9:FIN (consulté le 1er octobre 2019).

312 http://thedataface.com/

naturellement leur public. Si bien que l'enjeu de la réutilisation reste un horizon vague, relégué au second plan des projets d'ouverture. Il semble pourtant essentiel de prendre en compte le profil et les intentions des réutilisateurs lorsque l'on crée un dispositif de données ouvertes. Omettre ce paramètre, c'est compromettre l'adoption du dispositif par ses publics cibles. La prise en compte de ces derniers se traduit notamment dans la communication autour du dispositif (choix des publics auprès desquels communiquer) et dans le design du dispositif lui-même (l'ergonomie et la terminologie employée doivent être pensées en fonction d'un public prédéfini, qui a son propre domaine de connaissances et de compétences). C'est ce qu'analysaient Martin et al. (2013, p.10) dans le champ de l'Open Data, où la réutilisation des données est un sujet qui suscite d'ores et déjà des questionnements :

> *« La réutilisation dépend également des compétences des potentiels réutilisateurs. […] Les analyser peut permettre de comprendre comment faciliter la réutilisation et quels services développer à partir des données. »*[313]

L'ouverture des données et leur réutilisation interrogent également sur les compétences qu'elles mobilisent (2). Quelles sont ces compétences ? Lesquelles peuvent être endossées par les chercheurs ? Lesquelles doivent être déléguées à des personnels dédiés ? Les questions sont les mêmes quand on évoque les compétences liées à la réutilisation des données. Plus largement, ce sont des questions qui traversent l'ensemble des domaines concernés par les *data*. Arruabarrena et al. (2019) montrent que le traitement des données mobilise des compétences multiples – celles-ci sont informatiques, mathématiques, juridiques, éthiques, épistémologiques… Les sciences de l'information et de la communication rassemblent ces compétences sous le concept de *data literacy*, qui peut être défini comme la capacité des individus *« d'accéder, d'interpréter, d'évaluer de manière critique, de gérer, de manipuler et d'utiliser éthiquement les données »*[314] (Calzada-Prado et Marzal 2013). Ce qui nous intéresse, c'est d'explorer vers quel modèle d'organisation s'achemine l'ouverture des données : les tâches d'ouverture seront-elles confiées à des profils de poste dédiés (comme le montre la

313 Traduction de : « Reuse also depends on the skills of potential reusers. […] Analysing the skills of re-users can help understand how to facilitate the reuse and the type of services that can be developed on top of the datasets. » (Martin et al. 2013, p.10)

314 « Data literacy can be defined, then, as the component of information literacy that enables individuals to access, interpret, critically assess, manage, handle and ethically use data. » (Calzada-Prado et Marzal 2013, p.126)
Traduction de Arruabarrena et al. (2019)

création de la fonction d'ingénieur d'études chargé du traitement des données scientifiques[315]) ou bien seront-elles réparties entre différents corps de métiers (de l'informatique à la documentation, en passant par le droit) ? Dans ce second cas, comment structurer alors la coordination entre les différents acteurs des données ?

Comme le soulignait Marin Dacos, conseiller Science ouverte auprès du Directeur général de la recherche et de l'innovation, les enjeux ne sont pas seulement techniques et financiers. Ils sont aussi culturels.

> *« On peut décider tout ce que l'on veut au niveau ministériel, mais le risque est élevé qu'il ne produise pas de changement au cœur de la communauté scientifique. Si vous y ajoutez une mesure coercitive, vous augmenterez un peu vos chances de succès, mais vous aurez une courbe d'adoption forcée typique de ce qu'on appelle les innovations administratives. Si vous voulez atteindre 100 % d'accès ouvert et que tout le monde structure et partage ses données, il ne faut pas passer par la contrainte, mais par un changement culturel. »*[316]

« [S']appuyer sur la force des usages existants »[316] au sein des communautés scientifiques semble être une solution raisonnable et probablement celle qui sera la plus efficiente. L'interrogation est donc la suivante : **comment axer les services de données sur ce qui fait sens pour le chercheur (3) ?** Comment éviter l'écueil d'une offre de services décontextualisée de l'activité scientifique ? Les travaux de Roosendaal, Zalewska-Kurek et Geurts (2010) dans le domaine de l'édition scientifique sont intéressants sur ce point, car ils posent les bases d'un modèle économique de libre accès orienté vers la demande et centré sur la recherche. Roosendaal et al. considèrent en effet que l'édition scientifique doit avoir pour principal but le partage de l'information. Elle doit aider les chercheurs à rendre leurs résultats publics et à accéder aux productions de leurs pairs. Le chercheur et les motivations qui le poussent à partager l'information constituent donc le point charnière de l'analyse proposée par Roosendaal et al.. Les auteurs identifient trois motivations à publier (la reconnaissance, le partage de connaissances et les pressions extérieures) et analysent le marché de l'information

315 Emploi-type F2A43 du référentiel des emplois-types de la recherche et de l'enseignement supérieur III (REFERENS III) : https://data.enseignementsup-recherche.gouv.fr/pages/fiche_emploi_type_referens_iii_itrf/?refine.referens_id=F2A43#top

316 Propos tirés d'un article pour le magazine *Acteurs Publics* (Marzolf 2019).

scientifique en termes de forces et de fonctions. Dans cette perspective, une autre piste intéressante consisterait donc à appliquer le modèle d'analyse de Roosendaal et al. à l'étude des forces et fonctions d'un système de diffusion des données de la recherche. L'enjeu de ce travail serait ainsi de proposer un modèle qui soit le moins contraignant possible pour le chercheur et qui fasse sens avec ses pratiques de recherche et l'environnement dans lequel il évolue et se fait reconnaître.

Bibliographie

ARNOULD, P.-Y. ET JACQUEMOT, M.-C. (2016). *Guide de bonnes pratiques : Gestion et valorisation des données de la recherche*. OTELo et INIST-CNRS. https://hal.archives-ouvertes.fr/hal-01275841 (consulté le 15 septembre 2019).

ARROW, K. J. (1962). 'Economic Welfare and the Allocation of Resources for Invention'. In: NATIONAL BUREAU OF ECONOMIC RESEARCH (dir.). *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Princeton: Princeton University Press, pp. 609–26.

ARRUABARRENA, B., KEMBELLEC, G. ET CHARTRON, G. (2019). 'Data littératie & SHS : développer des compétences pour l'analyse des données'. In: H. BESTOUGEFF ET C. BOURRET (dir.). *Data Value Chain in Sciences and Territories. CODATA France Proceedings, 14 & 15 mars 2019*.

AWRE, C., BAXTER, J., CLIFFORD, B., COLCLOUGH, J., COX, A., DODS, N., DRUMMOND, P., FOX, Y., GILL, M., GREGORY, K., GURNEY, A., HARLAND, J., KHOKHAR, M., LOWE, D., O'BEIRNE, R., PROUDFOOT, R., SCHWAMM, H., SMITH, A., VERBAAN, E., WALLER, L., WILLIAMSON, L., WOLF, M. ET ZAWADZKI, M. (2015). 'Research Data Management as a "wicked problem"'. *Library Review*. 64 (4/5): 356–71. https://doi.org/10.1108/LR-04-2015-0043 (consulté le 19 septembre 2019).

BATES, M. .J. (2005). 'Information and knowledge: an evolutionary framework for information science'. *Information Research*. 10(4): 239. http://InformationR.net/ir/10-4/paper239.html (consulté le 14 octobre 2019).

BECARD, N., CASTETS-RENARD, C., CHASSANG, G., DANTANT, M., FREYT-CAFFIN, L., GANDON, N., MARTIN, C., MARTELLETTI, A., MENDOZA-CAMINADE, A., MORCRETTE, N., NEIRAC, C., JEAN, B. ET KASSEM, L. (2017). *Ouverture des données de recherche. Guide d'analyse du cadre juridique en France*. Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. https://www.ouvrirlascience.fr/wp-content/uploads/2018/11/Guide_Juridique_V2.pdf (consulté le 20 septembre 2019).

BENSAUDE VINCENT, B. (2014). 'The politics of buzzwords at the interface of technoscience, market and society: The case of "public engagement in science"'. *Public Understanding of Science*. 23 (3): 238–53. https://doi.org/10.1177/0963662513515371 (consulté le 19 septembre 2019).

BERNECKER, S. ET DRETSKE, F. (dir.) (2000). *Knowledge: Readings in contemporary epistemology*. Oxford: Oxford University Press.

BERTAUX, D. (1980). 'L'approche biographique : sa validité méthodologique, ses potentialités'. *Cahiers internationaux de sociologie*. LXIX: 197–225.

BORGMAN, C. L. (2012). 'The conundrum of sharing research data'. *Journal of the American Society for Information Science and Technology*. 63 (6): 1059–78.
https://doi.org/10.1002/asi.22634
(consulté le 19 septembre 2019).

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

BORGMAN, C. L., DARCH, P. T., SANDS, A. E. ET GOLSHAN, M. S. (2016). 'The durability and fragility of knowledge infrastructures: Lessons learned from astronomy'. *Proceedings of the Association for Information Science and Technology*. 53 (1): 1–10.
https://doi.org/10.1002/pra2.2016.14505301057
(consulté le 15 septembre 2019).

BOUCHEZ, J.-P. (2014). 'Autour de « l'économie du savoir » : ses composantes, ses dynamiques et ses enjeux'. *Savoirs*. 34 (1): 9–45.
https://www.cairn.info/revue-savoirs-2014-1-page-9.htm
(consulté le 15 septembre).

BOURDIEU, P. (1975). 'La spécificité du champ scientifique et les conditions sociales du progrès de la raison'. *Sociologie et sociétés*. 7 (1): 91–118.
http://www.erudit.org/fr/revues/socsoc/1975-v7-n1-socsoc122/001089ar/
(consulté le 15 septembre 2019).

BROUDOUX, E. (2018). *Dispositifs info-communicationnels numériques : éditorialisation et autorité*. Habilitation à diriger des recherches. Université Paris 8.

BRUNO, I. (2008). *À vos marques®, prêts... cherchez ! La stratégie européenne de Lisbonne, vers un marché de la recherche*. Bellecombe-en-Bauges: Éditions du Croquant.

BRUNO, I., GRAZULIS, S., HELLIWELL, J. R., KABEKKODU, S. N., MCMAHON, B. ET WESTBROOK, J. (2017). 'Crystallography and Databases'. *Data Science Journal*. 16(38): 1-17.
https://doi.org/10.5334/dsj-2017-038
(consulté le 8 octobre 2019).

BUCKLAND, M. K. (1991). 'Information as thing'. *Journal of the American Society for Information Science*. 42 (5): 351–60.
https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3
(consulté le 15 septembre 2019).

BYRNE, M. (2017). 'Making Progress Toward Open Data: Reflections on Data Sharing at PLOS ONE'. *PLOS Blogs*.
https://blogs.plos.org/everyone/2017/05/08/making-progress-toward-open-data/
(consulté le 17 septembre 2019).

CABRERA, F. (2014). *Les données de la recherche en sciences humaines et sociales : enjeux et pratiques. Enquête exploratoire*. Mémoire d'études de l'INTD. Paris: Conservatoire National des Arts et Métiers.
http://memsic.ccsd.cnrs.fr/mem_01128394/document
(consulté le 15 septembre 2019).

CACALY, S. (dir.) (1997). *Dictionnaire encyclopédique de l'information et de la documentation*. Paris: Nathan.

CALZADA PRADO, J. ET MARZAL, M.A. (2013). 'Incorporating data literacy into information literacy programs: core competencies and contents'. *Libri*. 63 (2): 123-134.
https://doi.org/10.1515/libri-2013-0010
(consulté le 1er octobre 2019)

CARTIER, A., MOYSAN, M. ET REYMONET, N. (2015). *Réaliser un plan de gestion de données : guide de rédaction*. Version 1.
https://www.fosteropenscience.eu/sites/default/files/pdf/2252.pdf
(consulté le 20 septembre 2019).

CHARTRON, G. (2010). 'Scénarios prospectifs pour l'édition scientifique'. *Hermes, La Revue*. 57 (2): 123–9.
https://www.cairn.info/revue-hermes-la-revue-2010-2-page-123.htm
(consulté le 15 septembre 2019).

CHARTRON, G. (2016). 'Stratégie, politique et reformulation de l'open access'. *Revue française des sciences de l'information et de la communication*. (8).
http://journals.openedition.org/rfsic/1836
(consulté le 15 septembre 2019).

CHARTRON, G. (2018). 'L'Open science au prisme de la Commission européenne'. *Education et sociétés*. N° 41 (1): 177–93.
https://doi.org/10.3917/es.041.0177
(consulté le 15 septembre 2019).

CHARTRON, G. ET SCHÖPFEL, J. (2017). 'Open access et Open science en débat'. *Revue française des sciences de l'information et de la communication*. (11). http://journals.openedition.org/rfsic/3331 (consulté le 15 septembre 2019).

CHIGNARD, S. (2012). *Open data, comprendre l'ouverture des données publiques*. Limoges: FYP Editions.

CHOWDHURY, G., BOUSTANY, J., KURBANOĞLU, S., ÜNAL, Y. ET WALTON, G. (2017). 'Preparedness for Research Data Sharing: A Study of University Researchers in Three European Countries'. In: S. CHOEMPRAYONG, F. CRESTANI, ET S. J. CUNNINGHAM (dir.). *Digital Libraries: Data, Information, and Knowledge for Digital Lives*. Springer, Cham, pp. 104–16. https://link.springer.com/chapter/10.1007/978-3-319-70232-2₉ (consulté le 15 septembre 2019).

COLEMAN, G. E. (2013). *Coding Freedom. The Ethics and Aesthetics of Hacking*. Princeton: Princeton University Press.

COLLECTIF P.É.C.R.E.S. (2011). *Recherche précarisée, recherche atomisée. Production et transmission des savoirs à l'heure de la précarisation*. Paris: Éditions Raisons d'Agir.

CORTI, L., VAN DEN EYNDEN, V., BISHOP, L. ET WOOLLARD, M. (2014). *Managing and Sharing Research Data: a Guide to Good Practice*. Londres : SAGE Publications Ltd.

COX, A. M. ET PINFIELD, S. (2014). 'Research data management and libraries: Current activities and future priorities'. *Journal of Librarianship and Information Science*. 46 (4): 299–316. https://doi.org/10.1177/0961000613492542 (consulté le 15 septembre 2019).

COX, A. M., PINFIELD, S. ET SMITH, J. (2016). 'Moving a brick building: UK libraries coping with research data management as a "wicked" problem'. *Journal of Librarianship and Information Science*. 48 (1): 3–17. https://doi.org/10.1177/0961000614533717 (consulté le 15 septembre 2019).

CUKIER, K. (2010). 'Data, data everywhere'. *The Economist*. https://www.economist.com/special-report/2010/02/27/data-data-everywhere (consulté le 15 septembre 2019).

DACOS, M. ET MOUNIER, P. (2015). *Humanités numériques : État des lieux et positionnement de la recherche française dans le contexte international*. Rapport de recherche. Institut français.
https://hal.archives-ouvertes.fr/hal-01228945
(consulté le 14 octobre 2019).

DEBOIN, M.-C. (2018). 'Découvrir de nouveaux métiers liés aux données de la recherche'. *CoopIST, Cirad*.
https://doi.org/10.18167/coopist/0061
(consulté le 20 septembre 2019).

DELAY-ARTOUS, C. (2017). 'Où sont les données de la recherche ? Essai de cartographie'. In: E. CHEVRY PÉBAYLE (dir.). *Systèmes d'organisation des connaissances et humanités numériques. Actes du 10ème colloque ISKO France 2015*. Londres: ISTE Editions, pp. 187–203.
https://halshs.archives-ouvertes.fr/halshs-01369745
(consulté le 16 septembre 2019).

DENIS, J. ET PONTILLE, D. (2012). 'Travailleurs de l'écrit, matières de l'information'. *Revue d'anthropologie des connaissances*. 6 (1): 1–20.
https://doi.org/10.3917/rac.015.0001
(consulté le 15 septembre 2019).

DILLAERTS, H. (2017). 'Ouverture et partage des résultats de la recherche dans l'économie de la connaissance européenne : quelle(s) liberté(s) de circulation pour l'IST ?'. *Communication & management*. (14): 39-54.
http://doi.org/10.3917/comma.141.0039
(consulté le 29 septembre 2019).

DUCHESNE, S. ET GARCIA, G. (2014). 'beQuali : une archive qualitative au service des sciences sociales'. In: M. CORNU, J. FROMAGEAU, ET B. MÜLLER (dir.). *Archives de la recherche. Problèmes et enjeux de la construction du savoir scientifique*. Paris: L'Harmattan, pp. 35–56.
https://halshs.archives-ouvertes.fr/halshs-00922690
(consulté le 16 septembre 2019).

EDWARDS, P. N. (2010). *A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge: The MIT Press.

EDWARDS, P. N., MAYERNIK, M. S., BATCHELLER, A. L., BOWKER, G. C. ET BORGMAN, C. L. (2011). 'Science friction: Data, metadata, and collaboration'. *Social Studies of Science*. 41 (5): 667–90.
https://doi.org/10.1177/0306312711413314
(consulté le 15 septembre 2019).

FABRE, I. ET GARDIÈS, C. (2008). 'L'accès à l'information scientifique numérique : organisation des savoirs et enjeux de pouvoir dans une communauté scientifique'. *Sciences de la Société*. 84–99.
https://hal.archives-ouvertes.fr/hal-00802763
(consulté le 15 septembre 2019).

FECHER, B., FRIESIKE, S., HEBING, M. ET LINEK, S. (2017). 'A reputation economy: How individual reward considerations trump systemic arguments for open access to data'. *Palgrave Communications*. 3: 17051.
https://www.nature.com/articles/palcomms201751
(consulté le 17 septembre 2019).

FECHER, B., FRIESIKE, S., HEBING, M., LINEK, S. ET SAUERMANN, A. (2015). 'A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing'. *DIW Berlin Discussion Paper No. 1454*.
http://dx.doi.org/10.2139/ssrn.2568693
(consulté le 15 septembre 2019).

FLORIDI, L. (2005). 'Semantic Conceptions of Information'. In: E. N. ZALTA (dir.). *The Stanford Encyclopedia of Philosophy*.
https://plato.stanford.edu/archives/sum2019/entries/information-semantic/
(consulté le 14 octobre 2019).

FORAY, D. ET LUNDVALL, B.-A. (1997). 'Une introduction à l'économie fondée sur la connaissance'. In: B. GUILHON, P. HUARD, M. ORILLARD, ET J.-B. ZIMMERMANN (dir.). *Économie de la connaissance et des organisations. Entreprises, territoires, réseaux*. Paris: L'Harmattan, pp. 16–37.

GAILLARD, R. (2014). *De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?* Mémoire de fin d'étude du diplôme de conservateur. Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques.
https://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche
(consulté le 15 septembre 2019).

GOËTA, S. (2016). *Instaurer des données, instaurer des publics : une enquête sociologique dans les coulisses de l'open data*. Thèse de doctorat. Télécom ParisTech. https://pastel.archives-ouvertes.fr/tel-01458098 (consulté le 17 septembre 2019).

GOËTA, S. (2018). '"Données recherchent publics" : les politiques d"open data à l"épreuve de la réutilisation'. In: G. GOURGUES ET A. MAZEAUD (dir.). *L'action publique saisie par ses 'publics" : Gouvernement et (dés)ordre politique*. Villeneuve d'Ascq: Presses universitaires du Septentrion, pp. 137–54. http://books.openedition.org/septentrion/37527 (consulté le 15 septembre 2019).

HAGSTROM, W. O. (1965). *The Scientific Community*. New York: Basic Books.

HANSSON, S. O. (2002). 'Les incertitudes de la société du savoir'. *Revue internationale des sciences sociales*. 171 (1): 43–51. https://www.cairn.info/revue-internationale-des-sciences-sociales-2002-1-page-43.htm (consulté le 15 septembre 2019).

HARNAD, S. (1994). 'Publicly Retrievable FTP Archives for Esoteric Science and Scholarship: A Subversive Proposal'. *Network Services Conference*, Londres, 28-30 Novembre 1994.

HEY, T., TANSLEY, S. ET TOLLE, K. (dir.) (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf (consulté le 19 septembre 2019).

IBEKWE-SANJUAN, F. ET PAQUIENSÉGUY, F. (2015). 'Open, Big, Collaboration : trois utopies de l'innovation au xxie siècle'. In: G. CHARTRON ET E. BROUDOUX (dir.). *Big Data - Open Data : Quelles valeurs ? Quels enjeux ? Actes du colloque 'Document numérique et société', Rabat, 2015*. Louvain-la-Neuve: De Boeck Supérieur, pp. 15–29. https://www.cairn.info/big-data-open-data-quelles-valeurs--9782807300316-page-15.htm (consulté le 16 septembre 2019).

IRSTEA (2017). *Guide pratique pour la gestion des données de recherche*. Version 2. https://donnees-recherche.irstea.fr/preambule-au-guide/ (consulté le 6 octobre 2019).

KAUFMANN, J.-C. (1996). *L'entretien compréhensif*. Paris: Nathan.

KINDLING, M., PAMPEL, H., SANDT, S., RÜCKNAGEL, J., VIERKANT, P., KLOSKA, G., WITT, M., SCHIRMBACHER, P., BERTELMANN, R. ET SCHOLZE, F. (2017). 'The Landscape of Research Data Repositories in 2015: A re3data Analysis'. *D-Lib Magazine*. 23 (3/4). https://doi.org/10.1045/march2017-kindling (consulté le 15 septembre 2019).

KNORR-CETINA, K. (1981). *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Oxford: Pergamon.

KNORR-CETINA, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge: Harvard University Press.

KOLTAY, T. (2016). 'Are you ready? Tasks and roles for academic libraries in supporting Research 2.0'. *New Library World*. 117 (1/2): 94–104. https://doi.org/10.1108/NLW-09-2015-0062 (consulté le 15 septembre 2019).

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M. ET FITZHUGH, W. ET AL. (2001). 'Initial sequencing and analysis of the human genome'. *Nature*. 409(6822): 860–921. https://doi.org/10.1038/35057062 (consulté le 8 octobre 2019).

LATOUR, B. (2001). *Le métier de chercheur : regard d'un anthropologue. Une conférence-débat à l'INRA*. Paris: INRA Editions.

LATOUR, B. ET WOOLGAR, S. (1979). *La vie de laboratoire. La production des faits scientifiques*. Paris: La Découverte.

LATRIVE, F. (2000). 'Les Barbares du Bazar : Une introduction aux faubourgs de la nouvelle économie'. In: O. BLONDEAU (dir.). *Libres enfants du savoir numérique. Une anthologie du 'Libre'*. Paris: Editions de l'Éclat, pp. 11–8. https://doi.org/10.3917/ecla.blond.2000.01.0011 (consulté le 16 septembre 2019).

LEJEUNE, C. (2014). *Manuel d'analyse qualitative. Analyser sans compter ni classer*. Louvain-la-Neuve: De Boeck.

LEONELLI, S. (2013a). 'Integrating data to acquire new knowledge: Three modes of integration in plant science'. *Studies in History and Philosophy of Biological and Biomedical Sciences*. 44 (4): 503–14. https://doi.org/10.1016/j.shpsc.2013.03.020 (consulté le 29 septembre 2019).

LEONELLI, S. (2013b). 'Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology'. *Bulletin of Science, Technology and Society*. 33 (1-2): 6–11.
https://doi.org/10.1177/0270467613496768
(consulté le 29 septembre 2019).

LEONELLI, S. (2015). 'What Counts as Scientific Data? A Relational Framework'. *Philosophy of Science*. 82 (5): 810–21.
https://www.jstor.org/stable/10.1086/684083
(consulté le 15 septembre 2019).

LIQUÈTE, V. (2010). 'Présentation générale : Formes et enjeux de la médiation'. In: V. LIQUÈTE (dir.). *Médiations*. Paris : CNRS Éditions, pp. 9-31.
http://doi.org/10.4000/books.editionscnrs.14712
(consulté le 29 septembre 2019).

MACHLUP, F. (1962). *The Production and Distribution of Knowledge in the United States*. Princeton: Princeton University Press.

MALINGRE, M.-L., MIGNON, M., PIERRE, C. ET SERRES, A. (2019). 'Construction(s) et contradictions des données de recherche en SHS'. *Recherche d'information, document et web sémantique*. 19-2 (1).
https://doi.org/10.21494/ISTE.OP.2019.0336
(consulté le 16 septembre 2019).

MARCIAL, L. H. ET HEMMINGER, B. M. (2010). 'Scientific data repositories on the Web: An initial survey'. *Journal of the American Society for Information Science and Technology*. 61 (10): 2029–48.
https://doi.org/10.1002/asi.21339
(consulté le 19 septembre 2019).

MARTIN, S., FOULONNEAU, M., TURKI, S. ET IHADJADENE, M. (2013). 'Risk Analysis to Overcome Barriers to Open Data'. *Electronic Journal of e-Government*. 11(1): 348-359.

MARZOLF, E. (2019). 'Marin Dacos : "Pour ouvrir la science, il ne faut pas contraindre, mais s'appuyer sur la force des usages existants"'. *Acteurs Publics*.
https://www.acteurspublics.fr/articles/marin-dacos-pour-ouvrir-la-science-il-ne-faut-pas-contraindre-mais-sappuyer-sur-la-force-des-usages-existants
(consulté le 17 septembre 2019).

MAUREL, L. (2015). 'Le statut juridique des données de la recherche : entre droit des bases de données et données publiques'. *Blog S.I.Lex*. https://scinfolex.com/2015/07/13/le-statut-juridique-des-donnees-de-la-recherche-entre-droit-des-bases-de-donnees-et-donnees-publiques/ (consulté le 11 octobre 2019).

MAUREL, L. (2018a). 'Données personnelles et recherche scientifique : quelle articulation dans le RGPD ?'. *Blog S.I.Lex*. https://scinfolex.com/2018/07/18/donnees-personnelles-et-recherche-scientifique-quelle-articulation-dans-le-rgpd/ (consulté le 16 septembre 2019).

MAUREL, L. (2018b). 'La réutilisation des données de la recherche après la loi pour une République numérique'. In: V. GINOUVÈS ET I. GRAS (dir.). *La diffusion numérique des données en SHS. Guide de bonnes pratiques éthiques et juridiques*. Aix-en-Provence: Presses Universitaires de Provence. https://hal.archives-ouvertes.fr/hal-01908766 (consulté le 19 septembre 2019).

MAUSS, M. (2007). *Essai sur le don : forme et raison de l'échange dans les sociétés archaïques*. Paris: PUF.

MAYÈRE, A. (1990). *Pour une économie de l'information*. Paris: C.N.R.S. Editions. http://doi.org/10.3917/cnrs.mayer.1990.01 (consulté le 14 octobre 2019).

MERTON, R. K. (1973). *The Sociology of Science. Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

NICHOLAS, D., RODRÍGUEZ-BRAVO, B., WATKINSON, A., BOUKACEM-ZEGHMOURI, C., HERMAN, E., XU, J., ABRIZAH, A. ET ŚWIGOŃ, M. (2017). 'Early career researchers and their publishing and authorship practices'. *Learned Publishing*. 30(3) : 205-217. https://doi.org/10.1002/leap.1102 (consulté le 14 octobre 2019).

O'REILLY, T. (2010). 'Government As a Platform'. In: L. RUMA ET D. LATHROP (dir.). *Open Government*. Sebastopol: O'Reilly Media. https://www.oreilly.com/library/view/open-government/9781449381936/ch02.html (consulté le 16 septembre 2019).

PASQUETTO, I. (2018). *From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses*. PhD Thesis. University of California. https://escholarship.org/uc/item/1s1814cj (consulté le 1er octobre 2019).

PEUGEOT, V. (2012). 'L'ouverture des données publiques : convergence ou malentendu politique ?' In: B. STIEGLER (dir.). *Confiance, croyance, crédit dans les mondes industriels*. Limoges: FYP Editions.

PINSON, G. ET SALA PALA, V. (2007). 'Peut-on vraiment se passer de l'entretien en sociologie de l'action publique ?'. *Revue francaise de science politique*. 57 (5): 555–97. https://doi.org/10.3917/rfsp.575.0555 (consulté le 15 septembre 2019).

REBOUILLAT, V. (2017). 'Inventory of Research Data Management Services in France'. In: L. CHAN ET F. LOIZIDES (dir.). *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices*. Amsterdam : IOS Press, pp.174-181. http://ebooks.iospress.nl/publication/46651 (consulté le 19 septembre 2019)

REBOUILLAT, V. (à paraître). 'Les données scientifiques face aux enjeux de la recherche en Sciences, Technologie et Médecine : enquête exploratoire à l'Université de Strasbourg'. *Etudes de communication : Langages, information, médiations*. 52. https://hal-cnam.archives-ouvertes.fr/hal-02321077/ (consulté le 14 octobre 2019).

REBOUILLAT, V. ET CHARTRON, G. (2019). 'Services de gestion et de partage des données de recherche : ce qu'en pensent les chercheurs'. *12ème colloque international d'ISKO-France*, Montpellier, 9-11 octobre 2019. https://hal-cnam.archives-ouvertes.fr/hal-02307085v1 (consulté le 14 octobre 2019).

REYMONET, N., MOYSAN, M., CARTIER, A. ET DÉLÉMONTEZ, R. (2018). *Réaliser un plan de gestion de données « FAIR » : modèle*. Version 2. https://archivesic.ccsd.cnrs.fr/sic_01690547 (consulté le 20 septembre 2019).

RITTEL, H. W. J. ET WEBBER, M. M. (1973). 'Dilemmas in a general theory of planning'. *Policy Sciences*. 4 (2): 155–69. https://doi.org/10.1007/BF01405730 (consulté le 16 septembre 2019).

ROOSENDAAL, H. E., ZALEWSKA-KUREK, K. ET GEURTS, P. (2010). *Scientific publishing: from vanity to strategy*. Oxford: Chandos Publishing.

ROSENBERG, D. (2013). 'Data before the Fact'. In: L. GITELMAN (dir.). *'Raw Data' Is an Oxymoron*. The MIT Press, pp. 15–40.
https://doi.org/10.7551/mitpress/9302.003.0003
(consulté le 16 septembre 2019).

SCHÖPFEL, J. (2012). 'Vers une nouvelle définition de la littérature grise'. *Cahiers de la Documentation*. 66 (3): 14-24.
https://hal.archives-ouvertes.fr/sic_00794984
(consulté le 5 octobre 2019).

SCHÖPFEL, J. (2018a). 'Hors norme ? Une approche normative des données de la recherche'. *Revue Communication, Organisation, Société du Savoir et Information*. (5).
https://revue-cossi.info/numeros/n-5-2018-processus-normalisation-durabilite-information/730-5-2018-schopfel
(consulté le 15 septembre 2019).

SCHÖPFEL, J. (2018b). *Vers une culture de la donnée en SHS. Une étude à l'Université de Lille*. Université de Lille.
https://hal.archives-ouvertes.fr/hal-01846849
(consulté le 15 septembre 2019).

SCHÖPFEL, J., KERGOSIEN, E. ET PROST, H. (2017a). 'Pour commencer, pourriez-vous définir 'données de la recherche' ? Une tentative de réponse'. *Colloque INFORSID 2017*, Toulouse, France.
https://hal.univ-lille3.fr/hal-01530937
(consulté le 15 septembre 2019).

SCHÖPFEL, J., PROST, H. ET REBOUILLAT, V. (2017b). 'Research Data in Current Research Information Systems'. *Procedia Computer Science*. 106 (2017): 305-320.
https://doi.org/10.1016/j.procs.2017.03.030
(consulté le 20 septembre 2019).

SCIENCE|BUSINESS CLOUD CONSULTATION GROUP (2018). *Priorities for the European Open Science Cloud. White paper*. Science|Business.
https://sciencebusiness.net/report/priorities-european-open-science-cloud
(consulté le 16 septembre 2019).

SERRES, A., MALINGRE, M. L., MIGNON, M., PIERRE, C. ET COLLET, D. (2017). *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2*. Rapport de recherche. Université Rennes 2.
https://hal.archives-ouvertes.fr/hal-01635186v2
(consulté le 8 octobre 2019).

SERRES, A., VIGNALE, F. (2016). 'Les données de la recherche'. *Formadoct*. Rennes : Université Européenne de Bretagne.
http://guides-formadoct.u-bretagneloire.fr/donnees_recherche
(consulté le 20 septembre 2019).

SI, L., XING, W., ZHUANG, X., HUA, X. ET ZHOU, L. (2015). 'Investigation and analysis of research data services in university libraries'. *The Electronic Library*. 33 (3): 417–49.
https://doi.org/10.1108/EL-07-2013-0130
(consulté le 16 septembre 2019).

SPARC EUROPE (2018). *An Analysis of Open Data and Open Science Policies in Europe*. Version 3.
https://sparceurope.org/latest-update-to-european-open-data-and-open-science-policies-released/
(consulté le 19 septembre 2019).

STAR, S. L. ET GRIESEMER, J. R. (1989). 'Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39'. *Social Studies of Science*. 19 (3): 387–420.
http://www.jstor.org/stable/285080
(consulté le 16 septembre 2019).

STUART, D., BAYNES, G., HRYNASZKIEWICZ, I., ALLIN, K., PENNY, D., LUCRAFT, M. ET ASTELL, M. (2018). *Whitepaper: Practical challenges for researchers in data sharing*.
https://figshare.com/articles/Whitepaper_Practical_challenges_for_researchers_in_data_sharing/5975011
(consulté le 16 septembre 2019).

TENOPIR, C., ALLARD, S., DOUGLASS, K., AYDINOGLU, A. U., WU, L., READ, E., MANOFF, M. ET FRAME, M. (2011). 'Data Sharing by Scientists: Practices and Perceptions'. *PLOS ONE*. 6 (6): e21101.
http://doi.org/10.1371/journal.pone.0021101
(consulté le 15 septembre 2019).

TENOPIR, C., BIRCH, B. ET ALLARD, S. (2012). *Academic libraries and research data services: Current practices and plans for the future. An ACRL White Paper*. Chicago: Association of College and Research Libraries.
http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf
(consulté le 15 septembre 2019).

TENOPIR, C., DALTON, E. D., ALLARD, S., FRAME, M., PJESIVAC, I., BIRCH, B., POLLOCK, D. ET DORSETT, K. (2015a). 'Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide'. *PLOS ONE*. 10 (8).
http://dx.doi.org/10.1371/journal.pone.0134826
(consulté le 15 septembre 2019).

TENOPIR, C., HUGHES, D., ALLARD, S., FRAME, M., BIRCH, B., BAIRD, L., SANDUSKY, R., LANGSETH, M. ET LUNDEEN, A. (2015b). 'Research Data Services in Academic Libraries: Data Intensive Roles for the Future?' *Journal of eScience Librarianship*. 4 (2).
http://dx.doi.org/10.7191/jeslib.2015.1085
(consulté le 15 septembre 2019).

TENOPIR, C., TALJA, S., HORSTMANN, W., LATE, E., HUGHES, D., POLLOCK, D., SCHMIDT, B., BAIRD, L., SANDUSKY, R. J. ET ALLARD, S. (2017). 'Research Data Services in European Academic Research Libraries'. *LIBER Quarterly*. 27 (1): 23–44.
http://doi.org/10.18352/lq.10180
(consulté le 15 septembre 2019).

VAN DEN EYNDEN, V., KNIGHT, G., VLAD, A., RADLER, B., TENOPIR, C., LEON, D., MANISTA, F., WHITWORTH, J. ET CORTI, L. (2016). *Towards Open Research: practices, experiences, barriers and opportunities*. Wellcome Trust.
https://dx.doi.org/10.6084/m9.figshare.4055448
(consulté le 16 septembre 2019).

WALLIS, J. C., ROLANDO, E. ET BORGMAN, C. L. (2013).'If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology'. *PLOS ONE*. 8 (7).
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332
(consulté le 15 septembre 2019).

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, IJ. J., APPLETON, G., AXTON, M., BAAK, A., BLOMBERG, N., BOITEN, J.-W., SANTOS, L. B. DA S., BOURNE, P. E., BOUWMAN, J., BROOKES, A. J., CLARK, T., CROSAS, M., DILLO, I., DUMON, O., EDMUNDS, S., EVELO, C. T., FINKERS, R., GONZALEZ-BELTRAN, A., GRAY, A. J. G., GROTH, P., GOBLE, C., GRETHE, J. S., HERINGA, J., HOEN, P. A. C. 'T, HOOFT, R., KUHN, T., KOK, R., KOK, J., LUSHER, S. J., MARTONE, M. E., MONS, A., PACKER, A. L., PERSSON, B., ROCCA-SERRA, P., ROOS, M., SCHAIK, R. VAN, SANSONE, S.-A., SCHULTES, E., SENGSTAG, T., SLATER, T., STRAWN, G., SWERTZ, M. A., THOMPSON, M., LEI, J. VAN DER, MULLIGEN, E. VAN, VELTEROP, J., WAAGMEESTER, A., WITTENBURG, P., WOLSTENCROFT, K., ZHAO, J. ET MONS, B. (2016). 'The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific Data*. 3 (160018).
https://www.nature.com/articles/sdata201618
(consulté le 16 septembre 2019).

YU, H. ET ROBINSON, D. G. (2012). 'The New Ambiguity of "Open Government"'. *UCLA Law Review*. 59: Discourse 158.
http://dx.doi.org/10.2139/ssrn.2012489
(consulté le 16 septembre 2019).

ZINS, C. (2007). 'Conceptual approaches for defining data, information, and knowledge'. *Journal of the American Society for Information Science and Technology*. 58 (4): 479–93.
https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20508
(consulté le 15 septembre 2019).

Sources

AGENCE NATIONALE DE LA RECHERCHE (2018). *Plan d'action 2019*. https://anr.fr/fileadmin/documents/2018/Plan-d-action-ANR-2019.pdf (consulté le 19 septembre 2019).

AGENCE NATIONALE DE LA RECHERCHE (2019). *Appel FLASH Science ouverte : pratiques de recherche et données ouvertes*. http://www.agence-nationale-recherche.fr/fileadmin/aap/2019/aap-data-2019.pdf (consulté le 19 septembre 2019).

Amsterdam Call for Action on Open Science (2016). https://www.ouvrirlascience.fr/amsterdam-call-for-action-on-open-science/ (consulté le 19 septembre 2019).

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003). https://openaccess.mpg.de/Berlin-Declaration (consulté le 16 septembre 2019).

Budapest Open Access Initiative (2002). https://www.budapestopenaccessinitiative.org/read (consulté le 16 septembre 2019).

COMMISSION EUROPÉENNE (1993). *Croissance, compétitivité, emploi : les défis et les pistes pour entrer dans le XXIe siècle. Livre blanc*. https://publications.europa.eu/fr/publication-detail/-/publication/0d563bc1-f17e-48ab-bb2a-9dd9a31d5004 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2000). *Communication de la Commission au Conseil, au Parlement européen, au Comité économique et social et au Comité des régions : Vers un espace européen de la recherche*. https://eur-lex.europa.eu/legal-content/fr/TXT/?uri=CELEX:52000DC0006 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2007). *Communication de la Commission au Parlement européen, au Conseil et Comité économique et social européen sur l'information scientifique à l'ère numérique : accès, diffusion et préservation*. https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52007DC0056&from=EN (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2010a). *Communication de la Commission. Europe 2020 : Une stratégie pour une croissance intelligente, durable et inclusive*. https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A52010DC2020 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2010b). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions. Une stratégie numérique pour l'Europe*. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52010DC0245R(01 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2010c). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions. Une Union de l'innovation*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2010:0546:FIN (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2011). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions. L'ouverture des données publiques: un moteur pour l'innovation, la croissance et une gouvernance transparente.* https://eur-lex.europa.eu/legal-content/fr/TXT/?uri=CELEX:52011DC0882 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2012a). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au comité des régions. Pour un meilleur accès aux informations scientifiques: dynamiser les avantages des investissements publics dans le domaine de la recherche*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:52012DC0401 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2012b). *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au comité des régions. Un partenariat renforcé pour l'excellence et la croissance dans l'Espace européen de la recherche*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:52012DC0392 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2012c). *Recommandation de la Commission relative à l'accès aux informations scientifiques et à leur conservation*. https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=uriserv:OJ.L_.2012.194.01.0039.01.FRA&toc=OJ:L:2012:194:TOC (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2015). *Communication de la Commission au Conseil, au Parlement européen, au Comité économique et social et au Comité des régions : Stratégie pour un marché unique numérique en Europe*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52015DC0192 (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2016a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: European Cloud Initiative - Building a competitive data and knowledge economy in Europe*. https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2016b). *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020*. Version 3.0. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2017a). *Communication de la Commission au Conseil, au Parlement européen, au Comité économique et social et au Comité des régions : Créer une économie européenne fondée sur les données*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:9:FIN (consulté le 1er octobre 2019).

COMMISSION EUROPÉENNE (2017b). *H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Version 3.2. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2017c). *H2020 Programme: Mono-Beneficiary General Model Grant Agreement*. Version 5.0.
http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-mono_en.pdf
(consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2018a). *Budget de l'Union : La Commission propose le programme de recherche et d'innovation le plus ambitieux à ce jour*. Communiqué de presse, 7 juin 2018, Bruxelles.
https://europa.eu/rapid/press-release_IP-18-4041_fr.htm
(consulté le 18 septembre 2019).

COMMISSION EUROPÉENNE (2018b). *Commission Staff Working Document. Impact Assessment of Horizon Europe*. SWD(2018) 307 final, partie 2/3.
https://ec.europa.eu/info/publications/horizon-europe-impact-assessment-staff-working-document_en
(consulté le 19 septembre 2019).

COMMISSION EUROPÉENNE (2018c). *Recommandation de la Commission du 25.4.2018 relative à l'accès aux informations scientifiques et à leur conservation*.
http://data.europa.eu/eli/reco/2018/790/oj
(consulté le 8 octobre 2019).

COMMISSION EUROPÉENNE (2019). *H2020 Programme: Annotated Model Grant Agreement.* Version 5.2.
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf
(consulté le 18 septembre 2019).

CONSEIL EUROPÉEN (2000). *Conclusions de la présidence*. Conseil du 23 et 24 mars 2000 à Lisbonne.
http://www.europarl.europa.eu/summits/lis1_fr.htm
(consulté le 19 septembre 2019).

Déclaration du gouvernement ouvert (2011).
https://www.opengovpartnership.org/fr/process/joining-ogp/open-government-declaration/
(consulté le 19 septembre 2019).

Décret n°2005-1309 du 20 octobre 2005 pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.
https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=20034927FA69108CF64868188CDA7239.tplgfr31s_3?cidTexte=JORFTEXT000000241445&dateTexte=20180812
(consulté le 20 septembre 2019).

DIRECTION GÉNÉRALE POUR LA RECHERCHE ET L'INNOVATION DE LA COMMISSION EUROPÉENNE (2019). *European Open Science Cloud (EOSC) Strategic Implementation Plan*.
https://publications.europa.eu/s/m1qV
(consulté le 19 septembre 2019).

Directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public. 345. (32003L0098).
http://data.europa.eu/eli/dir/2003/98/oj/eng
(consulté le 19 septembre 2019).

Directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013 modifiant la directive 2003/98/CE concernant la réutilisation des informations du secteur public. 175.
http://data.europa.eu/eli/dir/2013/37/oj
(consulté le 19 septembre 2019).

Directive (EU) 2019/1024 du Parlement européen et du Conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public. 172.
http://data.europa.eu/eli/dir/2019/1024/oj
(consulté le 19 septembre 2019).

EOSC Declaration (2017).
https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf
(consulté le 19 septembre 2019).

EUDAT, LIBER, OPENAIRE, EGI, ET GEANT (2015). *Position Paper: European Open Science Cloud for Research*.
https://doi.org/10.5281/zenodo.32915
(consulté le 19 septembre 2019).

EUROPEAN RESEARCH COUNCIL (2007). *European Research Council-Scientific Council Guidelines for Open Access*.
https://erc.europa.eu/sites/default/files/document/file/erc_scc_guidelines_open_access.pdf
(consulté le 19 septembre 2019).

EUROPEAN STRATEGY FORUM ON RESEARCH INFRASTRUCTURES (2018). *Roadmap 2018. Strategy Report on Research Infrastructures*.
http://roadmap2018.esfri.eu/
(consulté le 10 septembre 2019).

HIGH LEVEL-EXPERT GROUP ON SCIENTIFIC DATA (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data*. Rapport final.
https://ec.europa.eu/digital-single-market/en/news/digital-agenda-unlock-full-value-scientific-data-high-level-group-presents-report
(consulté le 19 septembre 2019).

Le nuage européen pour la science ouverte devient une réalité (2018). Actualités du site de la Commission européenne.
https://ec.europa.eu/commission/news/european-open-science-cloud-becomes-reality-2018-nov-23_fr
(consulté le 19 septembre 2019).

Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000886460
(consulté le 20 septembre 2019).

Loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal (dite « loi CADA »).
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000339241
(consulté le 19 septembre 2019).

Loi n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public (dite « loi Valter »).
https://www.legifrance.gouv.fr/eli/loi/2015/12/28/PRMX1515110L/jo/texte
(consulté le 19 septembre 2019).

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.
https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte
(consulté le 19 septembre 2019).

MAX PLANCK GESELLSCHAFT (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*.
http://openaccess.mpg.de/Berlin-Declaration
(consulté le 18 septembre 2019).

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION (2016). *Stratégie nationale des infrastructures de recherche*. Edition 2016. http://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/74/5/feuille_route_infrastructures_recherche_2016_555745.pdf (consulté le 19 septembre 2019).

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION (2018). *Plan national pour la science ouverte*. http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf (consulté le 18 septembre 2019).

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION (2018). *Stratégie nationale des infrastructures de recherche*. Edition 2018, n°2. http://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/70/3/Brochure_Infrastructures_2018_948703.pdf (consulté le 19 septembre 2019).

NATIONAL SCIENCE BOARD (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Foundation. https://www.nsf.gov/pubs/2005/nsb0540/ (consulté le 15 septembre 2019).

OBAMA, B. (2009). *Transparency and Open Government. Memorandum for the Heads of Executive Departments and Agencies*. https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government (consulté le 19 septembre 2019).

OFFICE OF MANAGEMENT AND BUDGET (1993). *Circulaire A-110*, amendée le 30 septembre 1999. https://www.govinfo.gov/app/details/CFR-2012-title2-vol1/CFR-2012-title2-vol1-part215 (consulté le 19 septembre 2019).

Open Government Data Principles (2007). https://public.resource.org/8_principles.html (consulté le 19 septembre 2019).

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (1975). *La mesure des activités scientifiques en techniques : méthode-type proposée pour les enquêtes sur la recherche et le développement expérimental. Manuel de Frascati*. https://hal.archives-ouvertes.fr/hal-01511852 (consulté le 20 septembre 2019).

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (1996). *L'économie fondée sur le savoir*. OCDE/GD(96)102. Paris : Editions OCDE.
http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=OCDE/GD(96)102&docLanguage=Fr
(consulté le 18 septembre 2019).

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2004). *Déclaration sur l'accès aux données de la recherche financée par des fonds publics*.
https://legalinstruments.oecd.org/fr/instruments/157
(consulté le 18 septembre 2019).

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2007). *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics.* Paris : Éditions OCDE.
http://www.oecd.org/fr/science/sci-tech/38500823.pdf
(consulté le 19 septembre 2019).

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2010). *La stratégie de l'OCDE pour l'innovation : pour prendre une longueur d'avance*. Paris : Éditions OCDE.
https://dx.doi.org/10.1787/9789264084759-fr
(consulté le 19 septembre 2019).

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. 119.
https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679
(consulté le 20 septembre 2019).

RESEARCH COUNCILS UK (2016). *Concordat on Open Research Data*.
https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/
(consulté le 18 septembre 2019).

SERVICES DU PREMIER MINISTRE (2015). *Objectifs et indicateurs de performance du programme Coordination du travail gouvernemental*. Annexe au projet de loi de règlement et RAP 2015 – Mission Direction de l'action du Gouvernement.
https://www.performance-publique.budget.gouv.fr/sites/performance_publique/files/farandole/ressources/2015/rap/html/DRGPGMOBJINDPGM129.htm
(consulté le 14 octobre 2019).

SKORDAS, T. (2018). 'European Open Science Cloud Council Conclusions'. *Digital Single Market Blog Posts*.
https://ec.europa.eu/digital-single-market/en/blogposts/european-open-science-cloud-council-conclusions
(consulté le 19 septembre 2019).

VALLS, M. (2015). *Présentation de la stratégie numérique du Gouvernement*.
https://www.gouvernement.fr/partage/4972-discours-de-manuel-valls-lors-de-la-presentation-de-la-strategie-numerique-du-gouvernement-a-la
(consulté le 19 septembre 2019).

VIDAL, F. (2018). 'Plan national pour la science ouverte'. Discours prononcé lors du *Congrès LIBER*, 4 juillet 2018, Villeneuve d'Ascq.
http://www.enseignementsup-recherche.gouv.fr/cid132531/plan-national-pour-la-science-ouverte-discours-de-frederique-vidal.html
(consulté le 18 septembre 2019).

Sitographie & Acronymes

Action, discours, pensée politique et économique (UMR 5206 Triangle)
http://triangle.ens-lyon.fr/

ADISP, Archives de Données Issues de la Statistique Publique
http://www.progedo-adisp.fr/

ANR, Agence Nationale de la Recherche
https://anr.fr/

ArchiPolis
https://archipolis.hypotheses.org/

ArchiPolis Catalogue
https://catalogues.cdsp.sciences-po.fr/dataverse/archipolis

Architecture et Fonction des Macromolécules Biologiques (UMR 7357 AFMB)
http://www.afmb.univ-mrs.fr/

ArkeoGIS
http://arkeogis.org/

Analyse et Traitement Informatique de la Langue Française (UMR 7118 ATILF)
http://www.atilf.fr/

Bases de Données sur la Biodiversité, l'Écologie, l'Environnement et les Sociétés (UMS 3468 BBEES)
https://bbees.mnhn.fr/

BBEES : Annuaire, devenu le portail InDoRES (Inventaire des Données de Recherche en Environnement et Sociétés)
http://www.indores.fr/

BeQuali, Banque d'enquêtes qualitatives en sciences humaines et sociales
https://bequali.fr/

BRGM, Bureau de Recherches Géologiques et Minières
https://www.brgm.fr/

EELS Data Base
https://eelsdb.eu/

Euro-BioImaging
http://www.eurobioimaging.eu/

CASD, Centre d'Accès Sécurisé aux Données
https://www.casd.eu/

Cat OPIDoR, Catalogue pour une Optimisation du Partage et de l'Interopérabilité des Données de Recherche
https://cat.opidor.fr/

CDPP, Centre de Données de la Physique des Plasmas
http://www.cdpp.eu/

CDS Portal
https://cdsportal.u-strasbg.fr/

CEA, Commissariat à l'énergie atomique et aux énergies alternatives
http://www.cea.fr/

Centre d'Études Européennes (UMR 8239 CEE)
http://www.sciencespo.fr/centre-etudes-europeennes/

Centre d'Études et de Recherches Administratives, Politiques et Sociales (UMR 8026 CERAPS)
http://ceraps.univ-lille2.fr/

CDS, Centre de Données astronomiques de Strasbourg
https://cds.u-strasbg.fr/

Centre de Données Socio-Politiques (UMS 828 CDSP)
https://cdsp.sciences-po.fr/

Centre de Recherches Politiques (UMR 7048 CEVIPOF)
http://www.sciencespo.fr/cevipof/

Centre de Sociologie des Organisations (UMR 7116 CSO)
http://www.cso.edu/

CEREMA, Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement
https://www.cerema.fr/

Céreq, Centre d'Etudes et de Recherches sur les Qualifications
https://www.cereq.fr/

CERN, Centre Européen de la Recherche Nucléaire
https://home.cern/

CINES, Centre Informatique National de l'Enseignement Supérieur
https://www.cines.fr/

CINES : Plateforme d'archivage
https://www.cines.fr/archivage/

Cirad, Centre de Coopération Internationale en Recherche Agronomique et pour le Développement
https://www.cirad.fr/

Cirad : Annuaire de données
https://datacatalog.cirad.fr

Cirad : Information, aussi appelé site CoopIST
https://coop-ist.cirad.fr/gerer-des-donnees

CNIL, Commission nationale de l'informatique et des libertés
https://www.cnil.fr/

CNRS, Centre National de la Recherche Scientifique
http://www.cnrs.fr/

CoCoON, Collections de Corpus Oraux Numériques
https://cocoon.huma-num.fr/

CORIOLIS
http://www.coriolis.eu.org

COUPERIN, Consortium Unifié des Établissements Universitaires et de Recherche pour l'Accès aux Publications Numériques
https://www.couperin.org

Data BRGM
https://data.brgm.fr/

Data Inra
https://data.inra.fr/

DataPartage
https://www6.inra.fr/datapartage

DataSPIRE
https://catalogues.cdsp.sciences-po.fr/dataverse/dataspire

DARIAH, Digital Research Infrastructure for the Arts and Humanities
https://www.dariah.eu/

DIME-SHS, Données, Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales
https://dime-shs.sciencespo.fr/

DINAMIS, Dispositif Institutionnel National d'Approvisionnement Mutualisé en Imagerie Satellitaire
https://dinamis.teledetection.fr/

DMPonline
https://dmponline.dcc.ac.uk/

DMP OPIDoR
https://dmp.opidor.fr/

DoRANum, Données de la Recherche : Apprentissage Numérique à la gestion et au partage
https://doranum.fr/

Dryad
https://datadryad.org/

Dynamique Musculaire et Métabolisme (UMR 866 DMEM)
https://www6.montpellier.inra.fr/dmem

ECOSCOPE, Pôle de données d'observation pour la recherche sur la biodiversité
http://ecoscope.fondationbiodiversite.fr/ecoscope-portal/

ECOSCOPE Metadata Portal of Biodiversity Research Observatories
http://ecoscope.fondationbiodiversite.fr/ecoscope-portal/

EGI, European Grid Infrastructure
https://www.egi.eu/

Elfe, Étude longitudinale française depuis l'enfance
https://www.elfe-france.fr/

ELIPSS, Etude Longitudinale par Internet pour les Sciences Sociales
https://www.elipss.fr/

ELIXIR
https://elixir-europe.org/

EMBL-EBI, European Bioinformatics Institute
https://www.ebi.ac.uk/

Enssib, École Nationale Supérieure des Sciences de l'Information et des Bibliothèques
https://www.enssib.fr/

EOSC, European Open Science Cloud
https://www.eosc-portal.eu/

EPRIST, Association des responsables de l'information scientifique et technique des organismes de recherche français publics ou d'utilité publique
https://www.eprist.fr/

ESTHER Database
http://bioweb.supagro.inra.fr/ESTHER/general?what=index

EU-SOLARIS
http://www.eusolaris.eu/

EUDAT, European Data
https://eudat.eu/

ESO, European Southern Observatory (Observatoire Européen Austral, en français)
https://www.eso.org/public/

Eurostat
https://ec.europa.eu/eurostat/

Figshare
https://figshare.com/

GANIL, Grand Accélérateur National d'Ions Lourds
https://www.ganil-spiral2.eu/

GAPHYOR, Gaz-Physique Orsay [l'entrepôt n'est plus disponible en ligne à ce jour]
https://cat.opidor.fr/index.php/GAPHYOR

GEANT, Gigabit European Advanced Network Technology
https://www.geant.org/

GenBank
https://www.ncbi.nlm.nih.gov/genbank/

GenomEast
http://genomeast.igbmc.fr/

GEO, Gene Expression Omnibus
https://www.ncbi.nlm.nih.gov/geo/

GDSI, Global Spatial Data Infrastructure Association
http://gsdiassociation.org/

Grand Collisionneur de Hadrons, ou Large Hadron Collider (LHC)
https://home.cern/fr/science/accelerators/large-hadron-collider

Hcéres, Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur
https://www.hceres.fr/

Huma-Num
https://www.huma-num.fr/

Ifremer, Institut Français de Recherche pour l'Exploitation de la Mer
https://wwz.ifremer.fr/

ILL, Institut Laue-Langevin
https://www.ill.eu/

ILL Data Portal
http://data.ill.eu

IMN, Institut des Matériaux Jean Rouxel
https://www.cnrs-imn.fr/

Ined, Institut National d'Etudes Démographiques
https://www.ined.fr/

InfoTerre
http://infoterre.brgm.fr/

Inist-CNRS, Institut de l'Information Scientifique et Technique
https://www.inist.fr/

INRA, Institut National de la Recherche Agronomique
http://www.inra.fr/

INSEE, Institut National de la Statistique et des Études Économiques
https://www.insee.fr

INSERM, Institut National de la Santé et de la Recherche Médicale
https://www.inserm.fr/

IO Data Science
https://io.datascience-paris-saclay.fr/

IPEV, Institut Polaire français Paul-Emile Victor
https://www.institut-polaire.fr/

Institut Pierre Louis d'Epidémiologie et de Santé Publique (UMR S 1136 IPLESP)
https://www.iplesp.upmc.fr/

IRSTEA, Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture
https://www.irstea.fr/

IrsteaData catalogue
https://data.irstea.fr/

Isidore
https://isidore.science/

Kinsources
https://www.kinsources.net/

LIBER, Ligue des Bibliothèques Européennes de Recherche
https://libereurope.eu/

Laboratoire d'Analyse et d'Architecture des Systèmes (UPR 8001 LAAS)
https://www.laas.fr/

Laboratoire de physique des gaz et des plasmas (UMR 8578 LPGP)
http://www.lpgp.u-psud.fr/

MédiHAL
https://medihal.archives-ouvertes.fr/

MISTRALS, Mediterranean Integrated Studies at Regional and Local Scales
http://www.mistrals-home.org/

MSHE Ledoux, Maison des sciences de l'homme et de l'environnement Claude Nicolas Ledoux
https://mshe.univ-fcomte.fr/

Nakala
https://www.nakala.fr/

Nakalona
https://www.nakalona.fr/

Observatoire Midi-Pyrénées (OMP)
http://www3.obs-mip.fr/omp/

Observatoire Sociologique du Changement (UMR 7049 OSC)
http://www.sciencespo.fr/osc/

Observatoire des Sciences de l'Univers Terre-Homme-Environnement-Temps-Astronomie (OSU THETA)
https://theta.obs-besancon.fr/

OpenAIRE
https://www.openaire.eu/

Ortolang, Outils et Ressources pour un Traitement Optimisé de la Langue
https://www.ortolang.fr/

OTELo, Observatoire Terre Environnement de Lorraine
http://otelo.univ-lorraine.fr/

PANGAEA
https://www.pangaea.de/

PerSCiDO
https://persyval-platform.univ-grenoble-alpes.fr/

PERSYVAL-Lab
https://persyval-lab.org/

PID OPIDoR
https://opidor.fr/identifier/

Pôle HPC Unistra
https://services-numeriques.unistra.fr/les-services-aux-usagers/hpc.html

Politiques publiques, Action politique, Territoires (UMR 5194 PACTE)
https://www.pacte-grenoble.fr/

Portail Epidémiologie France
https://epidemiologie-france.aviesan.fr/

Portail OPIDoR
https://opidor.fr/

PRACE, Partnership for Advanced Computing in Europe
http://www.prace-ri.eu/

PROGEDO, Production et Gestion des Données (USR 2006)
http://www.progedo.fr/

Protein Data Bank
https://www.rcsb.org/

RDA, Research Data Alliance
https://www.rd-alliance.org/

Re3data, Registry of Research Data Repositories
https://www.re3data.org/

Recolnat
https://www.recolnat.org/fr/

RECORD, Residential Environment Coronary Heart Desease
http://www.record-study.org/accueil.html

ResearchGate
https://www.researchgate.net/

Réseau Quetelet, aujourd'hui renommé Quetelet PROGEDO Diffusion
http://quetelet.progedo.fr/

RESIF, Réseau Sismologique et Géodésique Français
https://www.resif.fr/

RESIF Seismic Data Portal
http://seismology.resif.fr/

SEANOE
https://www.seanoe.org/

SIMBAD
http://simbad.u-strasbg.fr/

Site d'information sur les Données de la Recherche
http://www.donneesdelarecherche.fr/

Synchrotron SOLEIL
https://www.synchrotron-soleil.fr/

Télescope Canada-France-Hawaii
https://www.cfht.hawaii.edu/fr/

The Plant List
http://www.theplantlist.org/

Tropicos
https://www.tropicos.org/

Université de Lorraine
https://www.univ-lorraine.fr/

Université de Nice Sophia-Antipolis
http://unice.fr/

Université de Nice Sophia-Antipolis : Equipe d'accompagnement
https://bu.univ-cotedazur.fr/fr/utiliser-nos-services/services-a-la-recherche

Université de Strasbourg
https://www.unistra.fr/

Université de Strasbourg : Equipe d'accompagnement
https://bu.unistra.fr/opac/article/la-gestion-des-donnees-de-la-recherche-a-luniversite-de-strasbourg/services_donnees

Université Sorbonne Paris Cité : Equipe d'accompagnement
https://appui-recherche.univ-paris-diderot.fr/gerer-ses-donnees-scientifiques-le-plan-de-gestion-de-donnees

URFIST, Unité Régionale de Formation à l'Information Scientifique et Technique
https://urfistinfo.hypotheses.org/

USPC, Université Sorbonne Paris-Cité
http://www.sorbonne-paris-cite.fr/

VizieR
http://vizier.u-strasbg.fr/

VRP-REP, Vehicle Routing Problem Repository
http://www.vrp-rep.org/

Webservice-Energy Catalog
http://geocatalog.webservice-energy.org/

Zenodo
https://zenodo.org/

Annexes

Annexe 1 - Offre de poste pour la réalisation d'une cartographie nationale des services de données scientifiques

L'université de Strasbourg recrute un(e) chargé(e) d'études en Sciences de l'information et de la communication

Disponibilité : novembre 2015 - avril 2016
Statut : CDD à temps plein
Durée : 6 mois
Employeur : Université de Strasbourg
Rémunération : Rémunération selon la grille de la fonction publique - catégorie A (ingénieur d'études)
Niveau d'études : Master
Expérience requise : aucune
Lieu de travail : Strasbourg
Date limite de dépôt des candidatures : 01 septembre 2015
Branche d'activité professionnelle : BAP F : Information, Documentation, Culture, Communication, Edition, TICE
Corps : Ingénieur d'études

Mission

Dans le cadre des activités du segment 10 « Données de la recherche » de la Bibliothèque Scientifique Numérique, il s'agit d'établir une cartographie des centres de données et services de données sur l'ensemble du territoire national.

L'étude portera sur toutes les disciplines scientifiques et tous les organismes de recherche financés en tout ou partie par l'état.
L'analyse s'appuiera principalement sur des recensions sur internet, des recherches bibliographiques, des entretiens semi-directifs ainsi que sur des résultats d'enquêtes menées dans les organismes.

L'étude devra examiner les points de vue des différents acteurs au contact des données dans les processus de la recherche : scientifiques, informaticiens, professionnels de l'information scientifique et technique.
L'étude permettra de recenser le : lieu, les autorités de tutelle, le niveau d'intervention

(discipline, organisme, site, laboratoire, équipe...), le type de données..., l'insertion dans des réseaux européens, mondiaux…, l'interopérabilité, le type d'infrastructure, les moyens humains et financiers dédiés…Elle permettra aussi d'identifier les pratiques des champs disciplinaires (quand les scientifiques produisent-ils, où vont-ils déposer leurs données ?).

L'analyse s'attachera à dégager une vision prospective en faisant ressortir les difficultés, points bloquants ainsi que les opportunités de développement de ces centres.

Le travail s'effectuera avec l'appui des membres du segment BSN10 –données de la Recherche de la Bibliothèque Numérique.
Le chargé d'études sera intégré à l'équipe « Archives ouvertes de la Connaissance » de l'Université de Strasbourg.
Un rapport et un site wiki seront produits à l'issue de l'étude.

Activités principales

Production d'une grille d'analyse (liste de métadonnées) permettant de décrire les objets recensés.
Recension sur internet des centres de données et services de données.
Recherches et analyses bibliographiques
Définition d'une grille et gestion d'entretiens semi-directifs.
Des analyses approfondies seront effectuées sur une sélection de centres de données, dans une vision prospective.
Création et alimentation d'un wiki. Définition des principes d'alimentation et de mise à jour.
Rédaction d'un rapport.

Activités associées

L'étude sera menée sous la direction rapprochée d'un professionnel de l'information et le suivi d'un comité de pilotage composé de membres de BSN10.
Présentations régulières d'états d'avancement au comité de pilotage.
Présentation des résultats de l'étude aux membres du groupe BSN10.
Actions de communication sur les résultats de l'étude à l'échelle nationale et Internationale.

Compétences principales

Avoir une bonne connaissance des problématiques et pratiques de la communication scientifique.
Connaissance des normes et standards de l'Internet et du multimédia.
Bonne connaissance des enjeux et de l'actualité de la publication scientifique (open access, open data, open science).

Excellente maîtrise de l'expression écrite en français.
Maîtrise de l'anglais (compréhension et expression à l'écrit).

Formations et expérience professionnelle

Master en sciences de l'information et de la communication ou équivalent.
Pas d'expérience professionnelle requise.

Candidature

Adresser C.V. et lettre de candidature à l'attention de :

Université de Strasbourg
Service commun de la documentation
Madame Dominique Wolf, directrice du SCD de l'Université de Strasbourg
Contact : wolf@unistra.fr

Annexe 2 - Équipe projet de Cat OPIDoR (octobre 2016 – septembre 2017)

Copilotes du segment BSN10 : Francis ANDRÉ et Paul Antoine HERVIEUX

Chargée d'étude BSN10 : Violaine REBOUILLAT

Équipe Valorisation des données de la recherche (Inist-CNRS) : Ourida ABERKANE, Anne CIOLEK-FIGIEL et Marie-Christine JACQUEMOT

Équipe web (Inist-CNRS) : Catherine VERNISSON, Cécilia FABRY

Équipe informatique (Inist-CNRS) : Benjamin FAURE et Fabien VILLA

Annexe 3 - Exemples de grilles d'analyse de deux types de services de données

Exemple 1 : Grille d'analyse des services de type accompagnement

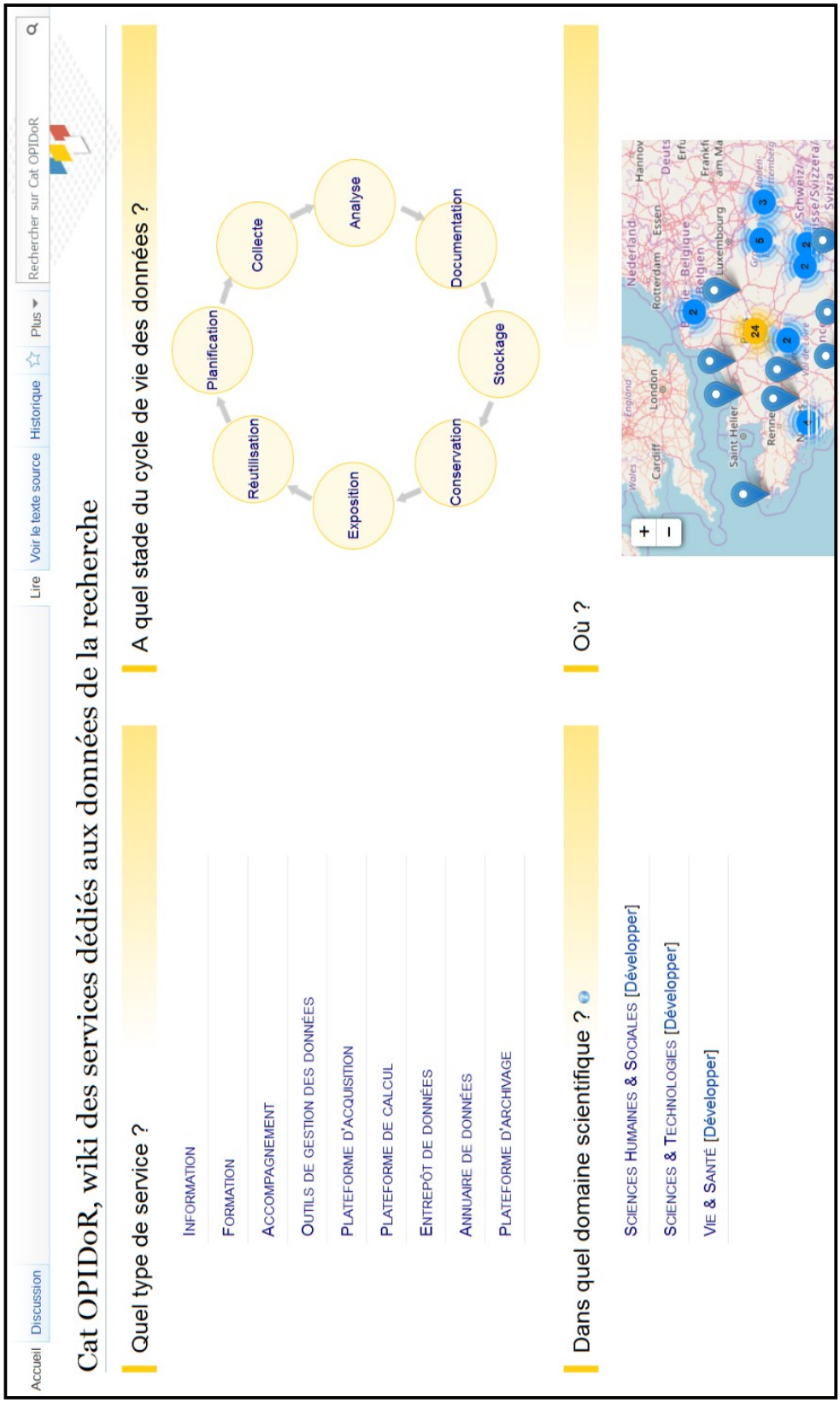
IDENTITÉ DU SERVICE	
Nom	
Contact	
Date de création	
GESTION DU SERVICE	
Opérateur(s)	Quelle(s) personne(s)/équipe(s)/service(s) en assure(nt) la gestion ?
Ressources financières	Quel est le budget dédié à ces outils et services ?
Financeur(s)	Quelle est l'origine du financement ?
Ressources humaines	Combien de personnes sont mobilisées (si possible en équivalent temps plein) ? Quels sont leurs statuts professionnels ?
FONCTIONS DU SERVICE	
Rôle(s)	Sur quel(s) aspect(s) de la gestion des données intervient cette équipe ?
Public cible	A quel public est/(sont) destiné(s) ce(s) service(s) et outil(s) ?
ÉVALUATION DU SERVICE	
Point(s) fort(s)	
Point(s) faible(s)	Quelles limites peuvent être soulevées par les membres de l'équipe et les destinataires du service ?
Perspectives de développement	Cette équipe est-elle amenée à évoluer dans son rôle ou sa composition ? Si oui, dans quel sens ?

Exemple 2 : Grille d'analyse des services de type entrepôt de données

IDENTITÉ DU SERVICE	
Nom	
Contact	
URL	
DOI Re3data	Si l'entrepôt est répertorié dans le Re3data, quel est son identifiant ?
Langue(s)	En quelle(s) langue(s) est accessible l'entrepôt ?
Contact	
Année de lancement	
GESTION DU SERVICE	
Opérateur(s)	Quelle(s) personne(s)/service(s)/équipe(s) assure la gestion du service ? Distinguer, si besoin, le responsable de la maintenance informatique du gestionnaire de l'entrepôt.
Ressources financières	Quel est le budget dédié à l'entrepôt ?
Financeur(s)	Quelle est l'origine du financement ?
Ressources humaines	Combien de personnes assurent la gestion de l'entrepôt (si possible en équivalent temps plein) ? Quels sont leurs statuts professionnels ?
FONCTIONS DU SERVICE	
Type d'entrepôt	Disciplinaire/multidisciplinaire/relatif à un projet/autre
Public cible	A quel public est/(sont) destiné(s) ce(s) service(s) et outil(s) ?
Domaine(s) disciplinaire(s)	De quelle(s) discipline(s) scientifique(s) relèvent les données traitées ? (Cf. classification du European Research Council)
Type(s) de données traitées	L'entrepôt accueille-t-il des données de forme(s) spécifique(s) ? Si oui, laquelle/lesquelles ? (Cf. formats listés par le Re3data)
Type(s) d'accès à l'entrepôt	L'accès à l'entrepôt est-il ouvert/fermé/restreint ? S'il est restreint, préciser : accès payant/après inscription/réservé aux membres de l'institution/autre.
Public cible	A quelle(s) communauté(s) d'utilisateurs est destiné l'entrepôt ? (membres de l'institution/communauté scientifique/communauté professionnelle/tout public/autre)
Solution logicielle	Sur quelle solution logicielle est basé le système informatique de l'entrepôt ?
Interface de programmation (API)	Quel type d'API est utilisé ? Quelle est son URL ?
Interopérabilité	Le système informatique est-il interopérable ?
Métadonnées associées	Quelles métadonnées sont renseignées pour décrire les données ? Citer les différents champs.
Standard(s) de métadonnées	Les métadonnées sont-elles renseignées dans un format connu ? Si oui le(s)quel(s) ?
Export des notices descriptives	Les métadonnées descriptives d'un jeu de données peuvent-elles être exportées ? Si oui, sous quel(s) format(s) ?
Lien avec la publication	Les données peuvent-elles être liées à une publication ?
Identification des données	Un identifiant est-il associé aux données ?
Identification des contributeurs	Un identifiant est-il associé aux contributeurs ?
Type(s) de stockage	

Mode(s) de dépôt des données	Le dépôt des données est-il ouvert/fermé/restreint ? S'il est restreint, préciser : dépôt payant/après inscription/réservé aux membres de l'institution/autre.
Responsable(s) du dépôt des données	Qui est chargé de déposer les données et de renseigner leurs métadonnées ?
Type(s) d'accès aux données	L'accès aux données est-il ouvert/fermé/restreint ? S'il est restreint, préciser : accès payant/après inscription/après une période d'embargo/réservé aux membres de l'institution/autre.
Conditions d'utilisation des données	L'entrepôt définit-il des conditions pour la réutilisation des données et des métadonnées ? Si oui, donner l'URL où en sont précisés les termes.
Licence(s) d'utilisation	La réutilisation des données et des métadonnées est-elle soumise à une licence ? Si oui, laquelle/lesquelles ?
Citation des données	L'entrepôt fournit-il des recommandations pour les utilisateurs qui souhaitent citer un jeu de données ? Si oui, donner l'URL.
UTILISATION DU SERVICE	
Taille de l'entrepôt	Combien de données sont contenues dans l'entrepôt ? Préciser la date du relevé. Distinguer éventuellement le nombre de notices du nombre de données.
Fréquentation de l'entrepôt	Quel est le nombre moyen de visites par jour/année ?
ÉVALUATION DU SERVICE	
Qualité des données et métadonnées	Comment est assurée la qualité des données et des métadonnées ?
Certification	La qualité de l'entrepôt est-elle certifiée par un organisme accrédité ? Si oui, le(s)quel(s).
Influences politiques	L'entrepôt s'aligne-t-il sur des recommandations institutionnelles/(inter)nationales/européennes ?
Visibilité	L'entrepôt est-il référencé sur un site web ou dans un catalogue en ligne ?
Partenaire(s)	L'entrepôt fait-il partie d'un réseau institutionnel/national/européen/international...?
Point(s) fort(s)	
Point(s) faible(s)	Quelles limites peuvent être soulevées par les administrateurs et les utilisateurs de l'entrepôt ?

Annexe 4 - Cat OPIDoR : Page d'accueil



Annexe 5 - Cat OPIDoR : Page de résultats du domaine Sciences & Technologies (affichage sous forme de tableau)

Page Discussion Lire Voir le texte source Historique Plus Rechercher sur Cat OPIDoR

Sciences & Technologies

Chimie de synthèse et matériaux Chimie physique et analytique Constituants fondamentaux de la matière Ingénierie des produits et des procédés Ingénierie des systèmes et de la communication Mathématiques

Physique de la matière condensée Sciences de l'Univers Sciences de la Terre Sciences informatiques et informatique

Afficher les entrées 10 Rechercher :

Services	Type de données	Thématique/Mots clés	Type de service
Agence DataCite	-	Identifiant pérenne DOI Métadonnées DataCite Citation Accessibilité	Outils de gestion des données
AVISO+	Données environnementales, Données altimétriques, Données océnographiques, Données géophysiques, Données instrumentales, Données marégraphiques, Données topographiques	Altimétrie spatiale Océanographie Climatologie Hydrologie Glaciologie Météorologie Hauteur de mer Hauteur des vagues Vitesse du vent	Entrepôt de données
BASS2000	Données d'observation, Images, Données audiovisuelles, Cartes synoptiques, Fichiers compressés, GIF, PNG, JPEG, MPEG, PS	Astronomie Astrophysique Soleil	Entrepôt de données
CALI	Données de simulation numérique	Calcul intensif Electronique Optique Communication	Plateforme de calcul

Annexe 6 - Cat OPIDoR : Page de résultats du sous-domaine Sciences du système Terre (affichage sous forme d'index)

Page | Discussion | Lire | Voir le texte source | Afficher l'historique | Rechercher dans Cat OPIDoR

Sciences du Système Terre

Géographie physique, géologie, géophysique, sciences de l'atmosphère, océanographie, climatologie, cryologie, écologie, changements environnementaux globaux, cycles biogéochimiques, gestion des ressources naturelles

Liste des services

A ce jour 74 services sont référencés dans Cat OPIDoR

A
- AVISO+
- Aeris portail

C
- C3I
- CALI
- CCRT
- CDGP
- CORIOLIS
- Carmen
- CeDONA
- Cirad : Annuaire de données
- Collec-Science
- Concordia

D
- DATA BRGM

D (suite)
- DYNALIT
- Data eaufrance.fr
- Data shom.fr
- DataSuds
- Dinamis

E
- E.cenaris

E
- ECORD/IODP
- EMSO-France : Plateforme d'acquisition
- ESPRI
- Earth Orientation Center

F
- FOF
- ForM@Ter portail
- France Grilles

G
- GEOSUD : Portail images satellites
- GeOrchestra
- Gestion et diffusion des données Irstea
- Ginco
- Géoportail

I
- IAGOS Data Portal
- ICARE
- IDOC
- IDRIS
- INPN
- ISGI
- Ifsttar Dataverse
- InfoTerre

I (suite)
- Intelespace
- IrsteaData

K
- Kalideos

M
- MISTRALS database
- MSH Clermont-Ferrand : Plateforme informatique - Bases de données
- MUST
- Mésocentre ESPRI
- Mésocentre SIGAMM : Plateforme de calcul

N
- Navigae

O

P
- PADC
- PCIM
- Peps
- Persée.fr
- Plateforme A2S
- Portail spatial

R
- RESIF Seismic data portal
- Recolnat
- ReefTemps

S
- S-CAPAD
- SAFIRE : Plateforme d'acquisition
- SAFIRE-data
- SANDRE

S (suite)
- SATMOS
- SCD Université Claude Bernard Lyon 1 : Formation
- SEANOE
- Sextant

T
- TGCC
- Theia portail
- Théia Animation régionale

W
- Webservice-Energy Catalog

Z
- Zenodo

Annexe 7 - Cat OPIDoR : Champs descriptifs d'un service

Page | Discussion | Lire | Modifier avec formulaire | Modifier | Afficher l'historique | Rechercher dans Cat OPIDoR

CINES Archivage numérique pérenne

La plateforme d'archivage du CINES a vocation à **archiver les données** et les documents numériques produits par la communauté française de l'Enseignement Supérieur et de la Recherche. Elle propose des solutions d'**archivage numérique payantes**, sur le moyen et le long terme, et offre à ses utilisateurs une **expertise** dans les domaines informatique et archivistique. La **sécurité** et l'**intégrité des données** sont garanties par un ensemble de procédures telles que l'attribution de **métadonnées**, le choix de **formats de fichiers pérennes**, la réplication des données et un environnement informatique protégé.

Domaines scientifiques :
Sciences Humaines & Sociales, Sciences & Technologies, Vie & Santé

Thématique et/ou mots clés :
Modèle OAIS | Archivage pérenne | Archivage intermédiaire

Type de données :
Données d'observation, Données d'expérimentation, Résultats de calcul

Communauté d'utilisateurs :
ESR français

Usagers et bénéficiaires :
Utilisateurs autorisés

Conditions d'usage :
Les demandes sont sélectionnées par le conseil d'administration du Département Archivage et Diffusion du CINES.

Modèle économique :
Payant

Certification/Label qualité :
https://www.datasealofapproval.org/

Conditions générales d'utilisation :
https://www.cines.fr/mentions-legales/

Services proposés par la structure d'appartenance [modifier]

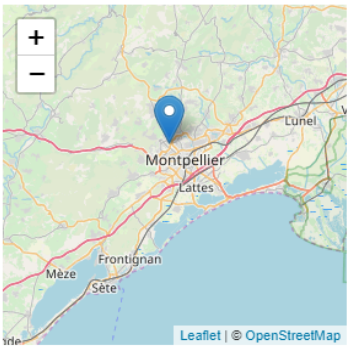
- CINES : Plateforme de calcul
- **CINES Archivage numérique pérenne**

CINES Archivage numérique pérenne

Type de service	Plateforme d'archivage
Statut	En production
Autres noms	Centre Informatique National de l'Enseignement Supérieur Archivage Numérique Pérenne , PAC - Plateforme d'Archivage au CINES, CINES PAC, CINES Preservatio, CINES Long Term Preservation, PAC - Archiving Platform at CINES
URL	https://www.cines.fr/archivage/
Contact	Mel administratif : contact-admin@cines.fr Mel technique : svp@cines.fr
Localisation	Montpellier
Structure d'appartenance	CINES
Identifiant dans un autre catalogue	http://doi.org/10.17616/R3R30N (re3data)

CYCLE DE VIE DES DONNÉES

Ce service intervient au cours des stades du cycle de vie suivants :



Annexe 8 – Tableau analytique des 44 services de gestion et d'ouverture des données étudiés dans la 3ème partie

Nom	**SEANOE**	**Ortolang**
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	http://www.seanoe.org/	https://www.ortolang.fr/
Structure d'appartenance	SISMER	Atilf
Date de création	2015	2013
Source de financement/Budget	IFREMER	EquipEx puis demande de financement au CNRS (estimée à 150 000€ par an)
Ressources humaines	35 ETP dans l'unité Informatique et Données marines	2 ETP + 5 chercheurs à temps partiel
Profils de postes	Informaticiens ; Scientifiques ; Techniciens	Informaticiens (Inist-CNRS) ; Chercheurs et ingénieurs (ATILF)
Description	SEANOE est une solution de publication des données scientifiques marines. Elle permet aux scientifiques de publier un jeu de données, en libre accès ou avec un embargo d'une période de 2 ans maximum, et de le citer de manière fiable et pérenne grâce à l'attribution d'un DOI (Digital Object Identifier). La réutilisation des données est soumise à une licence Creative Commons.	ORTOLANG est un équipement d'excellence, offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement. Il propose un accompagnement au dépôt des données, ainsi que différentes modalités de diffusion des données (en accès libre, en accès restreint ou sous embargo).
Disciplines	Vie & Santé	Sciences humaines et sociales
Thématique	Sciences marines	Sciences du langage
Communauté d'utilisateurs	ESR français et étranger	ESR français
Utilisateurs finaux	Tout public	Tout public
Conditions d'usage	Déposants : création d'un compte personnel Utilisateurs finaux : libre accès	Déposants : ils doivent au préalable créer un compte personnel. Usagers : ils peuvent consulter librement les notices descriptives de l'ensemble des données de l'entrepôt, mais ne peuvent accéder qu'aux fichiers de données qui ont été mis en libre accès par leur déposant.
Accès à l'entrepôt	Libre	Libre
Accès aux données	Libre ; Embargo	Libre ; Restreint ; Embargo
Dépôt de données	Restreint (registration)	Restreint (registration)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	100 à 200 jeux de données (le 22.03.2016)	40 corpus, 11 outils, 7 lexiques (le 21.03.16)
Utilisation (au 23/07/2019)	487 jeux de données	420 ressources

Nom	**Nakala**	**ArkeoGIS**
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	https://www.nakala.fr/	http://arkeogis.org/
Structure d'appartenance	Huma-Num	Archimède
Date de création	2014	2009
Source de financement/Budget	UMS 3598	80 000€ (LabEx + IdEx)
Ressources humaines	11 ETP (Huma-Num)	1 ETP
Profils de postes	Chercheurs ; informaticiens ; documentalistes ; gestionnaires de données	Chercheurs ; Ingénieurs
Description	NAKALA est un service mis en place par la TGIR Huma-Num pour déposer, documenter et diffuser les données de la recherche. Ce service permet à des équipes de recherche, qui en font la demande, de déposer leurs données numériques (fichiers texte, son, image, vidéo) dans un entrepôt sécurisé qui assure à la fois l'accessibilité aux données et leur citabilité (attribution d'identifiant pérenne de type handle) dans le temps. NAKALA propose des services intéropérables d'accès aux données elles-mêmes et de présentation des métadonnées descriptives (format standard Dublin Core étendu). Les données sont stockées et sauvegardées régulièrement sur les serveurs sécurisés de la TGIR Huma-Num au sein du centre de calcul de l'IN2P3-CNRS.	ArkeoGIS est un Système d'Information Géographique (SIG) multilingue, initialement développé afin de mutualiser les données archéologiques et paléoenvironnementales de la vallée du Rhin. Il permet aujourd'hui de mettre en commun les données scientifiques spatialisées concernant le passé, depuis la Préhistoire jusqu'à nos jours. Les bases de données sont issues de travaux de chercheurs institutionnels, de doctorants, d'étudiants en master, de sociétés privées et de services d'archéologie. Elles sont stockées sur la grille de services de la TGIR Huma-Num et archivées dans le cadre du service d'archivage à long terme Huma-Num/CINES. En raison de leur caractère sensible, qui pourrait conduire à un pillage des gisements archéologiques, l'accès à l'outil est réservé aux professionnels de l'archéologie, issus d'institutions de recherche ou d'organisations à but non lucratif.
Disciplines	Sciences humaines et sociales	Sciences humaines et sociales
Thématique		Archéologie, Histoire, Géographie
Communauté d'utilisateurs	ESR français	Professionnels de l'archéologie français et étrangers
Utilisateurs finaux	Déposants : Enseignants-chercheurs ; Chercheurs ; Doctorants	Chercheurs, Doctorants, Etudiants en Master, Services régionaux ou nationaux
Conditions d'usage	Pour utiliser Nakala : adresser une demande à cogrid@huma-num.fr.	Pour déposer ou consulter des données : faire une demande de création de compte
Accès à l'entrepôt	Restreint	Restreint (registration)
Accès aux données	Restreint	Restreint (registration)
Dépôt de données	Restreint (sélection + registration)	Restreint (registration)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	113 projets de recherche collectifs disposent d'une interface de gestion dans NAKALA (le 04.03.2016)	64 bases de données + environ 200 comptes utilisateurs ouverts (le 28.01.16)
Utilisation (au 23/07/2019)	NC	99 bases de données

Nom	Portail Data Inra*	PerSCiDO*
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En test	En test
URL	https://data-test.jouy.inra.fr	https://persyval-platform.univ-grenoble-alpes.fr
Structure d'appartenance	INRA	PERSYVAL-Lab
Date de création	2015	2015
Source de financement/Budget	100 000€ à la création	LabEx (50 000€ à la création)
Ressources humaines	Equipe projet de 15 personnes	2 ETP CDD
Profils de postes	Chercheurs ; Ingénieurs développement ; Professionnels de l'IST	Ingénieurs développement ; Chercheurs ; Professionnels de l'IST
Description	Le Portail Data Inra vise à référencer l'ensemble des données scientifiques de l'INRA. Sa fonction annuaire doit permettre de décrire les données gérées et partagées via des entrepôts internes ou externes existants. Les données non préservées dans un entrepôt peuvent éventuellement être déposées sur le portail et être diffusées en accès libre ou restreint. Pour chaque jeu de données déposé ou décrit, un identifiant pérenne lui est associé. Librement accessible, le portail favorisera la lisibilité et la visibilité des ressources produites par l'INRA.	PerSCIDO est une plateforme de partage de jeux de données de recherche initiée par le labex PERSYVAL-lab, centrée sur des métadonnées riches et flexibles. Son infrastructure et son modèle de métadonnées sont basés sur les standards du Linked Open Data (RDF) pour assurer l'intéropérabilité avec d'autres plateformes. PerSCIDO incite les chercheurs à adopter de bonnes pratiques : référencement par des Identifiants pérennes (DOI), droit d'usage (attribution licence licence Creative Commons) et la citation des jeux de données.
Disciplines	Sciences & Technologies ; Vie & Santé	Sciences Humaines & Sociales ; Sciences & Technologies ; Vie & Santé
Thématique	Ecologie	
Communauté d'utilisateurs	Membres de l'INRA	Membres du labex PERSYVAL-lab
Utilisateurs finaux	ESR français et étranger	Tout public
Conditions d'usage	Description et/ou dépôt de données réservé aux membres de l'INRA. Libre consultation des métadonnées. Accès restreint ou ouvert aux données, selon les conditions définies par les contributeurs.	Déposants : Membres de PERSYVAL-lab. Utilisateurs : Accès ouvert ou accès restreint.
Accès à l'entrepôt	Libre	Libre
Accès aux données	Libre ; Restreint ; Embargo	Libre ; Restreint
Dépôt de données	Restreint (membres)	Restreint (membres)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC (en cours de création au moment de l'étude)	NC (en cours de création au moment de l'étude)
Utilisation (au 23/07/2019)	78 183 jeux de données	34 jeux de données

Nom	**BeQuali***	**MédiHAL***
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	http://bequali.fr/fr/	https://medihal.archives-ouvertes.fr/
Structure d'appartenance	CDSP	CCSD
Date de création	2012	2010
Source de financement/Budget	EquipEx DIME-SHS	UMS 3668
Ressources humaines	5 ETP	15 ETP dans l'équipe du CCSD
Profils de postes	Ingénieurs de recherche ; Archivistes ; Professionnels de l'IST ; Ingénieurs développement de DIME-SHS	Ingénieurs développement ; Professionnels de l'IST
Description	La banque d'enquêtes qualitatives en sciences humaines et sociales (beQuali) fait partie de l'équipement d'excellence DIME-SHS. Elle offre la possibilité de valoriser les données de recherche qualitatives à travers la numérisation des corpus et l'enrichissement des données brutes au moyen de la documentation produite (classement, métadonnées, contextualisation). BeQuali propose, sous réserve d'autorisation, un accès à ces matériaux d'enquêtes et à une documentation restituant le contexte de leur production pour des analyses secondaires. Ce patrimoine scientifique est mis à disposition en accès sécurisé sur le site beQuali.fr et sur le portail Quetelet. Il est archivé de manière pérenne au Centre informatique national de l'enseignement supérieur (CINES).	MédiHAL a été créé en 2010 par le Centre pour la communication scientifique directe (CCSD). L'archive ouverte permet de déposer des données visuelles et sonores (images fixes, vidéos et sons) produites dans le cadre de la recherche scientifique. Les données (les fichiers et leurs métadonnées) sont archivées à long terme au Centre Informatique National de l'Enseignement Supérieur (CINES). La géo-localisation des images déposées est intégré au formulaire de dépôt par un accès cartographique couplé au référentiel géographique GeoNames.org. Les images comportant des métadonnées GPS sont automatiquement positionnées sur la carte.
Disciplines	Sciences Humaines & Sociales	Sciences Humaines & Sociales, Sciences & Technologies, Vie & Santé
Thématique	Sciences sociales, Sciences politiques, Ethnologie, Anthropologie	
Communauté d'utilisateurs	ESR français	ESR français
Utilisateurs finaux	L'accès aux enquêtes est restreint à la communauté scientifique.	Chercheurs, Enseignants-Chercheurs
Conditions d'usage	Déposants : La sélection des enquêtes s'effectue dans le cadre d'appels à proposition Utilisateurs : l'accès aux enquêtes est sécurisé et restreint à la communauté scientifique. La demande d'accès aux enquêtes se fait via le site Quetelet, portail français d'accès aux données pour les sciences humaines et sociales, après inscription.	Le déposant crée et renseigne les métadonnées et choisit la licence de diffusion. L'équipe du CCSD vérifie la conformité des informations. Les utilisateurs sont soumis aux règles du bon usage des données dans le monde scientifique : respect des travaux originaux, mention des auteurs originaux et du lieu de conservation, respect et accord des personnes photographiées.
Accès à l'entrepôt	Libre	Libre
Accès aux données	Restreint (registration)	Libre
Dépôt de données	Restreint (sélection + registration)	Restreint (registration)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	5 enquêtes (le 15.04.16)	22 476 documents (le 12.02.16)
Utilisation (au 23/07/2019)	12 enquêtes	42 046 jeux de données

Nom	RESIF Seismic data portal	CDPP: Entrepôt de données
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	http://seismology.resif.fr/	http://www.cdpp.eu/
Structure d'appartenance	RESIF	CDPP
Date de création	NC	NC
Source de financement/Budget	NC	NC
Ressources humaines	NC	NC
Profils de postes	NC	NC
Description	Le portail Réseau sismologique & géodésique français (RESIF) donne accès aux données sismologiques des réseaux permanents et mobiles des institutions de recherches françaises et de leurs partenaires. Les données sismologiques de RESIF sont en accès libre et sont distribuées selon les standards et formats internationaux spécifiques à chaque type de données (Standard for the Exchange of Earthquake Data).	Le Centre de données de la physique des plasmas (CDDP) assure la conservation à long terme des données pertinentes à la physique des plasmas naturels dans le système solaire et les rend accessibles et exploitables à la communauté internationale. Les données archivées ont été obtenues à bord des satellites ou des observatoires terrestres depuis plus de 40 ans. Le CDPP met à disposition des outils de visualisation, de manipulation et d'analyse des ensembles de données hétérogènes. Le CDPP est très impliqué dans le développement de l'interopérabilité et participe à plusieurs projets de l'Observatoire Virtuel.
Disciplines	Sciences & Technologies	Sciences & Technologies
Thématique	Géophysique, Géodésie, Sismologie	Physique des plasmas naturels, Magnétosphère terrestre, Astrophysique, Héliophysique
Communauté d'utilisateurs	Membres du consortium RESIF	Membres de la communauté d'astrophysique
Utilisateurs finaux	Tout public	Chercheurs
Conditions d'usage	Dépôt des données réservé aux membres du consortium RESIF. Utilisateurs : accès libre ou accès restreint après une période de 2 à 3 ans dans le cas des réseaux mobiles.	Dépôt des données issues des missions spatiales. Utilisateurs : accès libre avec création d'un compte.
Accès à l'entrepôt	Libre	Libre
Accès aux données	Libre ; Restreint (membres) ; Embargo	Restreint (registration)
Dépôt de données	Restreint (membres)	Fermé
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	NC
Utilisation (au 23/07/2019)	NC	NC

Nom	**DATA BRGM**	**Webservice-Energy Catalog**
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	https://data.brgm.fr/	http://geocatalog.webservice-energy.org/
Structure d'appartenance	BRGM	Centre O.I.E.
Date de création	2014	2008
Source de financement/Budget	BRGM	Mines ParisTech
Ressources humaines	NC	4 ETP
Profils de postes	NC	Ingénieurs ; Informaticien
Description	L'entrepôt DATA BRGM permet aux chercheurs du BRGM de conserver et de décrire les données de recherche qu'ils ont produites. Lorsqu'il dépose un jeu de données dans l'entrepôt, le chercheur renseigne un certain nombre d'informations de contexte, conformes au schéma de métadonnées de DataCite. Données et métadonnées sont ensuite stockées de manière sécurisée.	Webservice-Energy Catalog est un entrepôt de données basé sur la solution GeoNetwork, un environnement open source de gestion de l'information spatiale. Il permet : de rechercher des données géo-spatiales dans les domaines de l'énergie et de l'environnement ; d'afficher les données sous forme de carte géographique interactive ; d'accéder à leurs métadonnées et d'exporter celles-ci sous différents formats (XML, PDF,…). Les données de l'entrepôt sont régulièrement moissonnées par le Global Earth Observation System of System (GEOSS) et accessibles depuis le GEOSS Portal. Webservice-Energy Catalog s'appuie, par ailleurs, sur les standards de l'Open Geospatial Consortium (OGC) et du World Wide Web Consortium (W3C).
Disciplines	Sciences & Technologies, Vie & Santé	Sciences & Technologies
Thématique	Géographie physique, Géologie, Géophysique, Ecologie	Sciences de la Terre, Energies renouvelables, Environnement
Communauté d'utilisateurs	Membres du BRGM	Centre O.I.E. et ses partenaires
Utilisateurs finaux	Chercheurs	Tout public
Conditions d'usage	Le dépôt et la consultation de données sont réservés aux membres du BRGM, qui accèdent à l'entrepôt en s'authentifiant à l'aide d'identifiants de connexion.	Dépôt de données : l'équipe administratrice de la plateforme se charge de déposer les données fournies par le centre O.I.E. et ses partenaires. Consultation des données : en libre accès.
Accès à l'entrepôt	Restreint (membres)	Libre
Accès aux données	Restreint (membres)	Libre
Dépôt de données	Restreint (membres)	Restreint
Modèle économique	Gratuit	Le dépôt de données est : gratuit pour les membres du centre O.I.E. ; payant pour les contributeurs externes. La consultation des données est libre et gratuite.
Utilisation (en 2016)	NC	1699 notices (le 18.02.16)
Utilisation (au 23/07/2019)	NC	1 690 notices

Nom	CoCoON	BASS2000
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	https://cocoon.huma-num.fr/	http://bass2000.obspm.fr/
Structure d'appartenance	LLL ; LACITO	LESIA
Date de création	NC	NC
Source de financement/Budget	NC	NC
Ressources humaines	NC	NC
Profils de postes	NC	NC
Description	CoCoON (Collections de Corpus Oraux Numériques) est une plateforme technique qui accompagne les producteurs de ressources orales à créer, structurer, partager et archiver leurs corpus (i.e. des enregistrements audio ou vidéo, éventuellement accompagnés d'annotations textuelles). La plateforme est hébergée par Huma-Num et gérée conjointement par le Laboratoire de Langues et Civilisations à Tradition Orale (UMR7107 LACITO) et le Laboratoire Ligérien de Linguistique (UMR7270 LLL). L'auteur et son institution peuvent bénéficier d'un accès restreint et sécurisé à leurs données, pendant une période d'embargo définie, si le contenu de l'information est considéré sensible. L'entrepôt CoCoON est régulièrement moissonné par des fournisseurs de services tels que : la plateforme Isidore ; Le Language Resource Catalog de l'organisation OLAC ; le Virtual Language Observatory de l'infrastructure européenne CLARIN.	BASS2000 archive et donne accès à des données d'observation du soleil, fournies par différents observatoires en France. Les données sont issues d'instruments basés au sol, de type spectrohéliographes, radiohéliographes, coronographes et cartes synoptiques. Disponibles en libre accès, elles sont destinées à être réutilisées dans un but exclusivement scientifique ou pédagogique.
Disciplines	Sciences Humaines et Sociales	Sciences & Technologies
Thématique	Linguistique	Sciences de l'Univers
Communauté d'utilisateurs	Communauté de recherche en sciences humaines et sociales	ESR français et étranger
Utilisateurs finaux	Tout public	Enseignants-chercheurs, Doctorants
Conditions d'usage	Dépôt de données : tout contributeur doit au préalable demander la création d'un compte utilisateur, en écrivant à l'adresse cocoon_web@huma-num.fr. Consultation des données : elle est libre et gratuite ; toutefois certaines ressources, jugées sensibles, ne sont rendues accessibles qu'après une période d'embargo prédéfinie.	La base de données et les données qu'elle contient sont en libre accès.
Accès à l'entrepôt	Libre	Libre
Accès aux données	Libre ; Restreint (membres) ; Embargo	Libre
Dépôt de données	Restreint (sélection + registration)	Fermé
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	NC
Utilisation (au 23/07/2019)	11 939 notices	NC

Nom	**EELS Data Base**	**Kinsources.net**
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	https://eelsdb.eu/	https://www.kinsources.net/
Structure d'appartenance	IMN	MAE René-Ginouvès
Date de création	NC	NC
Source de financement/Budget	NC	NC
Ressources humaines	NC	NC
Profils de postes	NC	NC
Description	EELS DataBase est un entrepôt en libre accès de spectres d'excitation. Les données sont issues d'expériences en spectroscopie des pertes d'énergie et en spectroscopie d'absorption des rayons X. Chaque spectre est accompagné de ses paramètres d'enregistrement, afin de rendre compte des conditions d'expérimentation le plus précisément possible. L'entrepôt est utilisé par des spectroscopistes, des théoriciens, des étudiants et des entreprises privées comme catalogue de référence des structures fines. Il est également ouvert au dépôt de nouvelles données. La base de données et les données spectroscopiques sont couvertes par l'Open Database License (ODbL).	Kinsources.net est une plateforme web dédiée à l'archivage, au partage, à l'analyse et à la comparaison des données de parenté (généalogiques, terminologiques, résidentielles et relationnelles), alimentée par les chercheurs des sciences humaines et sociales. Cette plateforme collaborative est issue d'un projet réunissant quatre laboratoires en anthropologie et histoire : le Laboratoire d'Ethnologie et Sociologie Comparative (Univ. Paris X), le Laboratoire d'Anthropologie Sociale (Collège de France), le Laboratoire de Démographie et d'Histoire Sociale (EHESS) et le Centre Roland Mousnier (Univ. Paris IV). Elle est hébergée par le TGIR Huma-Num. Kinsource.net propose un outil de référence pour l'archivage (entrepôt pérenne, métadonnées, documentation, références scientifiques, interopérabilité) et la publication des données en accès libre (anonymisation des données nominatives sensibles, attribution de licence ouverte, validation scientifique, génération d'un lien permanent).
Disciplines	Sciences & Technologies	Sciences Humaines & Sociales
Thématique	Physique de la matière condensée, Chimie de synthèse et matériaux	Généalogie, Anthropologie, Histoire, Démographie
Communauté d'utilisateurs	ESR français et étranger ; Entreprises privées	Communauté de recherche en sciences humaines et sociales
Utilisateurs finaux	Enseignants-chercheurs, Doctorants, Etudiants, Ingénieurs	Tout public voire accès restreint aux chercheurs et étudiants
Conditions d'usage	Dépôt de données : tout contributeur doit au préalable créer un compte utilisateur. Consultation : la base de données et les données qu'elle contient sont librement consultables.	Dépôt de données : tout contributeur doit demander la création d'un compte utilisateur et ses données sont soumises à la validation scientifique avant la publication. Consultation des données : elle est libre et gratuite
Accès à l'entrepôt	Libre	Libre
Accès aux données	Libre	Libre
Dépôt de données	Restreint (registration)	Restreint (sélection + registration)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	NC
Utilisation (au 23/07/2019)	276 jeux de données	127 jeux de données

Nom	AVISO+	VRP-REP
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	http://www.aviso.altimetry.fr/en/home.html	http://www.vrp-rep.org
Structure d'appartenance	CNES	Université Catholique de l'Ouest
Date de création	NC	NC
Source de financement/Budget	NC	NC
Ressources humaines	NC	NC
Profils de postes	NC	NC
Description	AVISO+ (Archiving, Validation and Interpretation of Satellite Oceanographic data) archive et distribue les données des missions spatiales d'altimétrie ainsi que les produits Doris d'orbitographie et de localisation précises. C'est le portail de référence de l'altimétrie qui donne accès à des données d'observation concernant plusieurs thématiques : l'océan, l'hydrologie et les terres émergées, les zones côtières, la glace et la cryosphère, le climat, l'atmosphère, la géodésie, la géophysique et la biologie marine. AVISO+ fournit des outils d'aide à l'utilisation des données altimétriques.	VRP-REP est un entrepôt de données librement accessibles, dédié aux problèmes de tournées de véhicules. Les utilisateurs peuvent consulter ou déposer des exemples de problèmes-types ainsi que leur(s) solution(s). Pour signaler un nouveau cas, il est préconisé d'utiliser le schéma XML établi par VRP-REP.
Disciplines	Sciences & Technologies	Sciences & Technologies
Thématique	Océanographie, Climatologie, Hydrologie, Glaciologie	Sciences informatiques, Ingénierie des systèmes et de la communication, Ingénierie des produits et des procédés
Communauté d'utilisateurs	Communauté des chercheurs travaillant sur le développement et les applications de l'altimétrie	Communauté scientifique ou professionnelle travaillant sur les problèmes de tournées de véhicules
Utilisateurs finaux	Chercheurs	Tout public
Conditions d'usage	Plusieurs voies pour accéder aux données : accès libre ; accès par authentification	Dépôt de données : sur demande, auprès du comité de pilotage de VRP-REP. Consultation des données : La plupart des données sont en accès et téléchargement libres. Certaines peuvent être en accès restreint, lorsque le déposant ne souhaite pas les rendre visibles de tous.
Accès à l'entrepôt	Libre	Libre
Accès aux données	Libre ; Restreint (registration)	Libre
Dépôt de données	Restreint (membres)	Restreint (sélection + registration)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	NC
Utilisation (au 23/07/2019)	NC	85 jeux de données

Nom	**MISTRALS database**	**Coriolis**
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	http://mistrals.sedoo.fr/	http://www.coriolis.eu.org/
Structure d'appartenance	MISTRALS	SISMER
Date de création	NC	NC
Source de financement/Budget	NC	NC
Ressources humaines	NC	NC
Profils de postes	NC	NC
Description	MISTRALS Database est l'entrepôt des jeux de données issues de certaines thématiques du programme international de recherche et d'observation MISTRALS (Mediterranean Integrated Studies at Regional And Local Scales, initié en 2008) dédié à la Méditerranée. Il donne accès aux données des projets CORSiCA, EMSO et MOOSE provenant des observatoires. MISTRALS database rend accessibles les données d'observation et de recherche aux décideurs, acteurs territoriaux et gestionnaires. Chaque thématique possède sa politique de données (obligations pour les fournisseurs et utilisateurs des données, métadonnées).	CORIOLIS contribue au programme français d'océanographie opérationnelle, en offrant un accès unifié aux observations relevées in situ par des bateaux ou des systèmes autonomes fixes ou dérivants. Ces données peuvent être utilisées pour visualiser la structure et l'intensité de la masse d'eau à un instant t. Elles permettent de mieux surveiller et comprendre le fonctionnement de l'océan, de ses écosystèmes et de son rôle sur le climat. Les données sont vérifiées puis diffusées en ligne moins de 24h après leur acquisition. CORIOLIS propose ainsi à tout nouveau projet d'océanographie opérationnelle de : mettre en forme ses données ; contrôler la qualité des données en temps réel ; diffuser les données à la communauté scientifique en océanographie opérationnelle ou à des utilisateurs individuels ; garantir la découverte et l'accès aux données.
Disciplines	Sciences Humaines et Sociales ; Sciences & Technologies ; Vie & Santé	Sciences & Technologies ; Vie & Santé
Thématique	Climat, Hydrologie, Biodiversité, Agriculture, Sciences sociales	Sciences de la Terre, Biologie environnementale, Ecosystème marin
Communauté d'utilisateurs	Communauté des chercheurs travaillant sur le changement climatique du bassin méditerranéen	Projets et infrastructures effectuant des mesures in situ de l'océan
Utilisateurs finaux	Tout public	Communauté scientifique en océanographie opérationnelle
Conditions d'usage	Déposants : participants aux programmes MISTRALS. Utilisateurs : accès libre ou restreint selon les programmes	Transmission de données : tout projet ou infrastructure souhaitant bénéficier des services de CORIOLIS doit en faire la demande par mail. Consultation des données : une partie des données est en accès libre voire en libre téléchargement.
Accès à l'entrepôt	Libre	Libre
Accès aux données	Restreint (registration)	Libre ; Embargo
Dépôt de données	Restreint (membres)	Restreint
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	NC
Utilisation (au 23/07/2019)	NC	12 bases de données

Nom	ESTHER database	GAPHYOR
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	Arrêté
URL	http://bioweb.ensam.inra.fr/ESTHER/general?what=index	http://gaphyor.lpgp.u-psud.fr/index-fr.html (lien mort)
Structure d'appartenance	Dynamique Musculaire et Métabolisme	Laboratoire de Physique des Gaz et des Plasmas
Date de création	NC	NC
Source de financement/Budget	NC	NC
Ressources humaines	NC	NC
Profils de postes	NC	NC
Description	La base de données intégrative ESTHER (ESTerases, alpha/beta-Hydrolase Enzymes and Relatives) recense les analyses de la super-famille structurale des protéines à repliement de type alpha/beta hydrolase. Pour nombre des membres de cette super-famille, de nombreuses données expérimentales sur les mécanismes catalytiques, les mutations naturelles ou induites, la structure tridimensionnelle, etc. ont été obtenues via des approches conceptuelles et méthodologiques extrêmement diverses, mais ces données sont souvent dispersées et difficiles à corréler. ESTHER a été créée pour répondre à ces difficultés. La base de données est en accès libre.	GAPHYOR (GAz-PHYsique-ORsay) est une base de données bibliographiques, factuelles et numériques sur les atomes, les molécules, les gaz et les plasmas, incluant des réactions chimiques. Elle est basée sur l'analyse des publications couvrant les domaines de la physique atomique et moléculaire, la chimie physique et la physique des plasmas. Des experts scientifiques ont codé et indexé les informations issues des publications sous une forme concise et structurée. Les domaines suivants sont représentés : propriétés des atomes et des molécules isolés, collisions avec des photons, collisions avec des électrons, collisions et réactions entre atomes et molécules, propriétés macroscopiques des gaz.
Disciplines	Vie & Santé	Sciences & Technologies
Thématique	Génétique, Bioinformatique	Physique atomique et moléculaire
Communauté d'utilisateurs	ESR français et étranger	Membres du Laboratoire de Physique des Gaz et des Plasmas
Utilisateurs finaux	Chercheurs en agroalimentaire et en pharmacie	Communauté scientifique effectuant des recherches sur les plasmas
Conditions d'usage	Utilisateurs : accès libre	Pas d'accès pour les utilisateurs extérieurs au Laboratoire de Physique des Gaz et des Plasmas
Accès à l'entrepôt	Libre	Restreint (membres)
Accès aux données	Libre ; Restreint	Restreint (membres)
Dépôt de données	Restreint	Restreint (membres)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	NC
Utilisation (au 23/07/2019)	NC	Lien mort

Nom	**CDS Portal**	**ILL Data Portal**
Type de service	Entrepôt de données	Entrepôt de données
Statut du service	En production	En production
URL	http://cdsportal.u-strasbg.fr/	https://www.ill.eu/users/ill-data-policy/
Structure d'appartenance	CDS	Institut Laue Langevin
Date de création	1972 (portail web créé dans les années 1990)	NC
Source de financement/Budget	Observatoire astronomique	NC
Ressources humaines	30 personnes	NC
Profils de postes	Chercheurs ; Informaticiens ; Documentalistes	NC
Description	Le CDS Portal donne accès aux produits élaborés par le Centre de Données astronomiques de Strasbourg (CDS). Il fonctionne comme un méta-moteur, qui permet de rechercher dans les bases de données suivantes : SIMBAD, base de données de référence pour les identifications et la bibliographie des objets astronomiques hors système solaire ; VizieR, base de données de référence pour les grands relevés du ciel, les catalogues et les tables publiées dans les journaux académiques ; Aladin, atlas interactif du ciel permettant d'accéder, de visualiser et d'analyser la collection d'images de référence du CDS, ainsi que les images disponibles dans les archives des observatoires spatiaux et au sol.	L'Institut Laue-Langevin (ILL) offre aux chercheurs des faisceaux neutroniques parmi les plus intenses. L'essentiel du temps d'utilisation des faisceaux neutroniques est consacré à l'étude des matériaux. En 2011, l'Institut Laue-Langevin a mis en place une politique des données avec l'attribution d'un identifiant pérenne de type DOI pour les expériences sélectionnées par un comité d'experts. ILL Data Portal a été créé en 2014 pour gérer, conserver, consulter et partager les données expérimentales des scientifiques utilisant les installations de l'institut (spectromètres). L'accès aux données est réservé aux membres ayant participé aux expériences.
Disciplines	Sciences & Technologies	Sciences & Technologies ; Vie & Santé
Thématique	Sciences de l'Univers	Science des matériaux, Biologie, Chimie, Physique des particules
Communauté d'utilisateurs	Membres du CDS	Communauté de recherche internationale utilisant la diffusion des neutrons pour l'étude des matériaux
Utilisateurs finaux	Communauté de recherche internationale en astronomie	ESR français et étranger
Conditions d'usage	Dépôt de données : ce sont les documentalistes du CDS qui alimentent les différentes bases de données. Consultation des données : les données diffusées par le CDS sont en libre accès.	Déposants : scientifiques utilisant les installations de l'Institut Laue Langevin. Utilisateurs : accès restreint, après authentification et demande auprès du déposant ; embargo de 3 à 5 ans.
Accès à l'entrepôt	Libre	Restreint (registration)
Accès aux données	Libre	Restreint (registration) ; Embargo ; Restreint
Dépôt de données	Restreint	Restreint (membres)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	30 587 catalogues dans VizieR (22.02.16) ; 800 000 requêtes par jour (hommes + machines)	NC
Utilisation (au 23/07/2019)	18 764 catalogues	NC

Nom	CAZy Database	BBEES : Annuaire de données
Type de service	Entrepôt de données	Annuaire de données
Statut du service	En production	En production
URL	http://www.cazy.org/	http://www.bdd-inee.cnrs.fr/
Structure d'appartenance	Architecture et Fonction des Macromolécules Biologiques	BBEES
Date de création	NC	2011
Source de financement/Budget	NC	UMS 3468
Ressources humaines	NC	3,5 ETP
Profils de postes	NC	Chercheur ; Ingénieurs développement ; Gestionnaires de BDD
Description	La base de données Carbohydrate-Active enZYmes Database (CAZy) a été créée en 1998 par l'équipe Glycogénomique du laboratoire Architecture et Fonction des Macromolécules Biologiques (AFMB). L'objectif est d'établir les relations entre la séquence des enzymes et leur spécificité. L'équipe Glycogénomique a développé une classification en familles qui relie la structure et le mécanisme catalytique des CAZymes. Un couplage entre la base de données CAZy et les moyens en bioinformatique permet d'examiner le contenu en CAZymes de centaines de génomes eucaryotes et procaryotes. La classification CAZy est largement utilisée par la communauté scientifique pour ses données annotées (annotations structurales et fonctionnelles des protéines issues de GenBank, UniProt, Protein Data Bank).	L'annuaire de base de données de l'UMS BBEES permet de signaler et de rechercher une base de données sur la biodiversité, portée par l'InEE-CNRS et/ou le Museum National d'Histoire Naturelle. Il est modéré par l'Unité Mixte de Service et est hébergé sur les serveurs du CNRS et du Centre de Calcul de l'IN2P3. La saisie d'une nouvelle base de données se fait par le biais d'un formulaire, qui est publié dans l'annuaire après avoir été validé par l'administrateur. Chaque base de données signalée est accompagnée des coordonnées de son responsable, donnant ainsi aux utilisateurs la possibilité d'entrer directement en contact avec cette personne.
Disciplines	Vie & Santé	Vie & Santé ; Sciences Humaines et Sociales
Thématique	Biologie, Génétique, Bioinformatique	Ecologie, Environnement, Société
Communauté d'utilisateurs	Chercheurs en biologie structurale, génie génétique et bioinformatique	Chercheurs de l'InEE-CNRS, Chercheurs du Museum National d'Histoire Naturelle
Utilisateurs finaux	ESR français et étranger	Tout public
Conditions d'usage	Déposants : membres de l'équipe Glycogénomique du laboratoire Architecture et Fonction des Macromolécules Biologiques. Utilisateurs : accès libre.	Signalement : par le biais d'un formulaire de soumission. Accès : libre
Accès à l'entrepôt	Libre	Libre
Accès aux données	Libre	-
Dépôt de données	Restreint	Restreint (sélection + registration)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	Environ 180 bases de données référencées (le 14.04.16)
Utilisation (au 23/07/2019)	NC	2177 notices

Nom	**Portail Epidémiologie France**	**IrsteaData***
Type de service	Annuaire de données	Annuaire de données
Statut du service	En production	En cours d'élaboration
URL	https://epidemiologie-france.aviesan.fr	https://forge.irstea.fr/projects/irsteadata
Structure d'appartenance	Alliance Aviesan	Irstea
Date de création	2011	2016
Source de financement/Budget	250 000 à 350 000€ par an	Irstea
Ressources humaines	3 ETP	30 personnes dans le groupe projet
Profils de postes	Ingénieurs de recherche	Professionnels de l'IST ; informaticiens ; administrateurs de BDD ; juriste ; chercheurs
Description	Le portail Épidémiologie-France propose un catalogue en ligne des principales bases de données individuelles en santé de source française en santé publique (hors essais cliniques). Ce portail s'inscrit dans un mouvement de partage et de réutilisation des données épidémiologiques. Chaque base de données répertoriée dans le catalogue est décrite selon ses caractéristiques essentielles : objectifs, thématiques, populations couvertes, nature des informations recueillies, conditions d'accès, responsable…	L'Irstea a initié un projet d'annuaire destiné à répertorier l'ensemble des bases de données produites ou coproduites par ses chercheurs. L'objectif est de fournir une vision exhaustive des données de recherche de l'institut. Cette initiative s'inscrit dans le cadre de la démarche Qualité engagée par la direction de l'institut, ainsi qu'en réponse à la Directive Inspire de la Commission européenne. L'annuaire sera d'abord testé par 3 centres pilotes de l'Irstea, avant d'être déployé à l'ensemble des centres régionaux.
Disciplines	Vie & Santé	Sciences & Technologies ; Vie & Santé
Thématique	Santé publique, Epidémiologie	Ecologie, Biodiversité
Communauté d'utilisateurs	Chercheurs en sciences de la vie et de la santé	Membres de l'Irstea
Utilisateurs finaux	Tout public	ESR français et étranger
Conditions d'usage	Accès aux bases de données sur demande	Référencement : tout membre de l'Irstea souhaitant signaler une base de données devra s'authentifier au préalable. Consultation : contenu de l'annuaire en libre accès (à l'exception des données de recherche déclarées confidentielles, qui seront accessibles uniquement aux membres de l'Irstea).
Accès à l'entrepôt	Libre	Libre
Accès aux données	-	-
Dépôt de données	Restreint (registration)	Restreint (membres)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	901 notices (le 22.03.16)	NC (en cours de création au moment de l'étude)
Utilisation (au 23/07/2019)	914 notices	34 jeux de données

Nom	Cirad : Annuaire de données*	InfoTerre
Type de service	Annuaire de données	Annuaire de données
Statut du service	En cours d'élaboration	En production
URL	https://datacatalog.cirad.fr	http://infoterre.brgm.fr/
Structure d'appartenance	Cirad	BRGM
Date de création	2015	NC
Source de financement/Budget	Cirad	BRGM
Ressources humaines	5 à 6 personnes à temps partiel	NC
Profils de postes	Informaticiens ; Professionnels de l'IST ; Chercheurs	NC
Description	L'annuaire des données scientifiques du Cirad est un projet d'établissement initié en 2015. Il a pour objectif de répertorier les données produites par les unités de recherche de l'institut. Il est accessible depuis l'intranet du Cirad. Les données sont décrites selon le standard Dublin Core. A ces informations génériques s'ajoutent des métadonnées plus spécifiques de la production scientifique et des nomenclatures du Cirad (coordonnées géographiques, formats de fichiers, mode de sauvegarde, niveau de confidentialité, etc.). Lors de la première vague d'alimentation du portail, des correspondants ont été désignés dans chaque unité de recherche, afin de recueillir la liste des données qui y étaient produites. A terme, les pratiques d'inventaire seront amenées à se généraliser, avec la création d'un formulaire web qui permettra aux chercheurs de référencer directement leurs jeux de données sur le portail.	InfoTerre est le visualiseur de données géoscientifiques du Bureau de Recherches Géologiques et Minières (BRGM). Il donne accès aux couches de la Banque du Sous-Sol (BSS), aux cartes géologiques, ainsi qu'aux autres données relatives aux thématiques du BRGM. Il fournit également un ensemble de services avancés permettant de valoriser les données affichées (recherche avancée, informations/métadonnées sur chaque couche, téléchargement des données affichées…). Afin d'adapter son offre aux attentes du grand public comme de ses utilisateurs experts, InfoTerre propose deux visualiseurs cartographiques : une version standard (qui permet d'accéder à l'intégralité des données et de les exploiter au travers d'outils élaborés) et une version simplifiée (qui donne accès à une sélection de données épurées). Depuis 2003, il a fait le choix de l'interopérabilité. Cela se traduit par le respect des normes d'interopérabilité : W3C pour les aspects Web, OGC pour les aspects diffusion et échange de données.
Disciplines	Sciences & Technologies ; Vie & Santé ; Sciences Humaines et Sociales	Sciences & Technologies
Thématique		Géologie, Géophysique, Ecologie
Communauté d'utilisateurs	Membres du Cirad	Membres du BRGM
Utilisateurs finaux	Membres du Cirad	Tout public
Conditions d'usage	Accès réservé aux membres du Cirad	Le portail est en consultation libre.
Accès à l'annuaire	Restreint (membres)	Libre
Accès aux données	-	-
Dépôt de données	Restreint (membres)	Restreint
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	280 notices (le 11.03.16)	NC
Utilisation (au 23/07/2019)	NC	1 564 027 banques de données

Nom	**IO Data Science**	**ECOSCOPE : Annuaire de données**
Type de service	Annuaire de données	Annuaire de données
Statut du service	En production	En production
URL	https://io.datascience-paris-saclay.fr/	http://ecoscope.fondationbiodiversite.fr/ecoscope-portal
Structure d'appartenance	Paris-Saclay Center for Data Science	ECOSCOPE
Date de création	2014	2016
Source de financement/Budget	IdEx	FRB (organismes de recherche)
Ressources humaines	1 ETP + personnels titulaires	3 ETP
Profils de postes	Ingénieur développement ; chercheurs	NC
Description	IO Data Science référence des données de recherche produites par les laboratoires de l'Université Paris-Saclay. Il permet de lier, de découvrir et de réutiliser des données, dans l'objectif de créer de nouvelles synergies entre laboratoires. Les chercheurs peuvent décrire des jeux de données et choisir de partager cette description en accès libre ou bien en accès restreint (c'est-à-dire aux seuls membres de l'Université Paris-Saclay). Ils peuvent aussi lier plusieurs descriptions entre elles ou insérer un lien vers le jeu de données.	Le portail de métadonnées d'ECOSCOPE a pour objectif de fournir à une large communauté utilisatrice un accès à des métadonnées et jeux de données d'observation pour la recherche sur la biodiversité, des services et des outils de visualisation et d'analyses des données et des produits. Les métadonnées décrivant finement les données de recherche et les ressources biologiques issues d'observatoires et d'initiatives de recherche sur la biodiversité sont basées sur le standard Ecological Metadata Language (EML), conforme aux requis de la directive INSPIRE. Ce portail des métadonnées favorise la réutilisation des données de dans d'autres projets de recherche et constitue un état des lieux de référence des données disponibles dans le domaine de la biodiversité.
Disciplines	Sciences & Technologies ; Vie & Santé ; Sciences Humaines & Sociales	Vie & Santé
Thématique		Ecologie, Biodiversité
Communauté d'utilisateurs	Membres de l'Université Paris Saclay	ESR français, Associations de gestionnaires d'espaces et d'espèces, Décideurs publics, Entreprises privées
Utilisateurs finaux	ESR français et étranger	Tout public
Conditions d'usage	Référencement : tout membre de l'Université Paris Saclay peut décrire un jeu de données dans l'annuaire. Consultation : les descriptions diffusées en libre accès sont consultables par tous ; les descriptions diffusées en accès restreint ne sont visibles que des membres de l'Université Paris-Saclay (via leurs identifiants de connexion institutionnels).	Le portail de métadonnées est librement accessible en ligne
Accès à l'annuaire	Libre	Libre
Accès aux données	-	-
Dépôt de données	Restreint (membres)	Restreint (membres)
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	70 fiches décrivant des jeux de données (le 06.03.2017)
Utilisation (au 23/07/2019)	59 notices	53 notices

Nom	ArchiPolis*	Réseau Quetelet
Type de service	Annuaire de données	Annuaire de données
Statut du service	En production	En production
URL	https://catalogues.cdsp.sciences-po.fr/dataverse/archipolis	http://www.reseau-quetelet.cnrs.fr/
Structure d'appartenance	CDSP	PROGEDO
Date de création	2016	2014
Source de financement/Budget	Financement sur 4 ans en tant que consortium d'Huma-Num	UMS 3558
Ressources humaines	4 personnes + correspondants dans les unités de recherche	4 personnes
Profils de postes	Ingénieurs de recherche ; archivistes ; professionnels de l'IST ; chercheurs	Ingénieurs chargés des aspects juridiques et financiers ; de la communication ; de la coopération internationale ; Ingénieur développement
Description	Le consortium ArchiPolis rassemble 8 unités de recherche en sciences sociales du politique (CDSP, CEE, CEVIPOF, OSC, CSO, CERAPS, PACTE, Triangle). Archipolis a été labellisé par la TGIR Huma Num entre 2012 et 2016. L'objectif du Consortium est de proposer un outil commun d'inventaire des enquêtes qualitatives en sciences sociales, accompagné des méthodes et bonnes pratiques d'inventaire et d'archivage et d'une sensibilisation des chercheurs à ces démarches. Il s'agit de rendre ces enquêtes intelligibles grâce à une documentation et une mise en contexte conséquentes. Le travail du consortium est mené en collaboration avec les réseaux d'archivistes des universités et établissements d'enseignement supérieur et de recherche.	Le Réseau Quetelet est le portail français d'accès aux données pour les sciences humaines et sociales. Il permet aux chercheurs français et étrangers de commander des données extraites : de grandes enquêtes, de recensements et d'autres bases de données issues de la statistique publique française ; de grandes enquêtes françaises provenant de la recherche. Ces enquêtes sont mises à disposition par quatre unités partenaires du réseau Quetelet : l'ADISP du Centre Maurice Halbwachs ; le CDSP de Sciences Po ; le service des enquêtes de l'INED ; le CASD. La diffusion des données s'effectue dans le cadre d'un règlement d'ensemble, adapté par chacun des partenaires en fonction des spécificités des fichiers qu'ils diffusent.
Disciplines	Sciences Humaines & Sociales	Sciences Humaines & Sociales
Thématique	Sciences politiques	
Communauté d'utilisateurs	Chercheurs du consortium ArchiPolis	CASD, INED, CDSP, CMH-ADISP
Utilisateurs finaux	Chercheurs ESR	ESR français et étranger
Conditions d'usage	Alimentation du catalogue : ajout de notices réservé aux unités de recherche membres du consortium Archipolis Consultation : le catalogue est librement accessible en ligne	Tous les fichiers diffusés dans le cadre du Réseau Quetelet sont d'accès gratuit pour une utilisation dans une finalité de recherche. Les demandes de données se font sur le site du réseau Quetelet. Elles sont ensuite traitées par les fournisseurs de données.
Accès à l'annuaire	Libre	Libre
Accès aux données	-	-
Dépôt de données	Restreint (membres)	Restreint
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	222 notices (le 11.10.2016)	1274 notices (le 22.04.16)
Utilisation (au 23/07/2019)	228 notices	1 518 notices

Nom	**DMP OPIDoR***	**Agence DataCite***
Type de service	Outil de gestion de données	Outil de gestion de données
Statut du service	En cours d'élaboration	En production
URL	https://dmp.opidor.fr/	http://www.inist.fr/?Attribution-de-DOI&lang=fr
Structure d'appartenance	Inist-CNRS	Inist-CNRS
Date de création	2016	NC
Source de financement/Budget	CNRS	NC
Ressources humaines	NC	NC
Profils de postes	Professionnels de l'IST ; ingénieurs développement	NC
Description	DMP OPIDoR est un outil d'aide à la rédaction de plans de gestion de données. L'utilisateur a la possibilité de choisir le modèle de plan de gestion de données qui correspond à son institution d'affiliation ou au financeur de son projet de recherche. A défaut, il peut aussi utiliser le modèle proposé par la Commission européenne dans le programme de financement Horizon 2020. Chaque modèle est accompagné de recommandations, aidant l'utilisateur à répondre aux questions. Les institutions et communautés de recherche peuvent ajouter dans DMP OPIDoR leur modèle de plan de gestion de données, ainsi que leurs propres recommandations. DMP OPIDoR a été développé à partir du code source de DMPonline, son homologue britannique.	En tant que membre du consortium DataCite, l'Inist-CNRS propose un service d'attribution de DOI (Digital Object Identifier) aux objets issus de la recherche. Ce service assure l'établissement de contrats, l'attribution de préfixes et l'accompagnement pour la production de métadonnées DataCite. Le DOI (Digital Object Identifier) joue un rôle clé en assurant l'accès à long terme aux objets scientifiques comme les données de la recherche, les images, les vidéos etc.
Disciplines	Sciences & Technologies ; Vie & Santé ; Sciences Humaines & Sociales	Sciences & Technologies ; Vie & Santé ; Sciences Humaines & Sociales
Thématique		
Communauté d'utilisateurs	ESR français	ESR français
Utilisateurs finaux	Chercheurs, Professionnels IST	Producteurs/Gestionnaires de données
Conditions d'usage	L'utilisation de DMP OPIDoR est soumise à la création d'un compte utilisateur.	
Modèle économique	Gratuit	Payant
Utilisation (en 2016)	NC (en cours de création au moment de l'étude)	NC
Utilisation (au 23/07/2019)	16 modèles de DMP	NC

Nom	Equipe Valorisation des données de la r	USPC : Accompagnement*
Type de service	Accompagnement	Accompagnement
Statut du service	En production	En production
URL	http://www.inist.fr/?Donnees-de-la-recherche	https://appui-recherche.univ-paris-diderot.fr/gerer-ses-donnees-scientifiques-le-plan-de-gestion-de-donnees
Structure d'appartenance	Inist-CNRS	USPC
Date de création	2012	2015
Source de financement/Budget	CNRS	Pas de budget dédié
Ressources humaines	10 personnes	1 ETP
Profils de postes	Professionnels de l'IST	Professionnel de l'IST ; Archiviste ; Personnel administratif
Description	L'équipe Valorisation des données de la recherche met ses connaissances et ses compétences au service des chercheurs, laboratoires, groupements ou équipes de recherche de l'ESR, en proposant un accompagnement personnalisé à la gestion et à la valorisation des données de recherche. Cet accompagnement est réalisé soit à distance (visio- ou audio-conférence), soit en présentiel (d'une durée de quelques jours, réitérées si nécessaire).	Le Service Commun de la Documentation de l'Université Paris Descartes, le Bureau des archives ainsi que la Direction d'Appui à la Recherche et à l'Innovation de l'Université Paris Diderot se sont associés pour accompagner les chercheurs de la COMUE Université Sorbonne Paris Cité (USPC) dans la rédaction de plans de gestion de données (DMP). Un modèle de DMP a été conçu, en adéquation avec celui proposé par la Commission européenne dans le cadre d'Horizon 2020. L'objectif est d'offrir une liste de champs applicables, mais également d'identifier au sein de l'USPC les différents acteurs susceptibles d'accompagner les chercheurs dans la rédaction de leur DMP. Aussi le modèle de DMP proposé s'accompagne-t-il d'un workflow pointant, pour chaque section du plan de gestion de données, vers des personnes ressources.
Disciplines	Sciences & Technologies ; Vie & Santé ; Sciences Humaines et Sociales	Sciences & Technologies ; Vie & Santé ; Sciences Humaines et Sociales
Thématique		
Communauté d'utilisateurs	ESR français	Membres de l'Université Sorbonne Paris Cité
Utilisateurs finaux	Chercheurs ; Enseignants-chercheurs ; Doctorants	Chercheurs, Enseignants-chercheurs, Doctorants
Conditions d'usage	Sur demande	Sur demande Le modèle de DMP est librement réutilisable, dans le respect de la licence CC BY-NC-SA.
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	Accompagnement de 3 projets (le 04.06.2015)	Aide à la rédaction de DMP d'un projet de recherche H2020
Utilisation (au 23/07/2019)	NC	NC

Nom	**Université de Strasbourg : Accompagnement***	**DoRANum***
Type de service	Accompagnement	Formation
Statut du service	En production	En test
URL	https://bu.unistra.fr/.do?sysb=polo&cdArticle=SERVICES_DONNEES#0	http://www.doranum.fr/
Structure d'appartenance	Université de Strasbourg	Inist-CNRS
Date de création	2015	NC
Source de financement/Budget	IdEx	NC
Ressources humaines	1 ETP	NC
Profils de postes	Professionnel de l'IST	NC
Description	Ce service, encore expérimental, est conçu comme une interface entre les chercheurs, le Service Commun de la Documentation et la Direction informatique. Il aide notamment les chercheurs engagés dans des projets Horizon 2020 à assurer le libre accès à leurs données de recherche et propose les actions suivantes : accompagnement des projets de création de bases de données et d'outils de visualisation des données en ligne ; accompagnement dans la réalisation d'un plan de gestion des données ; conseil technique en matière de curation des données, de standards de métadonnées, de formats de fichiers, de choix d'entrepôts de données en ligne… ; conseil sur les aspects juridiques liés aux données de la recherche.	Le projet Données de la Recherche : Apprentissage NUMérique à la gestion et au partage a pour objectif de mettre en place un dispositif de formation à distance, intégrant différentes ressources d'auto-formation sur la thématique de la gestion et du partage des données de la recherche. Existantes ou créées dans le cadre du projet, ces ressources proposeront plusieurs parcours et modes d'apprentissage, à destination des enseignants-chercheurs, des doctorants et des professionnels de l'information. Le projet fait partie des actions de la Bibliothèque scientifique numérique (BSN). Il associe principalement le réseau national des URFIST et l'Inist-CNRS.
Disciplines	Sciences & Technologies ; Vie & Santé ; Sciences Humaines & Sociales	Sciences & Technologies ; Vie & Santé ; Sciences Humaines et Sociales
Thématique		
Communauté d'utilisateurs	Membres de l'Université de Strasbourg	ESR français
Utilisateurs finaux	Chercheurs, Enseignants-chercheurs	Chercheurs, Enseignants-chercheurs, Doctorants, Professionnels de l'IST
Conditions d'usage	Sur demande	Les ressources sont en accès libre. Elles sont réutilisables sous les conditions de la licence Creative Commons BY-NC-SA.
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	Accompagnement de 3 projets de valorisation en SHS	NC
Utilisation (au 23/07/2019)	NC	NC

Nom	Site Données de la Recherche*	Datapartage*
Type de service	Information	Information
Statut du service	En production	En cours d'élaboration
URL	http://www.donneesdelarecherche.fr/	https://www6.inra.fr/datapartage/
Structure d'appartenance	Inist-CNRS	INRA
Date de création	2012	2015
Source de financement/Budget	CNRS	100 000€ (à la création)
Ressources humaines	1 personne	Equipe projet de 15 personnes
Profils de postes	Professionnel de l'IST	Chercheurs ; Professionnels de l'IST ; Informaticiens
Description	Le site web Données de la Recherche est une plateforme d'information et de veille sur les données de la recherche. Les informations sont classées en six rubriques : Événements ; Normes ; Formats et protocoles ; Politiques et textes de référence ; Projets et initiatives ; Métiers et compétences ; Webographie. Il est possible de s'abonner au fil RSS du site.	Le site web Datapartage est une plateforme d'information proposée par l'INRA pour la gestion et le partage des données. Elle recense les services, outils et bonnes pratiques recommandés par l'institut, sous les 5 rubriques suivantes : Gérer ; Partager/Publier ; Réutiliser ; Technologies ; Documents de référence. Les services proposés sont : attribuer un DOI à un jeu de données INRA ; choisir un entrepôt pour déposer ses données ; écrire un plan de gestion de données ; citer des données.
Disciplines	Sciences & Technologies ; Vie & Santé ; Sciences Humaines et Sociales	Sciences & Technologies ; Vie & Santé
Thématique		
Communauté d'utilisateurs	ESR français	ESR français
Utilisateurs finaux	Professionnels de l'IST ; Chercheurs ; Enseignants-chercheurs ; Doctorants	Chercheurs ; Professionnels de l'IST ; Informaticiens
Conditions d'usage	Le site web est en consultation libre.	Le site web est en consultation libre.
Modèle économique	Gratuit	Gratuit
Utilisation (en 2016)	NC	NC (en cours de création au moment de l'étude)
Utilisation (au 23/07/2019)		

Nom	**Cirad : Information***	**CINES : Plateforme d'archivage***
Type de service	Information	Plateforme d'archivage
Statut du service	En production	En production
URL	http://coop-ist.cirad.fr/gestion-de-l-information/gestion-des-donnees-de-la-recherche	https://www.cines.fr/archivage/
Structure d'appartenance	Cirad	CINES
Date de création	2014	2014
Source de financement/Budget	Pas de budget dédié	Ministère de la Recherche
Ressources humaines	NC	13 ETP
Profils de postes	Professionnel de l'IST ; Chercheur	Archivistes ; Informaticiens
Description	Le site internet CoopIST propose des fiches didactiques, classées en rubriques thématiques, pour gérer, publier et évaluer l'information scientifique et technique. Une des rubriques est consacrée au thème de la gestion des données de recherche, avec des fiches didactiques telles que : - Découvrir des plans de gestion de données de la recherche - S'initier en ligne aux données de la recherche - Rendre publics ses jeux de données - Découvrir de nouveaux métiers liés aux données de la recherche - Rédiger et publier un data paper dans une revue scientifique - Citer un jeu de données scientifiques Les fiches sont conçues par la Délégation à l'Information Scientifique et Technique du Cirad et relues par un rédacteur scientifique ainsi qu'un membre du comité éditorial du CoopIST avant d'être publiées. Elles sont diffusées en libre accès et peuvent être exportées individuellement au format PDF.	La plateforme d'archivage du CINES a vocation à archiver les données et les documents numériques produits par la communauté française de l'Enseignement Supérieur et de la Recherche. Elle propose des solutions d'archivage numérique payantes, sur le moyen et le long terme, et offre à ses utilisateurs une expertise dans les domaines informatique et archivistique. La sécurité et l'intégrité des données sont garanties par un ensemble de procédures telles que l'attribution de métadonnées, le choix de formats de fichiers pérennes, la réplication des données et un environnement informatique protégé.
Disciplines	Sciences & Technologies ; Vie & Santé ; Sciences Humaines et Sociales	Sciences physiques et Ingénierie ; Sciences de la vie ; Sciences humaines et sociales
Thématique		
Communauté d'utilisateurs	ESR français	ESR français
Utilisateurs finaux	Chercheurs ; Professionnels de l'IST	Chercheurs
Conditions d'usage	Les fiches didactiques sont en libre consultation. Leur réutilisation est soumise à la licence Creative Commons BY-NC-SA 4.0.	La sélection des demandes est à l'appréciation du directeur et du conseil d'administration du Département Archivage et Diffusion. Les projets à volumétrie importante sont privilégiés, pour des raisons de rentabilité.
Modèle économique	Gratuit	Payant
Utilisation (en 2016)	5 000 visites (entre la création du site et mars 2016)	NC (difficilement mesurable puisque le CINES a vocation à archiver non seulement les données scientifiques, mais aussi les données patrimoniales et administratives)
Utilisation (au 23/07/2019)		

Annexe 9 - Formulaire de consentement éclairé

Formulaire de consentement

Présentation du doctorant

Cette recherche est réalisée dans le cadre du projet de doctorat de Violaine REBOUILLAT, dirigé par Ghislaine CHARTRON, du laboratoire Dicen-IDF au Conservatoire National des Arts et Métiers, et co-encadré par Joachim SCHÖPFEL, du laboratoire GERiiCO de l'Université Lille 3.

Avant d'accepter de participer à ce projet de recherche, veuillez prendre le temps de lire et de comprendre les renseignements qui suivent.

Nature et objectifs du projet

Mon projet de doctorat s'intitule « Open Research Data : comment naît une culture des données ? Enjeux, pratiques et compétences associés dans le contexte de la recherche française ». Il a pour but d'étudier la place des données scientifiques dans la pratique professionnelle des chercheurs du secteur académique français.

Déroulement du projet

Votre participation à cette recherche consiste à répondre à des questions que je vous poserai, dans le cadre d'un entretien semi-directif d'une durée d'environ 45 minutes. Cet entretien sera enregistré, de manière à ce que je puisse retranscrire vos propos le plus fidèlement possible.

Participation volontaire et droit de retrait

Vous êtes libre de participer à ce projet de recherche. Vous pouvez aussi mettre fin à votre participation sans conséquence négative ou préjudice et sans avoir à justifier votre décision. Si vous décidez de mettre fin à votre participation, il est important de m'en prévenir, en me contactant par mail (mes coordonnées figurent au bas de ce document). Tout le matériel permettant de vous identifier (incluant l'enregistrement de l'entrevue et les données que vous aurez fournies) sera alors détruit.

Confidentialité

Afin de garantir la confidentialité des données des participants, voici les mesures qui seront appliquées dans le cadre de la présente recherche :

- Votre nom et tous ceux cités durant l'entrevue seront remplacés par un code ;
- Moi seule aurai accès à la liste contenant les noms et les codes, elle-même conservée séparément du matériel de la recherche, des données et des formulaires de consentement ;
- Les noms des participants n'apparaîtront dans aucun rapport ni publication ;
- Les résultats seront présentés sous forme globale, de sorte que les résultats individuels des participants ne seront jamais communiqués ;
- Tout le matériel et toutes les données seront utilisés dans le cadre exclusif de cette recherche.

Remerciements

Votre collaboration est précieuse pour me permettre de réaliser cette étude. C'est pourquoi je tiens à vous remercier pour le temps et l'attention que vous acceptez d'y consacrer.

Signatures

Je soussigné(e) _____ déclare avoir pris connaissance du présent formulaire et consens librement à participer au projet de recherche sus-décrit.

_____	_____	_____
Signature du/de la participant(e)	Le	À

Je soussigné(e) _____ m'engage à respecter les termes du présent formulaire.

_____	_____	_____
Signature de la doctorante	Le	À

Personnes-ressources

Si vous avez des questions sur les aspects scientifiques du projet de recherche, vous pouvez contacter :

Violaine REBOUILLAT, doctorante (Dicen-IDF, Cnam) : violaine.rebouillat.auditeur@lecnam.net
Ghislaine CHARTRON, professeur (Dicen-IDF, Cnam) : ghislaine.chartron@lecnam.net
Joachim SCHÖPFEL, maître de conférences (GERiiCO, Université de Lille) : joachim.schopfel@univ-lille3.fr

Pour toute préoccupation sur vos droits ou sur les responsabilités des chercheurs concernant votre participation à ce projet, vous pouvez contacter le Correspondant Informatique et Libertés du Cnam :

Adresse mail : cil@cnam.fr
Adresse postale : Conservatoire National des Arts et Métiers
Direction des affaires générales
Service des affaires juridiques
Case 4DGS02S
292 rue Saint Martin
75003 PARIS

Annexe 10 - Guide d'entretien (1er panel, novembre 2017 – octobre 2018)

Introduction

→ Présentation de l'enquête
→ Objectif de l'entretien
→ Signature du formulaire de consentement éclairé

Contexte

→ Objectifs scientifiques du projet de recherche
→ Approche et méthodologie scientifiques
→ Partenaires du projet de recherche (internes/externes au laboratoire, disciplines scientifiques, statuts professionnels) ; répartition des rôles entre les partenaires

Utilisation des données de recherche

→ Utilisation du terme de « données » (oui/non, exemples, définition)
→ Mode(s) de collecte et de traitement des données
→ Documentation des procédures d'acquisition et de traitement (oui/non, lesquelles)
→ Circulation des données entre les partenaires du projet
→ Choix de conservation des données
→ Valeur des données

Ouverture des données de recherche

→ Incitations à l'ouverture (oui/non, lesquelles)
→ Opinion sur le mouvement d'ouverture des données scientifiques
→ Pratique personnelle
→ Conditions préalables à l'ouverture (enjeux éthiques/économiques...)

Annexe 11 - Guide d'entretien (2nd panel, mars – mai 2019)

Introduction

→ Présentation de l'enquête

→ Objectif de l'entretien

→ Signature du formulaire de consentement éclairé

Contexte

→ Thématique de recherche

→ Approche et méthodologie scientifiques

→ Utilisation du terme de « données » (oui/non, exemples, définition)

Utilisation de services de données (utiliser aussi des synonymes : outils, plateformes…)

→ Oui/non

→ Description du/des service(s) utilisé(s)

→ Mode d'accès à ce(s) service(s)

→ Utilité de ce(s) service(s)

Cat OPIDoR

→ Avis sur le catalogue (ergonomie, contenu, usage)

→ Connaissance des services répertoriés

Conclusion

→ Besoins en termes de services

→ Opinion sur le mouvement d'ouverture des données scientifiques

Annexe 12 – Exemples d'entretiens retranscrits

Seuls les entretiens pour lesquels les enquêtés ont donné leur accord de diffusion sont reproduits ici. Les données personnelles qu'ils contiennent ont été pseudonymisées.

La retranscription a été effectuée à l'aide du logiciel Sonal[317].

Entretien avec le chercheur 2 (chimie)

1 - 00:00 > 05:54 Objectifs scientifiques du projet

[>R1]: Est-ce que vous pourriez me résumer les objectifs du projet de recherche B. ?

[>R2]: L'objectif du projet est d'essayer de trouver des solutions pour traiter les douleurs chroniques neuropathiques, qui sont un groupe de maladies pour lequel il n'y a pas de traitement à l'heure actuelle. « Douleurs chroniques », ça veut dire des douleurs qui persistent sur des années et des années. Et « neuropathiques », parce que ce groupe de maladies est la conséquence d'une atteinte nerveuse. Cette atteinte nerveuse peut soit être causée par un accident/un traumatisme (un nerf est coincé ou coupé, suite à un accident de la route ou à une opération chirurgicale – les chirurgiens vous découpent et souvent ça peut laisser des petites traces), soit être la conséquence de maladies comme le sida, de maladies virales ou de diabète. Dans le cas d'atteintes nerveuses comme ça, une douleur issue de la lésion nerveuse va s'installer et persister dans le temps. Quand je dis « persister », ça peut être pendant des années et des années. Les sensations de membre fantôme sont un exemple assez classique : une personne ayant subi une amputation va continuer à sentir son membre amputé ; parfois c'est dans le même membre, parfois c'est dans le membre opposé. Enfin, c'est vraiment très compliqué. Ce sont des douleurs qui, au départ, sont des douleurs normales, et qui, après quelques mois, vont se présenter sous forme d'hyperréactivité à des sensations qui, normalement, ne devraient plus être douloureuses. Le patient va éprouver des picotements ou des sensations de brûlures ; et ceci en permanence. Il y a par exemple des gens pour qui le

317 https://sonal.hypotheses.org/

moindre contact avec le moindre vêtement va être insupportable. Ces douleurs chroniques touchent à peu près 15% de la population. Ça peut aller du banal mal de dos, qu'on ne sait pas diagnostiquer et qui traîne, à ce que l'on appelle des douleurs issues d'accidents de la route ou de suites post-opératoires. Or il n'y a pas de traitement pour ces maladies. Vous avez un mal de tête, vous prenez de l'aspirine ; quand c'est vraiment très aigu, vous prenez de la morphine. Les traitements sont symptomatiques : c'est-à-dire qu'on va uniquement traiter les symptômes et pas la cause. En réalité, ces traitements ne sont pas adaptés et marchent très, très mal, puisque les patients continuent à avoir mal. On estime qu'environ deux patients sur trois sont mal traités. Ces douleurs chroniques neuropathiques ont donc un coût socio-économique qui est très important : un coût en termes d'assurance maladie, (parce que les gens sont en arrêt maladie prolongé) et un coût sociétal (souvent ça conduit soit à du travail à domicile, soit à une perte d'emploi). Le coût global est très élevé : on estime que, pour les cinq pays industrialisés de l'Europe, plus les États-Unis, plus le Japon, il est de l'ordre de 200 milliards de dollars par an. C'est un coût qui est énorme, si vous cumulez les traitements médicaux, les hospitalisations, les arrêts maladie, les invalidités...

2 - 05:54 > 14:02 Financement du projet

[>R2]: C'est un projet qu'on a mené en collaboration avec un laboratoire de biologie, qui est à Montpellier.

[>R1]: C'est ce laboratoire qui fait partie de l'INSERM ?

[>R1]: Oui, tout à fait. C'est l'Institut des Neurosciences de Montpellier. A l'époque (le projet a démarré en 2011-2012), ce laboratoire avait une piste pour identifier le mécanisme à l'origine du déclenchement des douleurs chroniques neuropathiques et de leur maintenance dans le temps. Nous, on est un labo de chimie, on va dire. Donc la question, pour nous, était d'essayer de trouver une petite molécule qu'on pourrait administrer sous forme de comprimé et qui serait capable de stopper ce mécanisme. A ce moment-là, c'était la première fois qu'on essayait de s'attaquer aux causes de la maladie, et pas simplement aux symptômes. Ça c'est l'origine du projet. Et, comme souvent dans des projets de recherche, on démarre avec une idée mais pas de fonds. Donc, au tout début, on a travaillé avec très, très peu de fonds. On a obtenu un financement d'INSERM Transfert. « INSERM Transfert » c'est une cellule de

l'INSERM, qui est là disons pour donner des fonds d'amorçage (vraiment très peu d'argent), afin de développer une idée. Notre mission dans le cadre de ce projet préliminaire a été d'essayer de trouver les premières molécules (même si elles étaient imparfaites) capables de stopper ce mécanisme – sachant qu'on avait déjà identifié la protéine dont l'activation était responsable du mécanisme. Il fallait donc trouver des inhibiteurs de cette protéine. Assez rapidement, on a identifié des molécules qui étaient capables de l'inhiber. Elles étaient certes imparfaites, elles n'étaient pas très puissantes, pas très actives, mais elles étaient capables d'inhiber la protéine. A partir de là, on a soumis (comme tous les chercheurs) des demandes de financement. On a eu la chance d'avoir deux financements consécutifs. Le premier de la SATT de Montpellier. Les SATT sont des Sociétés Accélératrices de Transfert et de Technologie. Ce sont des sociétés de droit privé, financées par le CNRS, l'INSERM, l'Université... Il y a une SATT par région. Ces sociétés sont là pour valoriser les travaux académiques des laboratoires de la région. Nous, on a eu la chance d'avoir une aide de la SATT de Montpellier, qui nous a donné des financements, sous la forme d'un projet, qu'on appelle « projet de maturation », et qui nous a permis de commencer à travailler et de générer les premiers résultats. Ensuite on a eu un financement de l'Agence Nationale de la Recherche (ANR). Ce sont ces deux financements qui nous ont aidés à faire les découvertes académiques qui, par la suite, ont été transmises à une société que nous avons créée. Mais, au départ, ce sont vraiment ces deux sources de financement qui nous ont aidés.

[>R1]: Quelle était la durée du projet financé par la SATT ?

[>R2]: Le premier financement de la SATT était de 18 mois. Donc ça n'est vraiment pas beaucoup. C'était plutôt un projet pour faire ce qu'on appelle des « preuves de maturation », donc pour essayer de consolider la preuve de concept et avoir déjà des résultats préliminaires. Ensuite, on a eu un financement de l'ANR qui, lui, est de 4 ans et qui donc se termine l'an prochain. Au fur et à mesure de ces résultats, on a vu qu'on avait des molécules de plus en plus puissantes, de plus en plus sélectives. Et, comme il y a un potentiel de valorisation très important, on a décidé de créer une start-up pour valoriser ces résultats. Puisque les financements académiques/publics permettent de faire du travail de laboratoire, mais pas des travaux beaucoup plus coûteux comme les développements réglementaires en termes de préclinique/de clinique. Ça ce sont des choses qui requièrent des budgets qui ne sont plus atteignables dans la sphère académique. Il faut soit trouver un partenaire privé qui va nous

aider à développer ou co-développer le projet, soit il faut carrément créer une start-up, dont le but sera de lever des fonds pour développer ce projet et le valoriser.

[>R1]: Comment levez-vous des fonds avec cette start-up ?

[>R2]: Nous, les scientifiques, on est à l'origine de la start-up ; mais, parce qu'on est à la fois juge et partie, il y a une loi qui nous donne le droit d'être actionnaires de la société mais qui nous empêche d'être dans l'organigramme de la société (c'est-à-dire qu'on peut être consultants). Donc cette start-up a son personnel qui lui est propre. Elle finance du personnel dans nos laboratoires. Et il y a un président de la start-up, qui est responsable de développer sa société, à laquelle on est associés (puisque nous sommes actionnaires évidemment). Donc, là, il y a tout un tas de fonds, qui sont des fonds privés cette fois-ci. Ça peut quand même être des fonds publics, notamment des fonds régionaux. Mais, pour la plupart, ce sont des fonds privés, provenant par exemple de la Banque Publique d'Investissement (BPI) qui finance les start-up et les aide à se développer. Maintenant on commence à avoir des contacts avec des industries pharmaceutiques, qui seraient intéressées pour co-développer le projet avec nous.

[>R1]: Donc, pour l'instant, le but de la start-up c'est de faire des tests cliniques ?

[>R2]: En fait, pour développer un médicament, il y a plusieurs étapes. Il y a d'abord l'étape préclinique, c'est-à-dire l'étape où la molécule n'est pas encore administrée à l'homme (elle est administrée à des animaux). Bon, ça on l'a déjà. Mais il y a des étapes qui sont réglementaires ; il y a un cahier des charges très précis, on ne peut pas faire ce qu'on veut (ça n'est plus du travail de laboratoire). Ensuite, il y a les développements cliniques qui, eux, sont très élevés. Ils sont d'autant plus élevés que vous approchez du but final, à savoir la mise sur le marché. Sachant que développer un médicament coûte en moyenne deux milliards d'euros. Évidemment, il n'y a pas de fonds publics pour faire ça. Généralement, avec les financements d'une start-up, on peut aller jusqu'à administrer la molécule à un individu sain (ça c'est vraiment la première phase, pour montrer que la molécule n'est pas toxique, qu'elle ne tue pas les gens). Ensuite, vous administrez la molécule à des patients malades dans des phases en escalade avec une dizaine, une centaine, des milliers de patients. Là, ça devient évidemment très, très, très compliqué. Une petite société n'a quasiment plus les moyens de le faire. Donc, généralement, soit on s'associe avec un gros laboratoire, soit - ce qui arrive souvent - le gros laboratoire rachète la petite société pour développer la molécule qui l'intéresse.

3 - 14:02 > 17:19 Mode d'acquisition des données

[>R1]: Quel est l'objectif du financement ANR ?

[>R2]: L'objectif du financement ANR était de faire la preuve de concept que, en trouvant une petite molécule capable d'inhiber cette protéine, quand on l'administre à l'animal et qu'on génère chez celui-ci des douleurs neuropathiques (c'est un modèle animal qui mime la maladie chez l'homme)...

[>R1]: Vous lui faites une lésion ?

[>R2]: Oui, ce qu'on fait - ça n'est pas très sympa - c'est qu'on fait des ligatures du nerf sciatique. Des ligatures du nerf sciatique, ça fait mal. Quand vous faites une ligature du nerf sciatique à un animal, celui-ci va mettre à peu près une quinzaine de jours à développer une douleur neuropathique permanente. L'animal devient alors hyperactif à n'importe quel stimulus douloureux. Pour mesurer l'effet de ce genre de molécule, on va lui presser la patte. Tout ça, c'est calibré, c'est-à-dire qu'on va lui enfoncer des filaments dans la patte (des petits filaments avec une pression qui est calibrée). A partir du moment où l'animal a mal, il va lever la patte. Soit on lui applique des pressions de plus en plus importantes et on regarde à partir de quelle pression il lève la patte. Soit on lui applique une pression constante et on mesure le pourcentage d'animaux qui lèvent ou non la patte. Grosso modo, hein. Si l'animal est neuropathique, le moindre contact lui fait lever la patte. Quand vous administrez la molécule à l'animal (une des molécules qu'on a développées, parce qu'on en a plusieurs), pendant deux jours vous pouvez lui presser la patte, il n'a plus mal. Le but n'est pas de supprimer la sensation douloureuse. Par exemple, si vous prenez de la morphine, vous n'avez plus mal ; vous mettez votre main sur une plaque chaude, vous ne sentez rien, donc c'est dangereux. Nous, ce qu'on veut, c'est restaurer ce qu'on appelle un « seuil de nociception » qui soit normal. C'est-à-dire que vous n'avez mal que quand vous devez avoir mal. La douleur, c'est une alerte qui veut dire : « attention j'ai mal, donc il faut que je fasse attention ». Notre but est simplement que le patient ait de nouveau un seuil douloureux normal, comme tout un chacun. Et donc, quand on donne notre molécule (on l'administre une seule fois), pendant deux jours l'animal neuropathique va avoir la sensation de douleur d'une souris normale.

[>R1]: Et ensuite, il faudra à nouveau lui administrer la molécule ?

[>R2]: Oui. Ce sont des maladies chroniques, donc on ne sait pas pendant combien de temps il faudrait administrer la molécule pour que ce phénomène qui s'auto-entretient s'arrête. Est-ce qu'il faudra la donner pendant un an, cinq ans, toute la vie... ? On ne sait pas. Mais, à la limite, ces patients-là, ils souffrent tous les jours. Donc vous leur dîtes « vous prenez un comprimé tous les deux jours », peu leur importe.

[>R1]: Actuellement, vous savez que la durée d'effet de la molécule est de deux jours ?

[>R2]: A peu près deux jours, oui.

4 - 17:19 > 23:08 Méthode de travail

[>R1]: J'ai une question par rapport aux partenaires. Il y a trois partenaires...

[>R2]: En fait, il y en a deux : il y a ce laboratoire à Strasbourg et l'Institut des Neurosciences de Montpellier.

[>R1]: D'accord. Et l'UMS de Strasbourg - j'ai oublié son nom...

[>R2]: Ah oui, tout à fait, il y a une UMS qui était partenaire dans ce projet ANR. Puisque, pour trouver des molécules, on est parti d'une page blanche. Ici, une des spécialités de ce laboratoire, c'est de faire des cribles informatiques. C'est-à-dire qu'on va essayer de repérer informatiquement, sur la base de propriétés, les molécules qui, parmi des millions et des millions de molécules possibles, seraient capables d'inhiber la protéine qui nous intéresse. Les toutes premières molécules qu'on a identifiées ont donc été trouvées par crible informatique. On maintient une base de données d'environ 15 millions de molécules, toutes commercialement disponibles et qu'on peut acheter en poudre en trois semaines/un mois. C'est une sorte de panier de la ménagère, dans lequel on puise pour trouver des molécules de départ. Une fois qu'on a identifié ces molécules, on va ensuite les travailler au laboratoire pour en faire des analogues. C'est de là que démarre le travail de chimie puisque, adossée à la partie informatique, il y a toute la partie chimie. On va synthétiser des analogues de ces molécules, c'est-à-dire des molécules qui n'ont jamais été faites, qui ne sont pas commerciales et qu'on va pouvoir breveter – car, évidemment, vous ne pouvez pas breveter une molécule existante et c'est ça qui donne de la valeur au projet. Un peu plus tard, en parallèle de ce criblage virtuel par informatique, on a fait un vrai criblage expérimental. C'est là qu'est

intervenu notre partenaire PCBIS de l'UMS de V., qui a testé de manière robotisée 50 000 molécules. Il s'agit là d'un test physique. C'est ce que font toutes les compagnies pharmaceutiques pour identifier les premières molécules d'intérêt. Pour cela, ils utilisent des robots. Vous n'avez peut-être jamais eu l'occasion de voir comment ça se passe ; c'est assez impressionnant. Vous avez des robots avec des bras, qui vont aller manipuler des molécules. C'est fait dans des plaques qui ont 396 puits. Ce sont de toutes petites plaques ; on travaille vraiment sur des quantités minimes. On a donc criblé comme ça 50 000 molécules de la chimiothèque nationale, qui est une chimiothèque issue d'une trentaine de laboratoires académiques français. Ce sont des molécules qui appartiennent au patrimoine des universités françaises. Elles ont été informatisées et miniaturisées et on peut les cribler de manière robotisée. Ça c'est le travail qu'a fait V. Ce travail nous permet de faire ce qu'on appelle des « back up ». Développer un médicament c'est comme prendre sa voiture à Strasbourg pour aller à Lyon, avec le risque d'avoir un contrôle de police tous les 50 mètres. Tous les 50 mètres on peut vous prendre soit votre permis, soit votre voiture. Un médicament ça n'est pas un produit d'ingénierie ; c'est fait d'étapes qu'il faut suivre de manière chronologique et qui doivent toutes être positives. L'échec est interdit. C'est-à-dire que, si vous échouez à la première étape, vous ne pouvez pas vous dire : « ça n'est pas grave, je vais faire l'étape 2 et j'essayerai de résoudre l'étape 1 plus tard ». Non, ça n'est pas possible. Quand vous développez un médicament, toutes les étapes doivent être positives chronologiquement, de la première à la dernière. Et il y a un nombre incalculable d'étapes. En réalité, c'est pour des problèmes de sûreté, d'innocuité. C'est pour ça que c'est difficile et c'est pour ça que ça coûte cher. Donc, quand on développe une série chimique, comme celle issue de nos premiers travaux, on fait des tests de plus en plus poussés. Parce qu'il ne suffit pas que la molécule fasse ce qu'elle est censée faire. Il faut aussi qu'elle ne soit pas toxique ; il faut qu'elle soit bio-disponible (c'est-à-dire que, si vous l'avalez sous forme de comprimé, il faut qu'elle passe dans le sang et ne soit pas détruite) ; il ne faut pas qu'elle inhibe une autre protéine... Il y a énormément de facteurs à prendre en ligne de compte et beaucoup, beaucoup, beaucoup de tests à faire. Donc, ce qu'on fait toujours c'est que, quand on développe une série chimique comme ça, en parallèle, on va essayer de développer d'autres molécules appartenant à d'autres familles chimiques, au cas où il y ait des problèmes avec la première série. Si la première série a un problème, on est certes obligés de reculer, de repartir en amont, mais on peut tout de même

continuer le projet, en utilisant une des autres séries chimiques développées. Donc on ne développe jamais une seule série ; on en développe en général plusieurs à la fois. Le criblage robotisé expérimental qu'on a fait avec cette équipe de l'UMS, c'était pour apporter d'autres molécules actives/d'autres séries chimiques au support de notre projet.

5 - 23:08 > 25:25 Ressources humaines

[>R1]: Comment vous répartissez-vous les rôles entre le laboratoire de Montpellier et ici ?

[>R2]: Les rôles sont bien établis, puisqu'ils sont répartis en fonction de nos compétences. Nous, on est chimistes ; donc notre métier c'est d'imaginer des molécules et de les synthétiser. Le laboratoire de Montpellier, c'est un laboratoire de biologie ; donc leur rôle c'est de tester ces molécules sur tout un tas d'essais (ça peut aller d'une cellule à un animal). On est donc, de ce point de vue-là, très complémentaires.

[>R1]: Ici, à Strasbourg, combien de personnes travaillent sur le projet ?

[>R2]: C'est difficile à quantifier, parce qu'évidemment tout un tas de CDD se sont succédés. On va dire qu'en général on a en moyenne entre un et deux CDD pour faire de la chimie et un CDD pour faire le criblage expérimental. Le criblage expérimental a duré un an. Et la chimie, ça fait à peu près 3 ou 4 ans qu'on en fait.

[>R1]: Il y a des doctorants qui travaillent pour le projet ?

[>R2]: Non. Ce ne sont pas des doctorants, parce que c'est un projet de valorisation qui n'est pas très propice pour faire une thèse : il y aura dépôt de brevets, donc confidentialité des résultats ; c'est-à-dire qu'il n'y aura pas de publication immédiate, les publications seront différées. Là, la première publication va paraître d'ici 15 jours à 3 semaines, après 4 ou 5 ans de travail. Il faut évidemment qu'on protège les résultats, c'est-à-dire qu'on fasse une demande de brevet, que cette demande de brevet soit acceptée et, une fois qu'on a sécurisé l'aspect propriété intellectuelle, on peut se permettre de publier. Donc, pour un thésard, ça n'est pas un sujet idéal, puisqu'il aura du mal à valoriser ses travaux de recherche immédiatement. Ce qu'on fait, c'est qu'on fait appel à des CDD, qui sont des ingénieurs de valorisation (avec généralement un niveau ingénieur ou M2) voire des post-docs (mais qui savent en

connaissance de cause que c'est un projet de valorisation et qu'il n'y aura pas de publication immédiate). Ce genre de projet n'est pas très adapté pour recruter des doctorants.

6 - 25:27 > 29:32 Dépôt de brevets

[>R1]: Il y aura un dépôt de brevet ?

[>R2]: Là on a déjà cinq brevets déposés.

[>R1]: Ça correspond à un brevet par molécule... ?

[>R2]: Ça dépend. Là c'est vraiment stratégique et il est notamment important de s'adosser à une start-up, avec tout l'aspect juridique réglementaire, pour s'assurer qu'on va être capables de protéger nos résultats. Donc il y a ce qu'on appelle un brevet [inaudible] : si une quelconque personne dans le monde développe une molécule qui inhibe cette cible pour traiter des douleurs neuropathiques, ça nous appartient. Ce concept-là a été breveté. C'est-à-dire que ça nous protège de nos concurrents : si nos concurrents veulent faire la même chose, ils ne pourront pas, à moins qu'ils nous payent une licence d'exploitation. Ensuite, il y a des brevets par série chimique : on dit « voilà, nous revendiquons l'utilisation de telle formule chimique pour traiter des douleurs neuropathiques ». Toute molécule possédant cette formule chimique entre dans le cadre du brevet et – on va dire – nous appartient, ou plutôt l'utilisation de cette molécule nous appartient. C'est-à-dire qu'un tiers ne peut pas utiliser ce genre de molécule dans ce but-là, à moins de nous payer une licence d'exploitation. Et donc, pour chaque série chimique différente, on a un brevet qui protège. Ensuite, on peut aussi protéger des associations. Par exemple, on est en train de déposer un brevet qui revendique l'association de notre molécule avec de la morphine, car ça permet de diminuer les doses de morphine de manière considérable. Actuellement, il y a un gros problème aux États-Unis : c'est ce qu'on appelle la crise des opiacées (qu'on ne connaît pas trop en France). Les Américains consomment énormément d'opiacées/de morphine. Or la surconsommation de morphine induit de l'accoutumance, de l'intolérance et donc beaucoup de décès. C'est pourquoi les États-Unis essaient de trouver tous les moyens possibles et imaginables pour réduire les doses de morphine, sans en perdre l'efficacité. Donc, là, c'est une association nouvelle. A partir du moment où vous avez une inventivité et que ce que vous proposez n'est pas évident pour l'homme de l'art, c'est brevetable.

[>R1]: En fait, on peut breveter non seulement des données mais aussi un... ?

[>R2]: Vous pouvez breveter une idée, oui. C'est plus compliqué, parce que, généralement, quand vous essayez de breveter une invention, on vous demande la preuve expérimentale que ce que vous avancez est vrai. Donc, quand vous brevetez une idée, c'est compliqué de prouver que cette idée est bonne, tant que vous n'avez pas d'exemple à l'appui.

[>R1]: Comment avez-vous fait, par exemple, pour breveter le fait que vous travaillez sur tel inhibiteur dans le but de traiter des douleurs neuropathiques ?

[>R2]: On va décrire la structure et les propriétés de la molécule. On va aussi décrire les propriétés d'inhibition de la protéine et dire comment on mesure l'inhibition de cette protéine. Alors, on ne le fait pas pour toutes les molécules, mais on le fait pour quelques exemples – des exemples caractéristiques, qui montrent qu'on a bien fait le travail et que ce qu'on avance est fondé sur des données expérimentales.

[>R1]: D'accord. Donc il faut déjà être avancé dans le travail pour pouvoir déposer un brevet ?

[>R2]: Ah oui, oui, bien sûr. Parce que l'examinateur de votre brevet peut vous dire : « écoutez, selon moi, il n'y a pas assez d'évidences expérimentales que ce que vous avancez là est vrai ; donc vous ne pouvez pas revendiquer cette invention ».

7 - 29:32 > 36:22 Données de recherche

[>R1]: Je vais passer à la partie « données ». Est-ce que vous-même utilisez spontanément le terme « données » ?

[>R2]: Oui... Évidemment, on utilise le terme « données », puisqu'on va avoir – on va dire –...

[>R1]: Par exemple, qu'est-ce que vous désignez sous ce terme, si on prend l'exemple du projet B. ?

[>R2]: Je dirais qu'il y a peut-être trois types de données. Il y a, d'une part, des données de chimie, qui sont les structures des molécules. Ce sont des structures particulières, qui permettent à n'importe quel chimiste de savoir quelle est l'entité moléculaire de cette

structure/formule. Ensuite, on a des données attenantes à la manière dont on a synthétisé cette molécule. On synthétise la molécule en laboratoire et ça se fait par étapes.

[>R1]: Donc c'est plus le processus ?

[>R2]: Oui, tout à fait. C'est un processus de synthèse de la molécule. C'est la manière dont on l'a synthétisée, dont on l'a créée. Ça se fait en plusieurs étapes : parfois il y a dix étapes, parfois il y en a trois... C'est un processus qui peut être long. Ce sont des étapes de synthèse.

[>R1]: Et ça se présente sous quelle forme ?

[>R2]: Vous voulez que je vous montre un exemple ?

[>R1]: Oui, je veux bien.

[>R2]: Ça se présente sous forme de schémas réactionnels, qui permettent à un chimiste – c'est le principe de la science – de reproduire ce processus et de refaire cette molécule dans les conditions que l'on a décrites. Je vais vous montrer, par exemple, [il recherche sur son disque dur]... Je vais vous montrer un brevet. A quoi ressemble un brevet. [Il ouvre un fichier]. Ça c'est le brevet tel qu'il a été accepté. Là c'est la formule chimique qu'on revendique. On revendique également la famille/la combinatoire possible (c'est-à-dire en R1 on trouve tel groupement chimique, en R2...). Puis on développe la manière dont on a synthétisé la molécule finale depuis le plus petit bloc (qui généralement est un bloc commercialement disponible) : on dit « voilà, cette molécule-là, vous l'achetez chez un tel, vous la chauffez à 120°C en présence de ceci, de cela, elle donne telle molécule ; ensuite, cette nouvelle molécule, on la traite par ceci, cela, elle donne, etc. ». On donne exactement ce schéma réactionnel, qui permet de reproduire la synthèse de la molécule qu'on veut décrire. Ça c'est le procédé de synthèse, qui est aussi un type de données. Troisième type de données : ce sont les propriétés physico-chimiques de la molécule, qui permettent de la caractériser. Ce sont des propriétés physiques ou chimiques qui sont uniques et qui permettent d'identifier une molécule. Pour s'assurer qu'on a bien fait la bonne molécule, il faut notamment que les propriétés physico-chimiques concordent avec ce qu'on attend. On est souvent amenés à refaire de nouveaux lots. Or il faut qu'à chaque étape on ait exactement la molécule intermédiaire souhaitée, jusqu'à la molécule finale. Pour s'assurer qu'on a la bonne molécule, on fait des caractérisations physico-chimiques et il faut que notre molécule finale ait exactement les propriétés attendues. Donc ça ce sont des données de chimie : les structures,

les procédés de synthèse, les caractéristiques analytiques et les caractéristiques physico-chimiques. Ensuite, vous avez des données biologiques, qui vont décrire l'activité biologique de la molécule. Nos collègues biologistes vont tester la molécule sur tout un tas de protocoles différents, allant de l'étude de la molécule dans un tube à essai jusqu'à l'étude de la molécule administrée à une cellule, voire à un animal entier/vivant. On a donc des modèles qu'on appelle « in vitro » – dans des tubes à essai. On va regarder l'effet de la molécule sur une ou plusieurs protéines. C'est ce par quoi on commence, parce que c'est ce qui coûte le moins cher et que ça nous permet d'obtenir la preuve de concept. Ensuite, on passe à l'échelle de la cellule : quel est l'effet de la molécule sur une cellule. Ça peut être une cellule normale ou bien une cellule malade (dans laquelle on a introduit un déséquilibre quel qu'il soit). Et, à la fin, on teste l'effet de la molécule chez l'animal. On est alors sur des données « in vivo ». Ça ce sont les données de biologie.

8 - 36:22 > 46:54 Protection des données

[>R2]: Si on considère ces deux types de données séparément (seulement les données de chimie ou seulement les données de biologie), un tiers ne saura pas trop quoi en faire. Ces données sont difficilement exploitables en tant que telles. Évidemment, en elles-mêmes, elles sont importantes ; mais ce n'est que quand on couple et qu'on associe les données de chimie aux données de biologie que vient la valeur ajoutée, à savoir : « la molécule 27 a l'activité décrite dans cet essai-là », « la molécule 43 a l'activité décrite dans cet essai-là » ou « quand on modifie la molécule 27 en la molécule 43, on améliore/on perd l'activité décrite ou la molécule a un effet plus/moins durable », etc. C'est cette association-là qui donne de la valeur. Je vais vous montrer à quoi ressemblent des données de biologie [il recherche dans le fichier montré précédemment]. Voilà, vous avez ici une table avec des données de biologie : « la molécule 1 a ... » et vous avez des valeurs. Bon, et bien ça, ce n'est pas exploitable. Si vous ne connaissez pas la structure de 1, ça n'est pas exploitable. Donc, ce genre de tableaux, on peut se les échanger librement par mail, il n'y a pas de problème. A la limite, on pourrait faire la même chose avec les structures. On ne le fait pas, parce qu'on tient quand même à ce que nos structures soient originales. Donc, généralement, quand on les communique par mail, les documents sont cryptés avec des mots de passe. Ce qu'on ne fait jamais – au grand jamais –,

c'est envoyer un document où il y a à la fois la structure et la biologie. Ça, on ne le fait jamais. Car c'est de la très, très haute valeur ajoutée.

[>R1]: Comment faites-vous, dans ce cas ?

[>R2]: Ça reste en interne.

[>R1]: Donc ce sont les équipes qui se déplacent l'une vers l'autre ?

[>R2]: Moi je stocke des données ici. Je vais essayer de vous trouver un article, où vous avez l'association des deux types de données… Voilà, là vous avez tout : vous avez à la fois les structures, les variations structurales et les activités. Alors, on ne s'envoie jamais ça par mail, jamais.

[>R1]: C'est une publication ?

[>R2]: Au départ, c'est un brevet. Une fois qu'on a consolidé la propriété intellectuelle, on peut le publier. Mais, tant qu'il n'a pas été déposé, on ne communique pas dessus. Jamais. On ne mélange jamais les données biologiques et les données de chimie. Jamais. On peut les envoyer séparément et la personne qui les reçoit fait l'association des deux fichiers. Mais on n'envoie pas le travail tout fait. Parce que si vous vous faites intercepter, vous n'avez plus rien. N'importe quel chimiste peut refaire et, alors, vous n'avez plus que vos yeux pour pleurer. Donc, ça, on ne le fait jamais.

[>R1]: Vous parliez de la sauvegarde des données. Comment ça se passe pour ce projet ?

[>R2]: On fait ce qu'on peut. C'est-à-dire que le CNRS nous impose – mais ça c'est une politique du CNRS – que chaque ordinateur portable soit crypté. En théorie, ils nous déconseillent fortement de voyager avec nos ordinateurs portables. Parce que, si on vous vole votre ordinateur portable et qu'il n'est pas protégé, vos résultats se trouvent dans la nature. Donc, à tout le moins, les disques durs des ordinateurs sont cryptés/protégés ; il faut un mot de passe particulier pour avoir accès aux fichiers (ce n'est pas le mot de passe de Windows). On fait des sauvegardes journalières avec, idéalement, une sauvegarde sur site et une sauvegarde à l'extérieur. Ça, c'est ce qui est recommandé. Il y a par exemple des clouds. L'université en propose un, qui s'appelle Seafile. Le CNRS en a un également : MyCore, qui permet d'externaliser vos données et de les synchroniser avec un espace sécurisé, où des

informaticiens professionnels sont là pour protéger l'accès à vos données. Parce qu'évidemment on ne va pas mettre ça sur Amazon ou sur le web.

[>R1]: Est-ce que Seafile ou MyCore vous paraissent adaptés d'un point de vue sécurité ? Vous les utilisez ?

[>R2]: C'est limité. Par exemple, Seafile est limité à 15 Go ; MyCore à 20 Go. Le CNRS a pour ambition de passer à 100 Go pour chaque agent CNRS. C'est compliqué... A partir du moment où un hacker veut pirater vos données, s'il a envie de les pirater, il piratera. Le but c'est quand même de faire en sorte qu'il y ait suffisamment de portes fermées. Et, surtout, nous aussi on se prémunit, on se protège. C'est important pour nous de garder nos données protégées. C'est important aussi vis-à-vis de nos instituts – que ce soit l'Université ou le CNRS. Vous pouvez imaginer que, si on se fait pirater nos données, alors qu'il n'y avait même pas de mot de passe sur l'ordinateur et que les données étaient ouvertes à tous les vents, c'est compliqué… Il y a aussi des niveaux de sécurité et d'accès à certains laboratoires qui sont en train de se mettre en place à partir de cette année. On va passer dans une zone (on a un terme assez barbare pour ça) qui s'appelle « ZRR ». Je ne sais pas si vous en avez entendu parler. Ce sont des « Zones à Régime Restreint ». Le Ministère de la Défense a évalué le potentiel scientifique des laboratoires de la nation et a estimé que certains laboratoires étaient sensibles. Que ce soit pour des questions de travaux (il y a des expériences qui sont dangereuses, avec par exemple des molécules toxiques ou des réactions nucléaires) ou parce qu'il y a un potentiel de valorisation vers l'extérieur. Ces Zones d'accès à Régime Restreint sont imposées par le Ministère de l'Intérieur. Ça régit les entrées et sorties dans le laboratoire. Nous, officiellement, on devait passer en ZRR au 1er janvier. Pour l'instant on n'y est pas encore. Sinon, vous n'auriez pas pu rentrer. C'est-à-dire que vous ne pouvez pas rentrer sans avis et vous devez signer un registre d'entrée et de sortie. C'est pour qu'on sache qui rentre et qui sort. Ici, vous êtes dans une fac : il y a des étudiants ; c'est ouvert à tous les vents ; n'importe qui rentre, n'importe qui sort. A un moment, ça n'est plus possible. Donc, si quelqu'un de l'extérieur doit venir, il faut qu'on en fasse la demande au Ministère de l'Intérieur. Ils font une enquête et nous disent « oui, cette personne a le droit de venir chez vous » ou « non, cette personne n'a pas le droit de venir chez vous ».

[>R1]: Par exemple, moi, s'il y avait déjà eu...

[>R2]: Non, ça concerne les personnes qui restent plus de deux mois. Puisqu'il y a du pillage industriel. Même à l'Unistra, c'est déjà arrivé. Je veux dire, vous avez des gens qui viennent faire de l'espionnage industriel. Vous avez l'impression que ce sont des choses qui n'existent pas ; ce sont des choses qui existent. Je ne sais pas si, dans le cadre de votre thèse, vous aurez l'occasion de vous attacher à la sécurité/à la protection des données en général (c'est quand même un élément très important en recherche) et à l'aspect qui est derrière, à savoir le pillage des données… Moi je vous le conseille, si ça vous intéresse. Vous entendrez des choses qui sont assez ahurissantes. Et ça n'est pas de la science-fiction. Vraiment, ça arrive.

[>R1]: Oui, on en entend parler mais ça nous paraît assez rare.

[>R2]: Ah non, ça n'est pas rare du tout. Une fois par an, on a la visite d'agents de la DGSI (la Direction Générale de la Sécurité Intérieure), qui viennent nous sensibiliser aux risques de pillage et aux mesures qu'il faut prendre pour éviter au maximum de se faire piller. Par exemple, si vous allez à un congrès, vous n'amenez pas votre ordinateur avec vous. Car vous avez des gens (ça arrive) qui viennent visiter votre chambre d'hôtel, pendant que vous n'êtes pas là, et qui désossent votre ordinateur. Donc, généralement, on nous dit « voyagez juste avec une clé USB et votre conférence sur votre clé USB ». On a l'impression que c'est de la science-fiction. Ça n'est pas du tout de la science-fiction. Vous avez des pays qui font de l'espionnage à grande échelle. Je ne vais pas les mentionner, mais vous avez un certain nombre de pays qui sont sur des listes rouges (vous allez à l'Unistra, ils vous diront lesquels). C'est-à-dire qu'un étudiant, qui vient d'un de ces pays et qui demande à venir étudier à l'Unistra, va être surveillé – enfin, il va faire l'objet d'une enquête.

[>R1]: Qui a intérêt à ce que les données soient protégées : l'État, le...?

[>R2]: Tout le monde. Je veux dire, le chercheur a intérêt à ce que ses données soient protégées, parce qu'il n'a pas envie que cinq années de sa vie partent en fumée, comme ça, en l'espace d'une après-midi, parce que quelqu'un aurait eu accès à ses données. L'employeur a également intérêt à ce que les données soient protégées, pour qu'il puisse continuer à en revendiquer la paternité. Les données ne nous appartiennent pas ; les données qu'on génère appartiennent au CNRS (puisque moi je suis CNRS) et à l'Université de Strasbourg (puisque je suis accueilli dans des locaux de l'Université de Strasbourg). Les résultats ne nous appartiennent pas. On a souvent tendance à croire que les résultats nous appartiennent ; les

résultats ne nous appartiennent pas. On est locataires de ces résultats. Notamment quand on a déposé un brevet pour telle ou telle molécule et que ce brevet est acheté par une compagnie qui en veut les droits, une partie des royautés va atterrir au CNRS et à l'Université de Strasbourg, une partie au laboratoire et une partie aux inventeurs.

[>R1]: Aux « Inventeurs », c'est-à-dire à vous, les chercheurs ?

[>R2]: Alors, non. « Inventeur » c'est inventeur au sens juridique du terme. Un inventeur c'est une personne qui a eu une activité inventive.

[>R1]: Donc un enseignant-chercheur ou un chercheur ?

[>R2]: Pas forcément. C'est quelqu'un qui a eu une activité inventive. C'est-à-dire que le résultat final n'aurait pas pu se faire, si cette personne n'avait pas été là.

[>R1]: Comment peut-on le déterminer ?

[>R2]: Oui, c'est compliqué. Souvent, les offices de brevet considèrent que les doctorants ou les post-docs n'ont pas d'activité inventive, puisqu'ils ne font que faire ce que leur disent leurs chefs. Et donc, généralement, ce sont les chefs les inventeurs. A moins que l'étudiant ou le post-doc ait eu une idée particulière, qui témoigne que, sans cette idée, on n'aurait pas pu faire cette molécule. Là, vous pouvez témoigner de votre inventivité. Donc tout le monde a intérêt à ce que les données soient protégées. Archivées, évidemment, sauvegardées, protégées. Parce que c'est le patrimoine – le nôtre et celui de nos employeurs.

9 - 49:20 > 53:44 Conservation des données

[>R1]: Est-ce que vous conserverez toutes les données à la fin du projet ? Le fait qu'il y ait des brevets vous oblige-t-il à…?

[>R2]: Non. A partir du moment où vous avez un brevet qui est déposé, celui-ci est accessible sur le web. Donc il n'y a pas de problème.

[>R1]: Il n'y a pas besoin d'autres données ?

[>R2]: Non. En fait, vous avez la paternité du brevet pendant 20 ans. Une fois que le brevet vous est accordé, pendant 20 ans vous avez l'exclusivité pour exploiter commercialement cette découverte. Alors, je ne sais pas quelle est la législation… Je sais, par exemple, qu'il y a

une législation pour les hôpitaux : dès que vous avez des données de patients, la loi vous impose de conserver ces données pendant 40 ans (il me semble mais je ne suis pas sûr). Je crois qu'il faut que les données soient pérennes pendant une durée assez longue. Nous, nous n'avons pas cette exigence légale. Il y en a peut-être une, mais je ne la connais pas. En tout cas, on conserve toutes nos données. On les stocke, on les archive toutes. On ne jette rien. Alors, on ne jette rien au niveau des données dépouillées, archivées, analysées.

[>R1]: Quelles sont les autres données ?

[>R2]: Il y a les données brutes. C'est-à-dire, par exemple, ce que va vous cracher une machine. Ce sont des fichiers Excel. Ce sont des données brutes, qui ne sont pas analysées.

[>R1]: Ces données brutes, vous ne les gardez pas ?

[>R2]: On les garde pendant la durée du projet et, une fois que le projet est terminé, on peut éventuellement les supprimer. Ces données brutes, il faut les analyser. Soit c'est l'homme qui le fait, soit c'est une machine, qui va analyser ces résultats et vous dire ce que vous voulez y voir. Ce qu'on obtient, ce sont des données analysées. Ensuite, il y a les données analysées dépouillées, desquelles on a extrait la quintessence des résultats...

[>R1]: Vous avez interprété...

[>R2]: On a interprété les résultats. Souvent, ce sont des publications. Mais tout n'est pas publié. On garde également les données dans les cahiers de laboratoire.

[>R1]: Oui, vous avez aussi des cahiers de laboratoire ? Ils sont au format papier ?

[>R2]: Non. On a les deux. Maintenant on a des cahiers de laboratoire électroniques.

[>R1]: Qui vous les fournit ?

[>R2]: Il y a tout un tas de fournisseurs. En fait, il y a une start-up alsacienne qui s'appelle Novalix, qui fournit des cahiers de laboratoire électroniques. C'est un ancien du labo. Toutes nos expériences, tout ce que je vous ai montré là, est stocké sous forme numérique dans des cahiers de laboratoire électroniques. On peut également faire des copies sur des cahiers de laboratoire papier. Dans un cahier de laboratoire, le chercheur recense de manière journalière l'activité de chaque projet. Ça se présente comme ça [il me montre un exemple de cahier de laboratoire papier]. En fait, ça, c'est une preuve. On l'utilise en cas de contestation d'une invention. On a la date, les signatures ; on sait qu'on a tels résultats, qu'on a fait cette

molécule tel jour ; et donc on peut revendiquer l'antériorité de la découverte. Donc ça c'est très important. Tous les résultats sont consignés sur des cahiers de laboratoire papier et/ou électroniques, puisque maintenant, sur les cahiers de laboratoire électroniques, on peut stocker à la fois du texte, des images, des vidéos… On peut stocker tout ce qu'on veut.

[>R1]: Est-ce que vous avez abandonné les cahiers papier ?

[>R2]: On abandonne de plus en plus les cahiers papier, oui, puisque, quand on stocke nos résultats d'expérience, quand on décrit nos expériences, c'est de toute façon sous forme électronique. Donc vous n'avez que des copier/coller à faire. C'est beaucoup plus facile d'archiver ses données dans un cahier de laboratoire électronique. Et le lien est plus facile aussi. C'est-à-dire qu'il y a des moteurs de recherche pour savoir « là, telle molécule, je l'ai faite quand ». S'il faut aller fouiller dans quinze cahiers de laboratoire, ça n'est pas évident. Alors que là, vous faites une requête électronique, vous dites « je veux savoir, pour cette donnée-là, à quelle date je l'ai générée, combien j'en ai d'exemplaires… » et vous avez des moteurs de recherche qui vous permettent d'extraire les données automatiquement.

10 - 53:44 > 55:02 Valeur économique des données

[>R1]: Je ne sais pas si c'est la peine de revenir sur cette question. Vos données ont-elles un enjeu économique ?

[>R2]: Dans le cadre de ce projet, oui. Mais on a des projets qui sont complètement académiques, dans lesquels on essaie de défricher une idée, sans savoir quel sera son potentiel de valorisation. Dans ces projets, on veut simplement faire progresser la science. Alors, c'est vrai qu'on est dans un domaine d'activité qui est très proche de la valorisation. Plus proche que si on s'intéressait à l'histoire de l'art dans l'Égypte ancienne, par exemple. Là on est vraiment très près de la valorisation, parce qu'on crée des molécules actives avec un enjeu en santé humaine. Donc le pas vers la valorisation est relativement tenu. Mais tous les projets ne deviennent pas des projets de valorisation. Pas forcément.

[>R1]: Peut-être ont-ils une valeur économique indirecte ?

[>R2]: La valeur indirecte ça peut être... Si vous faîtes des publications de très haut niveau, où est mentionnée l'Université de Strasbourg, celle-ci va être reconnue internationalement. Ça va

remonter au classement machinchouette et donc il y aura quand même un impact économique derrière.

11 - 55:02 > 57:53 Ouverture des données

[>R1]: Ça ne vous concerne probablement pas encore, mais il existe des injonctions pour la diffusion/l'ouverture des données scientifiques au niveau de l'Union européenne, et peut-être bientôt de l'ANR. Avec la possibilité de placer les données en accès restreint quand, effectivement, elles sont confidentielles. Je voulais savoir ce que vous en pensiez…

[>R2]: Les accès libres (open access, etc.) c'est compliqué à gérer. Je vais vous donner un exemple. Il y a tout un tas de publications qui sont en open access…

[>R1]: Vous voulez dire parmi vos publications ou dans votre domaine ?

[>R2]: J'en ai peut-être quelques unes parmi mes publications. Mais on a une pression de notre environnement – de l'Université, plus que du CNRS, je pense – pour diffuser de l'information en open access. Nous, par exemple, on communique essentiellement de deux manières : ce sont soit des communications orales, soit des articles dans des journaux spécialisés/scientifiques. Parfois, c'est dans la presse mais c'est rarissime. Parce que, généralement, quand on communique dans la presse, c'est plutôt une conséquence qu'une cause (c'est la conséquence d'un travail qui a été particulièrement exposé). Publier en open access, ça a un coût. Si on prend l'exemple des *PLOS*, ça coûte à peu près 2000 à 2500 dollars. Bon, très bien : votre article est en open access mais c'est 2000 à 2500 dollars. Dans leur grande sagesse, évidemment, ni l'Université, ni le CNRS, ni l'État ne nous donnent des budgets pour publier en open access. C'est-à-dire qu'on sera obligés de prendre sur nos petits budgets de recherche, qu'on a déjà du mal à consolider, pour publier en open access. Donc si vous vous posez la question : « est-ce que je publie en open access ou est-ce qu'avec ces 3000€ je fais une voire deux expériences supplémentaires pour publier chez Wiley ou Elsevier, où la publication sera gratuite »… Alors, évidemment, il y a des abonnements qu'on paie. Ça je suis d'accord. C'est un peu là, où le système est déséquilibré. Moi, ça ne me dérangerait pas qu'on publie en open access, mais il faudrait qu'on se donne les moyens ou qu'on nous donne les moyens de le faire. Et s'il y a une politique de publication en open

access, qu'il y ait également les moyens financiers pour supporter cette politique. Ce qui n'existe pas vraiment.

12 - 57:53 > 1:02:04 Politiques des éditeurs en matière de données scientifiques

[>R1]: Est-ce que les éditeurs scientifiques, chez lesquels vous publiez, vous demandent parfois d'associer à la publication les données liées (sous forme de fichier joint) ?

[>R2]: Oui.

[>R1]: C'est obligatoire ou c'est plutôt conseillé ?

[>R2]: Ça dépend. Par exemple, ça peut être quand vous publiez la structure d'une nouvelle protéine (ça c'est plutôt pour les gens qui font de la biologie structurale). On va vous demander de déposer la structure soit conjointement à votre publication (c'est-à-dire dans un format qui permet de la lire), soit dans une base de données publique avant la publication. Et votre publication n'est acceptée qu'à cette condition-là.

[>R1]: Pourquoi doit-on déposer la structure au préalable dans la base de données publique ?

[>R2]: Parce que la base de données va faire un « check » de cette structure. Elle va regarder si les données sont correctes, s'il n'y a pas de problème. Comme ça, les éditeurs s'affranchissent du contrôle. Le contrôle est fait en amont.

[>R1]: Comment s'appelle cette base ?

[>R2]: Alors, il y a plein de bases de données. Il y en a une qui s'appelle la Protein Data Bank (PDB). La Protein Data Bank est une base de données qui rassemble l'ensemble des structures de protéines libres d'exploitation dans le monde. Enfin, « libres »… Les structures sont libres. Quand vous voulez publier une structure, vous êtes obligés dans votre publication de donner un code, qui est le code d'accès à cette base de données. Donc, si votre structure n'a pas été enregistrée par cette base de données en amont, votre publication ne sera pas acceptée. Ou bien, si c'est du code informatique, par exemple, les éditeurs peuvent vous demander de fournir une copie du code informatique. Ils vont vous dire : « donnez-nous une copie de votre code, pour que les lecteurs puissent le télécharger » ou « donnez-nous un site où l'on peut télécharger le logiciel que vous avez développé ». Ça arrive assez souvent.

[>R1]: Vous même développez parfois des logiciels ?

[>R2]: Oui, on développe aussi des logiciels. Et ça arrive que le logiciel doive être en accès libre. En accès libre, simplement pour une utilisation non commerciale/académique (ça c'est toléré). Puisque, sinon, vous avez aussi le risque que des sociétés privées pillent votre travail et vous pilonnent vos logiciels. Donc, généralement, les éditeurs demandent l'accès libre pour des motifs de recherche académique/non commerciaux. Vous donnez le code source du logiciel ou vous donnez la structure de la protéine qui vous intéresse. De plus en plus souvent maintenant, dans les publications, la partie « matériel supplémentaire » est bien plus touffue que l'article en lui-même. Puisque, souvent, dans les journaux scientifiques, on est limités à une certaine taille d'article (si vous avez un article de 50 pages, personne ne va le lire). Donc vous êtes limités en nombre de figures, en nombre de tables, en nombre de caractères... Tout ça est très calibré. Par exemple, toutes les publications du groupe *Nature* sont calibrées au caractère près : votre abstract doit faire 150 caractères, pas un de plus ; vous avez droit à quatre figures, pas plus. C'est vraiment très calibré. Donc, quand vous avez de grosses données, tout est mis dans les matériels supplémentaires. La personne qui s'y intéresse va aller voir dans les matériels supplémentaires. Alors, c'est sûr, l'open access, c'est intéressant en SHS. Pour nous, c'est un peu plus compliqué, puisqu'on est très proches de la valorisation. On va dire ça comme ça.

[>R1]: Oui, il vous faut protéger en amont les résultats.

[>R2]: Oui. A partir du moment où c'est protégé, il n'y a aucun souci, vous pouvez mettre en libre accès.

13 - 1:02:04 > 1:09:44 Parcours professionnel

[>R1]: J'aurais une dernière question, peut-être plus personnelle. Dans quel établissement avez-vous effectué votre formation universitaire ?

[>R2]: Dans beaucoup ! Au départ, j'ai fait des études de pharmacie à l'Université de Rennes. Ensuite, j'ai fait une thèse ici à l'Université de Strasbourg.

[>R1]: Et ensuite vous êtes devenu chercheur ici ?

[>R2]: Ensuite, je suis allé à l'étranger pendant dix ans. Puis je suis revenu.

[>R1]: C'était dans quel pays ?

[>R2]: J'étais en Allemagne et en Suisse.

[>R1]: D'accord. C'était pour savoir un peu si, par exemple en Allemagne ou en Suisse, certaines pratiques de gestion des données avaient pu vous influencer...

[>R2]: Ça ne m'a pas trop influencé, puisque je n'étais pas jeune chercheur. En fait, l'évolution qu'on voit – notamment dans le domaine de la santé – c'est le big data : le nombre/ le volume de données qui a explosé au cours des cinq dernières années. Quand j'étais en thèse, on travaillait sur un volume de données très restreint. Il n'y avait pas beaucoup de données, parce que ces données étaient difficiles à générer. Maintenant, vous avez eu des progrès en miniaturisation, en automatisation, etc. Ça vous permet de générer des données à un débit considérablement plus important qu'il y a vingt ans. Pour faire une comparaison, il y a vingt ans, si on voulait cribler informatiquement une molécule, on mettait peut-être une semaine par molécule. Alors que maintenant, avec les centres de calcul, on peut faire dix millions de molécules en une nuit. Donc tout ça génère des données – c'est le big data (le volume de données est énorme). Et ça a changé considérablement notre manière de travailler : on est passé progressivement d'une approche basée sur la génération de connaissances à une approche qui est de plus en plus dans l'exploitation des connaissances déjà connues. Ne serait-ce qu'exploiter de manière rationnelle les tonnes de connaissances aujourd'hui disponibles vous permet d'en générer de nouvelles. Et donc ça, ça a changé ; ça n'existait pas il y a vingt ans. Maintenant, il y a quinze millions de molécules disponibles dans le monde ; il y a des centaines de millions d'activités biologiques disponibles à tous, que vous pouvez exploiter. Ensuite, c'est votre intelligence et la manière dont vous exploitez ces résultats qui fait que vous allez générer de nouvelles connaissances. Alors, ça a des avantages et des inconvénients, puisque dans ces données il y a du signal – évidemment – mais il y a aussi beaucoup, beaucoup de bruit de fond. Et donc, comment différencier le signal du bruit de fond ? C'est quasiment une discipline en soi. Vous irez peut-être voir des gens qui travaillent sur les machines d'apprentissage ou sur l'intelligence artificielle, dont c'est le métier d'essayer d'extraire la quantité infinitésimale de signal dans un océan de bruit de fond.

[>R1]: Ce bruit, ça ne ralentit pas votre travail ? Par exemple, quand vous criblez des molécules.

[>R2]: Ça ralentit notre travail, si on le fait mal. C'est-à-dire qu'il faut bien vérifier au départ les données que l'on a. Parce qu'un des gros problèmes du big data, c'est la fiabilité des données qui sont disponibles. Une donnée scientifique, pour qu'elle soit exploitable, doit être fiable et reproductible. Donc, si vous avez dix valeurs pour une même donnée dans dix labos différents, c'est quelque chose que vous ne pouvez pas exploiter.

[>R1]: Qu'en est-il des molécules de la banque nationale, par exemple ? Est-ce qu'il vous arrive de rencontrer des problèmes liés à la qualité des données ?

[>R2]: A partir du moment où vous travaillez avec du vivant, il peut toujours y avoir des variations de pureté et d'impureté. Mais, dans le big data, il y a des gens dont le but est d'essayer d'exploiter les données disponibles sur le web, pour générer de nouvelles connaissances. Le travail c'est le travail du traitement des données – même si maintenant le calcul n'est plus un problème (il y a des ordinateurs qui vous permettent de faire des puissances de calcul phénoménales). Le problème, c'est « checker » que les données à analyser sont correctes/fiables. Les données proviennent de droite et de gauche. Est-ce que c'est une donnée brute que vous analysez ? Est-ce que c'est une moyenne ? Etc. Vous avez tout et n'importe quoi. L'objectif c'est d'avoir des données homogénéisées qui soient fiables et que vous pouvez vraiment comparer. Il y a des logiciels qui permettent de le faire. Mais ça reste quand même votre décision. C'est à vous de dire à partir de quel moment vous considérez une donnée comme fiable ou pas. Ça relève du choix de l'utilisateur – même s'il y a des systèmes d'auto-apprentissage qui vont sélectionner votre donnée. Moi, j'ai malheureusement déjà vingt-cinq ou trente ans d'expérience et j'essaie toujours d'attirer l'attention des jeunes (des doctorants, des post-docs, des jeunes chercheurs) sur l'importance de prendre du recul par rapport aux données qu'ils génèrent. Notamment de prendre du recul par rapport à la force d'une image ou d'une vidéo (parce que vous avez beaucoup de données qui sont stockées sous forme d'image ou de vidéo). Souvent, j'observe que les jeunes ont tendance à croire qu'une image est vraie, si elle est belle. En fait, il n'y a pas de relation entre la véracité d'une image et sa beauté artistique. Évidemment, l'œil est attiré par une image. Vous allez être beaucoup plus impressionnés par une image ou une vidéo que par du texte. Mais ça n'est pas pour autant que cette donnée est vraie. Donc il faut se détacher de ce genre de données et, pour cela, il faut de l'expérience. Il faut avoir subi des échecs, pour se dire « okay, là, il y a une jolie image, c'est super ; mais ça n'est pas pour autant que c'est vrai ». Il

faut faire attention. Et ce n’est qu'avec l'expérience qu'on l'acquiert. Ça n’est pas facile, parce que maintenant vous avez des outils, vous avez des données d'imagerie/de microscopie électronique, des trucs extraordinaires. Ça n’est pas pour autant que c'est pertinent pour le problème qui vous intéresse. Pas forcément. Donc il faut toujours se replacer dans le contexte : « est-ce que cette image m'apporte une information ? Est-ce que c'est vérifié, quantifié, reproductible dans le contexte qui m'intéresse ? ». Et ça n’est pas toujours évident.

14 - 1:09:44 > 1:13:09 Méthode de travail

[>R1]: Merci beaucoup !

[>R2]: Je vous en prie. J'espère que j'ai fait avancer votre travail.

[>R1]: Oui. C'était vraiment intéressant de voir comment fonctionne un projet avec de la valorisation économique derrière.

[>R2]: Disons que ça nous oblige à avoir des pratiques légèrement différentes des pratiques académiques. C'est-à-dire qu’on va travailler quasiment comme une start-up. Notamment, avec tout ce qui est « reporting » : la manière dont on va communiquer sur nos résultats, la manière dont on va planifier nos réunions... Tout ça se professionnalise, comme on le ferait si on était dans une start-up, avec les mêmes pratiques de communication (pas simplement vers l'extérieur mais aussi en interne). Ça ne se fait plus de manière empirique, comme on le faisait dans les labos il y a vingt ans. Ça se rationalise. C'est de la gestion de projet, qui se professionnalise, qui n'est plus académique. On utilise les mêmes méthodes de management de projet que dans une société.

[>R1]: Pour quelle raison ? Parce qu'il y a des enjeux économiques ?

[>R2]: Oui, aussi. Et parce qu'à partir du moment où on va vouloir valoriser ces résultats, qu'on va trouver un partenaire industriel qui s'intéresse à ces résultats, il va falloir que l'industriel soit convaincu des données que nous avons générées. Ça suppose de mettre en place un certain nombre de pratiques (notamment l'archivage des données). Si les données sont ouvertes à tous les vents, l’industriel va vous dire : « bon, vous êtes bien gentils mais au revoir ». C’est du « reporting ». Est-ce que les données sont bien archivées ? Est-ce qu'aucune étape n’a été manquée ? Donc vous avez un cahier des charges qui vous impose tout ça en

matière de communication en interne. Ça permet que le jour où vous voulez valoriser vos résultats, ils soient immédiatement valorisables. C'est plus facile si vous avez utilisé les mêmes techniques que la personne qui va acheter votre brevet ou votre molécule. Donc ça nous oblige à une communication en interne qui est légèrement différente. En fait, c'est de la professionnalisation de la communication en interne. C'est du management de projet, tout simplement. Alors qu'il y a trente ans on faisait ça au petit bonheur la chance. Là, ça nous pousse à utiliser des méthodes de travail/de management de projet qui sont peut-être légèrement différentes d'un projet académique pur et dur, où vous vous diriez « bon, ça n'est pas grave ».

[>R1]: Oui, ça implique plus de rigueur et de procédures.

[>R2]: Un peu plus. Après, c'est aussi plus facile, puisque vous vous dîtes : « s'il faut que je revienne sur ce travail deux ans plus tard, si l'information a été bien stockée, bien archivée, je la retrouverai beaucoup plus facilement ».

[>R1]: Oui, vous y trouvez un intérêt aussi.

[>R2]: Bien sûr. On a intérêt à le faire, parce que les projets vivent tout de même plusieurs années. Si vous engagez un post-doc et qu'il veut consulter les données générées par son collègue il y a trois ans, il faut évidemment qu'il ait accès aux données, que ça soit clair, net et précis et qu'il n'ait pas besoin d'aller chercher dix tomes dans une salle poussiéreuse pour vérifier l'information qui l'intéresse. Donc c'est bénéfique pour tout le projet.

Entretien avec le chercheur 16 (éthologie)

1 - 00:00 > 08:53 Objectifs scientifiques du projet

[>R1]: Est-ce que vous pourriez me résumer le projet de recherche H. ?

[>R2]: Ah, vous voulez parler spécifiquement de H. ! Okay. Parce que moi je travaille sur différents aspects : sur les réseaux de manière générale et, au sein de ma recherche sur les réseaux sociaux, je travaille sur les insectes, les primates et les humains. Là, du coup, je ne vais vous parler que des humains, mais après je pourrai vous parler du reste, si vous voulez. Le projet H. c'est essayer de comprendre le bien-être des personnes âgées : qu'est-ce qui va influencer leur bien-être en fonction de là où elles habitent (i.e. le type d'urbanisation), comment ce type d'urbanisation va influencer leur réseau et comment le réseau va influencer le bien-être à la fois physique et mental des personnes âgées. Donc c'est essayer de comprendre un peu comment la coévolution de l'urbanisation et de tout ce qui va avec la mobilité va influencer le réseau social et comment cet ensemble va influencer le bien-être des personnes.

[>R1]: Ce sont les réseaux « de type social » que vous étudiez ?

[>R2]: Oui, ça n'est pas le média social. Pour les personnes âgées, on a quand même rajouté dans le questionnaire quelques questions sur tout ce qui est e-santé (les personnes ont-elles accès à des systèmes d'e-santé via des tablettes, des logiciels spécifiques... ?).

[>R1]: Comment procédez-vous pour étudier l'influence du type d'urbanisation et du réseau social ?

[>R2]: En fait, il y a une cohorte qui existe à Paris (on travaille aussi avec Montréal et Luxembourg). La cohorte existe depuis peut-être 20 ans. Ces personnes ont déjà répondu à différentes vagues d'étude. Là on va les recontacter pour savoir lesquelles d'entre elles veulent participer à l'étude. On va sélectionner environ 400 à 500 personnes d'un certain âge dans cette cohorte. D'abord on les appelle. Ensuite on va les voir. Puis on leur demande si elles sont d'accord pour participer et si les membres de leur foyer sont également d'accord pour participer. En fonction de ça, elles sont intégrées ou non dans l'étude.

[>R1]: Comment s'appelle la cohorte ?

[>R2]: R-----.

[>R1]: Elle fait partie d'un dispositif particulier ?

[>R2]: C'est l'équipe Inserm N. à Paris, qui est responsable de cette cohorte. C. en est le coordinateur.

[>R1]: Que faites-vous faire aux personnes qui acceptent de participer ?

[>R2]: Elles répondent à deux questionnaires. Un premier questionnaire portant sur leurs habitudes (alimentaires, etc.), sur leur quartier de vie, sur leur ressenti par rapport à leur quartier, sur leur santé physique et mentale, sur ce qu'elles font tous les jours, sur leurs déplacements, sur le nombre de personnes qu'elles voient... Ensuite, pendant 7 jours, elles vont porter des capteurs, de même que les membres de leur foyer. Ce sont des capteurs GPS, accélérométrie et de proximité. Les chats et chiens sont aussi équipés. Pendant ces 7 jours, les personnes tiennent également un carnet de bord, afin qu'on puisse recouper ce qui s'est passé avec le GPS (il y a parfois des petits soucis au niveau du GPS). Ensuite on donne aux personnes un second questionnaire, qui est un questionnaire veritas, dans lequel les personnes vont décrire les trajets qu'elles ont faits durant cette semaine ou de manière générale. Ça nous permet de voir un peu dans quels endroits elles vont, comment elles y vont, avec qui, qui elles rencontrent là-bas... Ça nous permet de lier vraiment le réseau spatial au réseau social.

[>R1]: En quoi consiste l'accélérométrie ?

[>R2]: L'accélérométrie ce sont des capteurs de niveau et de mouvement – un peu comme une manette Wii. C'est pour savoir comment sont placées les personnes et quel type d'activité elles font : est-ce qu'elle est en train de se déplacer, et si oui, est-ce qu'elle marche, est-ce qu'elle court, est-ce qu'elle fait du vélo... ? On connaît un peu tout ça. Si elle est chez elle, on peut savoir comment elle est : est-ce qu'elle est assise, est-ce qu'elle est debout, allongée, combien de temps par jour ? Donc tout ça va nous permettre de connaître toute l'activité des personnes et de lier la mobilité à ce type d'activité qu'elles vont faire tous les jours.

[>R1]: Quel était le troisième type de capteur ?

[>R2]: Des enregistreurs de proximité. Pour savoir à quel moment les personnes sont à proximité des membres du foyer et donc savoir si elles les voient souvent. Ça nous permet de

voir si le réseau social qu'on mesure via le questionnaire est comparable au réseau social qui va être mesuré durant la semaine. Car, en général, ce que disent les personnes, quand elles répondent à un questionnaire, ne correspond pas forcément au nombre de fois où elles ont réellement vu un proche durant la semaine. Donc ça nous permet de corriger un petit peu.

[>R1]: A quoi sert le carnet de bord ?

[>R2]: Quand les individus se déplacent avec le GPS, il arrive qu'il y ait des erreurs avec l'appareil. Parce que ça passe sous un pont, parce que la couverture dans le bus est mal faite... Ou bien le GPS va dire que les personnes sont à tel endroit, mais ça n'est pas exactement le bon endroit. Ou bien les personnes vont enlever le GPS pendant deux heures. Du coup, dans ces cas-là, on va leur demander pourquoi elles ont enlevé le GPS. Le carnet de bord permet de savoir exactement ce qu'elles ont fait chaque jour. Quand il y a une faute, on va pouvoir regarder ce carnet de bord (au lieu de rappeler les personnes à chaque fois). Par exemple, une personne va dire qu'elle était à Uniqlo, alors que le GPS indique qu'elle était à Printemps (les deux magasins sont à côté). Ça n'est pas la même chose, vous voyez. Donc là on va pouvoir corriger et dire : « le GPS a dit qu'elle était à Printemps ; ça n'était pas Printemps, c'était à Uniqlo ». Et quand on va voir que la personne ne porte plus le GPS pendant deux heures, elle va avoir marqué : « j'étais à la piscine pendant deux heures ». On sait qu'on va pouvoir ajouter une période de deux heures, pendant laquelle la personne était à la piscine et n'a pas pu porter l'appareil.

[>R1]: Je comprends. En fait, ce sont les personnes qui remplissent le carnet de bord ?

[>R2]: Oui.

[>R1]: Quel type de données recueillez-vous par le biais des questionnaires : des données quantitatives ou qualitatives ?

[>R2]: Les deux. Il va y avoir des oui et des non, et puis il va y avoir des sortes d'échelles permettant de mesurer de 1 à 5 et de faire un score sur la dépression par exemple. Et puis parfois ça va être des questions quantitatives, du style : combien de verres d'alcool buvez-vous par semaine, combien de cigarettes fumez-vous par jour, etc. ?

2 - 08:53 > 16:21 Répartition des rôles entre les partenaires du projet

[>R1]: Sur le site de l'ANR, j'ai vu qu'il y avait quatre laboratoires associés sur le projet et, sur le site d'un des partenaires, j'ai vu qu'il y avait aussi Montréal. Est-ce que Montréal fait partie du projet ANR ?

[>R2]: Il est dedans, oui. Ils sont associés avec Luxembourg. Ils ne reçoivent pas d'argent, mais ils sont associés. Normalement, le but ultime de l'ANR était d'avoir les mêmes résultats ou protocoles sur une cohorte à Montréal, une cohorte à Luxembourg et une cohorte à Paris et d'essayer de comparer les trois résultats pour ce qui était similaire et ce qui était différent.

[>R1]: C'est effectivement ce qui va se passer ?

[>R2]: Oui, si on arrive à avoir toutes les données qu'il faut et à les comparer. Parce qu'il y a ce qu'on veut faire dans l'idéal et ce qu'on arrive à faire à la fin !

[>R1]: Comment vous répartissez-vous les rôles entre laboratoires ? Vous dites que chacun a une cohorte. Est-ce que vous avez exactement les mêmes questions et les mêmes dispositifs ?

[>R2]: Ça a été assez difficile justement, parce que pour moi c'était la première fois que je travaillais sur ce type d'étude. C'était la première fois que je faisais de l'humain. Et je pensais par exemple que les questionnaires étaient très rodés, très précis, qu'ensuite on n'avait pas besoin de les retravailler, que le questionnaire qui avait été fait à Montréal était le même que celui de Luxembourg et que, du coup, on aurait réutilisé le même. En fait, pas du tout. On a refait tout un questionnaire. Parce que les personnes avec qui je travaillais considéraient que certaines questions posées à Montréal n'étaient pas forcément applicables à Paris et que, nous, on avait des petites questions en plus. C'est là où ça devient compliqué, parce que, du coup, ça n'est pas totalement comparable. C'est à nous de chercher dans les données ce qui peut être comparable ou pas. Et on a passé beaucoup de temps à faire ça. C'était assez incroyable.

[>R1]: Vous êtes partis des questionnaires déjà réalisés ?

[>R2]: Oui. On a repris les questionnaires de Montréal et Luxembourg. On a regardé les questions qui étaient posées et on en a repris certaines, d'autres ont été modifiées, d'autres ont été rajoutées, certaines ont été supprimées. Mais c'est vrai que, moi personnellement, ça m'a étonné, parce que je trouvais qu'on aurait pu être plus efficace dans ce travail-là.

[>R1]: Est-ce qu'il y aura une phase de comparaison à la fin du projet ? Ou bien est-ce que vous échangez déjà certains résultats entre laboratoires ?

[>R2]: Pas encore. Je pense qu'on le fera à la fin. On réutilisera tout ce qui a été fait à droite et à gauche. Par exemple, tout ce qui a déjà été fait sur les réseaux sociaux à Montréal sera beaucoup plus facilement applicable à Paris, parce qu'on n'aura pas toute la méthodologie à refaire. On pourra refaire directement ce qui a été fait et appliquer la même méthodologie.

[>R1]: C'est-à-dire la méthodologie d'analyse ?

[>R2]: Oui. Là on a développé des réseaux multimodaux : il y a le réseau social, qui est « comment les personnes sont connectées entre elles », et il y a le réseau spatial, qui est « comment les différents endroits où va la personne sont connectés entre eux » (par le train, etc.). Donc il y a ce qu'on appelle les réseaux bipartites : on lie le réseau spatial au réseau social, en regardant comment la personne se connecte à ses amis grâce au réseau spatial. Grâce aux capteurs et aux questionnaires, on essaie de voir où les personnes sont allées, qui les y a conduites et qui elles ont rencontré à cet endroit. On arrive comme ça à reconstruire tous les réseaux.

[>R1]: Vous dites que vous représentez le résultat sous forme de graphes. Quel logiciel utilisez-vous pour analyser les données des capteurs et des questionnaires ?

[>R2]: Avec les capteurs, tous les jours vous allez recevoir un fichier qui vous dira : « de tel moment à tel moment, les deux personnes étaient en contact ou face à face ». Si c'est bien fait, vous pouvez facilement mettre ça dans un logiciel, qui va analyser et sortir le réseau de la semaine. Après, c'est à vous de mettre des filtres, en disant « je ne veux relever que les périodes de plus de tant de temps » ou « je veux relever les émissions qui étaient de plus de telle intensité ». Et puis c'est à vous de construire le réseau. Ensuite, avec le GPS, vous pouvez regarder où les personnes se sont rencontrées. Là c'est plus difficile, parce que les deux appareils sont déconnectés. Ça veut dire qu'à un moment donné vous allez avoir le temps de telle personne, qui a rencontré telle autre personne, et vous allez voir sur le GPS qu'au même moment la personne était à la piscine. Du coup, vous pouvez essayer de recouper. Mais c'est à vous de le faire dans des logiciels – dans R par exemple.

[>R1]: D'accord, vous utilisez R. Est-ce que vous utilisez d'autres logiciels ?

[>R2]: Non. Maintenant c'est surtout R qui est utilisé. Parce que même les logiciels qui permettaient de mesurer le type d'activité que les gens faisaient n'étaient pas assez précis. L'accéléromètre va vous donner différents signaux. En fonction de ces différents signaux, vous en déduirez : « ça c'était de la marche, ça c'était de la course, etc. ». Les logiciels n'étant pas suffisamment précis, le doctorant, avec qui je travaille, a dû tout redévelopper lui-même dans R, pour avoir des positions spécifiques, qui n'existaient pas dans le logiciel d'origine de l'accéléromètre, et pour avoir des résultats plus fins. Il y a beaucoup de packages qui existent sous R. R est un langage de programmation. Avant c'est plus Python qui était utilisé. Moi j'utilisais NetLogo pour faire de la modélisation de systèmes multi-agents. C'est vrai que maintenant on peut tout faire avec R. S'il nous manque quelque chose, le package existe, on le remet dedans, on prend la ligne de code... C'est l'open access qui permet ça. C'est la gratuité de l'open access qui permet le développement incroyable de ce type de logiciel.

3 - 16:21 > 22:35 Ressources humaines

[>R1]: Combien de personnes travaillent sur le projet H., ici, dans cette équipe ?

[>R2]: Ici, il y a moi. Je suis seul. Il y a B. qui s'est rajoutée dessus. C'est une autre chercheuse CNRS. Après, au niveau de l'Université, il y a deux doctorants qui travaillent ici : I., qui est ici, et A., qui est Québécois et qui fait sa thèse entre Strasbourg et Montréal. Lui a vraiment développé le double aspect. Avec A. je travaille sur les données de Montréal. Avec I., je travaille sur des données de R-----, donc de Paris. Après, il y a un autre doctorant mais en sociologie : H. Lui travaille sur l'e-santé. Il m'a contacté, car il voulait avoir un peu mon aide. Donc on travaille aussi ensemble sur le sujet. Il est à D., je crois. Son encadrant c'est Z.

[>R1]: Sur quoi porte la thèse de I. à Strasbourg ?

[>R2]: Au début il devait analyser les données du projet. Mais les données ont été prises trop tard. Donc là il analyse d'autres données sur la même cohorte. C'est un peu similaire, sauf qu'il n'y a pas l'aspect réseau social. Isaac travaille vraiment sur l'accélérométrie, donc sur l'activité des gens. Il a à la fois des indices de santé (est-ce que les personnes sont en diabète/en surpoids, comment elles fument, comment elles boivent…), des données sociodémographiques (où les personnes vivent, quel est leur métier…) et les données de l'accélérométrie. Sa première question de recherche consiste à essayer de voir comment

l'activité physique influence la santé et quels sont les liens directs ou indirects. La deuxième question de recherche va être justement de comprendre comment la sociodémographie influence l'activité des personnes (il y a des choses toutes simples comme « plus il y a d'éducation, plus on passe de temps assis »).

[>R1]: Donc vous réutilisez des données de la cohorte, établies par l'Inserm ?

[>R2]: Oui. Normalement ça n'est pas ce qu'Isaac aurait dû faire. Normalement il aurait dû travailler sur les données qu'on aurait dû prendre et qui n'ont pas été prises ou qui sont en train d'être prises. Du coup, ça aurait été trop tard pour sa soutenance. Donc on lui a donné d'autres données à analyser. Moi, quand je prends des doctorants, j'essaie toujours d'avoir une roue de secours.

[>R1]: Vous disiez que vous travailliez sur les données de Montréal avec A. En quoi consiste le travail de A. ?

[>R2]: A. traite l'aspect réseau social, c'est-à-dire qui la personne voit, où elle va, comment elle y va, avec qui, qui elle voit là-bas. Et là on lie vraiment l'aspect social à l'aspect spatial. On voit que la personne est liée à différents lieux et que dans ces lieux elle arrive à voir telle ou telle personne. On voit que des sous-groupes se créent. Par exemple, une personne va voir telle personne à tel endroit, mais elle va voir telle autre personne à tel endroit. Et puis, il y a des personnes qui ne voient pas beaucoup de monde. Parce qu'elles habitent à un endroit, où elles ne peuvent pas beaucoup se mouvoir. Donc elles ne voient que leur famille. On voit qu'ici il y a des réseaux qui sont pauvres, on voit qu'il y a des réseaux en clusters, et puis on voit des réseaux, où la personne est avec beaucoup de monde et qu'en plus tous ces gens-là sont connectés entre eux. On a développé toute une méthodologie autour de ça, pour essayer de comprendre un peu comment l'analyser. Parce que, pour l'instant, ça n'avait pas été fait (ça avait été proposé, mais ça n'avait pas encore été analysé). Donc, nous, on est en train de l'analyser. Ensuite on essaiera de comparer tout ça à la santé mentale et physique des personnes âgées. Pour l'instant, A. n'a fait sa thèse que sur l'aspect social, pour voir comment les gens sont connectés et quel type d'indice on pourrait en extraire. On n'a pas encore comparé ces données avec les indices de santé. On va le faire là.

[>R1]: A. applique cette méthodologie aux données de Montréal et aux données de Paris ?

[>R2]: Non, il ne travaille que sur Montréal, I. que sur Paris et il y a une doctorante qui ne travaille que sur Luxembourg. Pour l'instant, on n'a personne pour comparer les données des différentes villes. C'est trop tôt, trop complexe. Mais il faudrait que, par la suite, il y ait quelqu'un qui traite différents aspects. On ne pourra pas tout « merger ». Il faudra prendre, par exemple, seulement l'aspect social et essayer de voir comment ça se passe entre les trois villes. Ou alors prendre seulement l'aspect santé. Ce qui a été fait peut continuer à l'être pendant au moins 5 ans. Toutes les données sont là, on peut les analyser sans aucun souci. On a encore des questions qu'on pourrait traiter.

4 - 22:35 > 27:17 Confidentialité des données

[>R1]: Vous-mêmes, avez-vous accès aux données de Montréal ? Via un espace informatique partagé, par exemple ?

[>R2]: Alexandre a les données sur son ordinateur. Sauf que les personnes sont codées avec des numéros, donc on ne sait pas qui elles sont. On pourrait savoir, parce que le lieu de vie est quand même assez bien identifié. Quand c'est en ville, dans un immeuble, c'est encore peu localisable. Mais c'est quand même indiqué parce que, si c'est une personne âgée, on veut savoir à quel étage elle vit, si elle a un ascenseur ou des escaliers... On pose tellement de questions, pour savoir comment la personne vit, qu'on pourrait facilement retrouver de qui il s'agit. Bon, on s'en fout de qui c'est. Mais là, par exemple, A. compare des personnes qui habitent dans le centre-ville avec des personnes qui habitent dans des campagnes. Dans les campagnes, du coup, c'est très localisé. Les maisons sont assez espacées, donc, s'il y a un point, on peut voir à peu près de quelle maison il s'agit.

[>R1]: Est-ce qu'au Canada il existe un organisme du type de la CNIL ?

[>R2]: Oui. Je pense qu'il doit y avoir ça à peu près partout.

[>R1]: Pour les données de Paris par exemple, est-ce que la CNIL vous demande autre chose que l'anonymisation des participants ?

[>R2]: Par exemple, vous ne pouvez pas avoir l'adresse exacte dans le fichier. On peut avoir le nom de la personne et son adresse exacte dans le fichier d'identification, avec le numéro [d'anonymisation] correspondant. Mais, dans le questionnaire, on ne peut pas demander

l'adresse exacte de la personne. On peut seulement demander à la personne si elle habite dans un immeuble et, si oui, à quel étage. Bon, après, on a le GPS... Mais les données du GPS et les réponses au premier et deuxième questionnaires sont séparées physiquement. Après, il y a l'aspect carnet de bord. Donc c'est vrai qu'on pourrait retrouver la personne. On pourrait recouper. Par contre, dans la demande CNIL, on est obligé de dire que le nom et l'adresse précise de la personne sont stockés dans un fichier séparé, qui donne un numéro de code et que ce numéro de code ne figure qu'ici et là. On doit dire également que dans le questionnaire on ne connaît que l'étage et que les données du GPS nous donneront l'endroit précis où la personne vit, mais pas l'étage. Donc on sépare aussi ces fichiers-là.

[>R1]: Est-ce que la CNIL exige un certain niveau de sécurité ?

[>R2]: On ne peut pas partager les fichiers par internet, par exemple. Normalement, tout est sur CD ou sur clé, lesquels sont enfermés dans un placard à code, auquel seul C., le responsable de la cohorte, a accès. Quand on va chercher les données, C. nous donne le fichier. On a donc ensuite le fichier sur notre ordinateur, mais on ne peut pas partager ce fichier par mail.

[>R1]: Vous êtes obligés d'aller à Paris pour récupérer les données ?

[>R2]: Oui. Moi je n'ai pas le fichier sur mon ordinateur. C'est I. qui l'a. Si je veux avoir accès aux données, il faut que je lui demande, avec normalement l'aval de C. Alors, on pourrait ne pas respecter ces recommandations. C. pourrait nous envoyer le fichier par mail ou je pourrais aller le chercher auprès d'I. sans demander à C. Ce sont des choses qui sont faciles à ne pas respecter. On les respecte quand même, parce que le jour où il y a une fuite, au moins on peut montrer que ça a été fait de manière très précise (en espérant que ça n'arrive jamais).

[>R1]: La CNIL vous demande-t-elle de supprimer les données au terme du projet ?

[>R2]: On peut garder les données pendant cinq à dix ans, si d'autres travaux sont ré-effectués dessus. C'est une période qui est renouvelable, c'est-à-dire que, si un nouveau travail est effectué sur ces données, on peut prolonger à nouveau la conservation. Les données devront toujours être gardées par une seule personne. Et de notre côté, dès qu'on n'aura plus besoin de ces fichiers-là, on devra les supprimer de notre ordinateur.

5 - 27:17 > 29:43 Contraintes de gestion des données

[>R1]: Le financement ANR se termine fin 2018, il me semble.

[>R2]: Oui, on doit prolonger d'un an. Mais on a demandé à ce que les données soient stockées pendant dix ans, pour qu'on puisse travailler à nouveau dessus.

[>R1]: Pensez-vous effectuer une nouvelle demande de financement ensuite ?

[>R2]: Il faut qu'on voit… Pas nécessairement. A moins que ça ne soit pour des doctorants – pour qu'ils puissent travailler sur les données. Mais si la prise de données est faite, ça ne nous coûte plus tellement cher. Ce qui coûte cher, c'est la prise de données, le matériel, les gens qui prennent les données (les enquêteurs), le système informatique pour accéder aux données et les analyser... Mais ensuite, une fois que les données sont là, ça ne nous coûte rien (excepté du temps) de les analyser.

[>R1]: Les doctorants ont-ils des contrats doctoraux ?

[>R2]: Oui, tous. C'est très précis chez nous en sciences. Ailleurs, ça n'est pas forcément le cas (en sociologie, par exemple). Ici, à l'ED 414, il faut que ça soit un contrat (ou une bourse pour des personnes étrangères), d'un minimum de 1400€ par mois pendant 3 ans.

[>R1]: Les doctorants ont-ils une contrainte au niveau des données ? Doivent-ils les céder au laboratoire au terme de leur thèse ?

[>R2]: Rien ne leur appartient. Dans la charte des doctorants, il y a une signature de confidentialité, qui dit que rien ne leur appartient. De notre côté, si on publie sur quelque chose qu'ils ont récolté, on doit publier avec leur nom. Mais eux non plus ne peuvent pas publier sans demander notre accord et sans mentionner le CNRS comme tutelle.

[>R1]: Utilisez-vous des règles de nommage de fichiers ou une arborescence ?

[>R2]: Normalement, il y a un seul fichier avec toutes les données. Après, on en extrait ce qu'on veut et on se crée un autre fichier à côté.

6 - 29:43 > 31:29 Définition d'une donnée scientifique

[>R1]: Je passe à la partie « Données ». J'ai une question un peu plus théorique. Utilisez-vous spontanément le terme « données » ? Si oui, comment le définiriez-vous ? Que désigneriez-vous sous ce terme, si on prenait l'exemple du projet H. ?

[>R2]: Une donnée c'est une information qui a été récoltée sur un participant, je dirais. Voilà, tout simplement. Après, cette information est de différentes natures. Quand on dit « les données », c'est le fichier. Mais chaque case de ce fichier est pour nous une donnée, puisque les lignes correspondent aux différents participants et les colonnes aux différentes questions qui ont été posées. Il y a donc dans ce fichier autant de données qu'il y a de personnes, qu'il y a de questions.

[>R1]: Ce serait la même chose pour un projet sur des primates ?

[>R2]: Oui, c'est pareil. Par contre, la mentalité sur le partage des données n'est pas la même.

7 - 31:29 > 43:52 Ouverture des données

[>R1]: Comment concevez-vous le partage des données ?

[>R2]: Il y a la mentalité de chacun, qui est : quand on va publier, est-ce qu'on va partager les données ? On n'est pas obligé de partager les données brutes. On peut donner le fichier qui nous a servi à l'analyse par exemple. Ça c'est quelque chose que je fais presque spontanément maintenant, mais qui est de toute façon de plus en plus demandé pour les bons papiers – surtout chez les humains (moins chez les primates). Si on publie dans des journaux plus généralistes comme *Proceedings of the Royal Society*, on va nous demander le fichier. Moi je travaille aussi sur les piétons (comment les piétons traversent, pourquoi ils traversent, quelle est leur prise de décision en fonction de leur culture, de leur genre, de leur âge…) et là c'est pareil, on a déposé les fichiers. Mais on n'a pas mis le fichier global ; on a seulement mis le fichier qui nous a servi à analyser les données. Je fais aussi de la modélisation. Quand je développe un modèle, naturellement je mets le code ou la modélisation en ligne.

[>R1]: Oui, c'est une culture en informatique.

[>R2]: Nous, les primatologues (mais c'est pareil ici pour ceux qui travaillent sur les humains), on ne va pas partager le fichier de données global. On va attendre cinq ou dix ans,

le temps de travailler dessus. Et après on va le partager. Mais c'est vrai que je travaille sur des méta-analyses, où on fait des analyses de différents groupes de primates (90 groupes par exemple). Là c'est impossible de tout récolter soi-même. Parce qu'il faut aller sur le terrain pour étudier chaque groupe de primate (il faut au minimum 3 à 6 mois pour étudier un groupe de primates). Du coup, on va demander aux gens si on peut avoir accès à leurs données. Généralement ils nous disent non. Certains disent oui. Moi j'arrive à partager certaines données. C'est ce que j'ai fait pour mes données de thèse par exemple. Mais ça a été assez difficile, parce que mes encadrants ne voulaient pas forcément.

[>R1]: Pour quelle raison ?

[>R2]: Quand ils ont passé trois mois à observer un groupe de primates, ils ne veulent pas forcément que leurs données soient récoltées comme ça et que des gens, qui sont derrière leur ordinateur, puissent travailler dessus sans mentionner leur nom. Nous, quand on fait une méta-analyse comme ça et qu'on décide de publier, on prend toutes les personnes qui ont participé et on leur demande quels noms on doit faire figurer. Donc on va faire des publications à 20 auteurs par exemple. Ça n'est pas autant qu'en physique, où ils peuvent avoir mille co-auteurs, mais 20 auteurs c'est déjà beaucoup en primatologie.

[>R1]: D'où proviennent les données des 90 groupes de primates ?

[>R2]: Généralement ce sont des doctorants, qui ont acquis ces données sur le terrain. Ils ont publié sur le sujet, mais n'ont pas ajouté le fichier de données original à leur publication. Parce qu'ils veulent en garder l'exclusivité. C'est une mentalité un peu différente de la mienne. Mais ça dépend aussi de la difficulté que les gens ont à publier. Moi je suis plus dans un système de collaboration, où on peut publier beaucoup, parce qu'on collabore tous ensemble et qu'on s'échange les données. Les gens qui partagent moins, du coup, ont un peu plus de mal à publier aussi. Moi j'aime bien le fait de partager les données, parce que souvent les gens reviennent vers vous (bien que les données aient été publiées). Parce que ce ne sont pas exclusivement les données qui les intéressent ; c'est aussi votre expertise. Moi ça ne me dérange pas de partager mes données, parce que je n'en ai pas forcément besoin pour publier. Mais je pense qu'il y a des gens qui ont besoin de garder l'exclusivité de leurs données – surtout les doctorants et les jeunes chercheurs. Parce qu'ils vont avoir ce besoin de publier derrière. C'est aussi aux chercheurs plus avancés de reconnaître qu'il faut impliquer les jeunes

chercheurs dans davantage de projets. Donc c'est toute une mentalité qu'il peut y avoir sur l'open access, en primatologie en tout cas. Parce que c'est vrai qu'en informatique c'est différent. En primatologie ou en sociologie, on ne partage pas forcément tout de suite les données, tandis qu'en informatique les modèles sont tout de suite mis à disposition.

[>R1]: Selon vous, d'où vous vient cette mentalité du partage, de la collaboration ?

[>R2]: C'est personnel, c'est-à-dire que c'est ma mentalité qui est un peu comme ça. Après, j'ai eu une expérience au niveau de la thèse, qui a été assez prolifique. En fait, nous étions deux thésards à travailler sur le même sujet de thèse. Lui était en Angleterre, moi en France. On ne travaillait pas sur la même espèce, mais le sujet de thèse était exactement le même. Donc on s'est dit : soit on est compétition et on va publier l'un contre l'autre ; soit on collabore et on travaille ensemble. On s'est mis à travailler ensemble et là on a fait des « special issues » dans les journaux, on a publié des papiers qui ont été très cités. En fait, on a explosé le nombre de nos publications. Cette expérience a suffit à me convaincre qu'il faut collaborer avec les gens. Ensuite vous acquérez une certaine réputation. Le système de réputation va assez vite : dans un petit monde comme la primatologie, où il n'y a pas beaucoup de chercheurs, si vous êtes reconnus comme quelqu'un qui collabore et qui fait du très bon travail, les gens vont vouloir collaborer avec vous ; si vous êtes connus comme quelqu'un qui ne veut pas collaborer, qui s'approprie certaines idées, voire qui pique parfois les idées des autres, personne ne va collaborer avec vous. Ça se sait très, très vite. Le système de réputation est assez important dans le partage de données, je trouve. C'est tout un ensemble, qui est assez intéressant. Ça n'est pas juste partager les données pour partager les données. Il y a toute une mentalité derrière, qui est concomitante avec le partage de données. La modélisation, que j'ai faite durant ma thèse sur les primates, a ensuite été réutilisée par d'autres personnes, qui ont utilisé mes formules chez les humains et qui m'ont posé des questions auxquelles j'ai répondu. Du coup, c'est intéressant, parce que vous n'êtes pas forcément co-auteur de ces publications mais vous êtes cités. Et ensuite les personnes vous connaissent. Ça veut dire que, quand elles ont besoin d'organiser un symposium, elles vont faire appel à vous. Donc vous n'êtes pas récompensés tout de suite, mais un peu plus tard.

[>R1]: C'est peut-être aussi plus facile de partager, parce qu'il n'y a pas d'enjeux économiques ? Les données, dans votre domaine, n'ont pas de valeur économique ?

[>R2]: Je connais beaucoup de collègues allemands qui n’ont pas de position fixe. En Allemagne, les positions fixes sont très rares.

[>R1]: Qu’entendez-vous par « position » ?

[>R2]: Les postes permanents. Ici, en fait, on a très peu de post-docs mais, dès qu'on est un bon chercheur, on peut facilement avoir une position au CNRS ou dans une université. Même si, avec la crise, ça devient de plus en plus difficile. Alors qu'en Allemagne ils vont avoir plus de post-docs (ils ont des sous pour des post-docs, on peut candidater beaucoup de fois à un post-doc), mais je connais des gens qui ont 45 ans et qui n'ont pas encore de position fixe. Donc il y a un enjeu économique derrière, mais pour eux, pour leur propre vie. C'est-à-dire qu'il faut qu'ils publient un maximum. Du coup, ils sont un peu moins partageurs sur certaines données. On sent qu'il y a un enjeu personnel pour eux. Alors que quand on a une position permanente, on n'a plus forcément cet enjeu personnel.

[>R1]: Je pensais plutôt aux domaines, où les données vont ensuite être rachetées. J'ai rencontré un chercheur qui écrivait des codes informatiques. Il disait que, pour lui, ça n'était pas possible de partager le code, parce qu’il avait une valeur économique. Il allait par la suite vendre ce code à des entreprises. Dans votre domaine, ce problème ne se pose pas ?

[>R2]: Non. Mais je pense qu'il y a les deux aspects en informatique : soit vous développez des codes, vous les vendez ensuite et vous faites votre propre argent ; soit vous développez des codes, vous les mettez en ligne et les gens voient que c'est vous. Ils peuvent se les approprier, mais ils ne sauront pas les redévelopper eux-mêmes ou en faire autre chose. Donc ils vous appellent, parce qu'ils veulent vous embaucher. En fait, je pense qu’en informatique il y a les deux aspects. J'ai un copain qui est informaticien. Il mettait toujours tout en open access (il a toujours tout fait comme ça). Et c'est parce que les gens voyaient ce qu'il faisait en open access, qu'ils décidaient de l'embaucher directement derrière. Parce qu'ils savaient la valeur qu'il avait et ils voulaient absolument cette personne.

[>R1]: D'accord, ça donne de la visibilité.

[>R2]: Voilà.

8 - 43:52 > 47:08 Collaboration scientifique

[>R1]: Vous me disiez que ça vous était arrivé de demander des données à d'autres chercheurs. Est-ce qu'on vous a déjà refusé l'accès à des données ?

[>R2]: Oui. La personne disait qu'on travaillait sur le même sujet et qu'on était en compétition. Cash ! Alors que, pour moi, partager les données, parce qu'on travaillait sur le même sujet, nous permettait d'écrire plus d'articles ensemble et d'être plus productifs. Mais la personne ne voyait les choses de la même manière. C'était un Allemand.

[>R1]: Qui n'avait pas encore de poste ?

[>R2]: Oui. Je pense que c'est très important, ça. Donc il y a des gens qui vous répondent « non, on travaille sur le même sujet, on est en compétition ».

[>R1]: Avez-vous eu d'autres motifs de refus ?

[>R2]: Non. Parfois on nous disait « on ne veut plus participer à cette étude », mais sans forcément de motif particulier. Après, c'est intéressant. Par exemple, là ça fait un moment qu'on travaille sur les réseaux sociaux chez les macaques, on est un noyau dur de quatre ou cinq personnes et ensuite il y a les jeunes chercheurs qui gravitent autour. Et parce qu'on travaille souvent ensemble, on s'est dit qu'il fallait qu'on fasse un consortium, donc qu'on continue à travailler systématiquement tous ensemble et qu'on multiplie le partage de données. Mais c'est vrai qu'il y a ceux qui comprennent et ceux qui ne comprennent pas. Parmi ceux qui ne comprennent pas, on sent le côté compétition. Je pense que c'est parce qu'ils n'ont pas forcément de position fixe. En tout cas, ça joue. Surtout quand tu commences à avoir 45 ans. Ça n'est pas facile.

[>R1]: En même temps, il y a tout un aspect positif qu'ils ne voient pas : cet aspect collaboration, qui donne davantage de visibilité.

[>R2]: Il y a aussi la mentalité des gens. Ça ne vaut pas seulement pour le partage des données, mais aussi pour le partage d'idées. Par exemple, moi on me dit « j'aimerais que tu travailles avec nous sur cet article », je suis capable de dire : « Okay, pas de souci. On n'a pas du tout la même façon de penser sur ce sujet-là, mais ça va être intéressant. Ça va être un challenge, mais ça va être intéressant ». Moi je suis très ouvert : on va construire, on va échanger. Et au pire, si ça ne marche pas, ça ne marche pas. Mais d'autres personnes m'ont

dit : « ah non, moi je pense qu'on ne va pas être d'accord sur ce truc-là, donc je préfère ne pas participer ». Et tu lui dis : « Mais non, justement, ça va être intéressant. Si je publie cet article-là, c'est sûrement toi qui va être reviewer. Donc autant que tu participes tout de suite et qu'on construise un article super bien construit ». Du coup, quand j'ai dit ça à la personne, elle m'a dit « oui, d'accord, je comprends » et elle a accepté. Mais, à mon avis, c'est une façon de penser.

9 - 47:08 > 50:06 Éditeurs scientifiques

[>R1]: A votre avis, pourquoi les éditeurs demandent d'associer le fichier de données à la publication, quand il s'agit d'humains, et pas forcément quand il s'agit d'animaux ?

[>R2]: A cause du côté appliqué. Chez les humains il y a quand même une approche très appliquée derrière – ce qui n'est pas forcément le cas pour les primates, où la recherche est relativement fondamentale. Mais je pense que c'est quelque chose qui va se généraliser, pour qu'on puisse répliquer les études. Pour moi, c'est déplacer le problème. Qu'on me donne un fichier qui puisse être réutilisé. Si dans ton étude tu me dis que tu as fait une corrélation entre ces deux colonnes et que tu obtiens R^2 de 96%, moi, si je reprends ton fichier, je fais la corrélation entre les deux colonnes, j'obtiens R^2 de 96%, c'est okay. Ça ne veut pas dire que ce fichier n'est pas totalement faux non plus. Peut-être qu'il a été modifié de manière à obtenir ce qu'on voulait. Et là, personne n'ira nous demander les données brutes. Tu peux les balancer, mais personne ne comprendra, parce que c'est un énorme fichier texte. Donc, pour moi, c'est mettre une sorte de filtre à la fraude scientifique. Ça limite un peu le problème, mais ça ne le limite pas totalement. Je pense qu'il y a aussi ce problème d'inattention. Parfois tu publies un truc, et puis tu te rends compte qu'il y avait une faute dans ce fichier-là. Donc tu corriges. Voilà, je pense que c'est pour répliquer les études et pour limiter les fraudes. Parce que la course à la publication est tellement intense qu'elle peut mener à ce genre de soucis. Mais personne n'est là pour vérifier ensuite. Il m'est arrivé qu'un reviewer me dise : « Tiens, j'ai repris les données que tu avais mises dans ton fichier, mais je ne retrouve pas le même pourcentage. Comment ça se fait ? ». Je lui avais répondu : « ah ben non, c'est parce qu'il faut prendre ce truc-là à la place et ensuite on retrouve le même résultat ». Normalement c'est leur boulot. Les reviewers sont censés vérifier rapidement – certains le font.

[>R1]: Ça leur fait du travail en plus ?

[>R2]: Oui.

10 - 50:06 > 52:21 Parcours professionnel

[>R1]: J'ai une toute dernière question. J'essaie de savoir si la formation universitaire a une influence sur la manière dont les chercheurs mènent leurs recherches et gèrent les données. Dans quel établissement avez-vous fait votre formation universitaire (master, thèse…) ?

[>R2]: J'ai fait mon master et ma thèse à Strasbourg. J'ai eu deux approches différentes, parce que j'ai fait ma thèse en cotutelle avec l'Université Libre de Bruxelles, avec quelqu'un qui faisait de la modélisation. Avec cette personne, j'ai appris à modéliser, j'ai été dans des réseaux et des colloques, où il y avait des informaticiens. C'est grâce à ça que j'ai cette approche de partage de données, que beaucoup de mes collègues n'ont pas forcément. Grâce à cette double formation. Du côté informatique, c'est le partage à fond.

[>R1]: A votre avis, pour quelle raison y a-t-il cette culture du partage en informatique ?

[>R2]: Pourquoi pas un autre domaine ? Je n'en sais rien. Ce sont des mentalités/des cultures différentes, je dirais. C'est intra culture. C'est-à-dire c'est une culture qui s'est développée dans les écoles d'informatique, qui a évolué et qu'on acquiert en étant dans cette formation. On le voit bien aussi à notre façon de travailler avec un doctorant. Quand le doctorant fait une thèse avec tel encadrant, il va intégrer la culture du labo et de l'encadrant dans sa façon de travailler. Et s'il va voir d'autres personnes, il va acquérir différentes cultures. C'est pour ça que j'adore que les thèses que j'encadre soient en cotutelle – avec au moins deux encadrants, pour que le doctorant puisse avoir différents sons de cloches et apprenne beaucoup plus qu'en étant avec un seul encadrant.

Entretien avec le chercheur 18 (neurosciences)

1 - 00:00 > 15:17 Objectifs scientifiques du projet

[>R1]: Est-ce que, rapidement et simplement, vous pourriez me résumer les objectifs scientifiques du projet et comment sont répartis les rôles entre les différents partenaires ?

[>R2]: Je vais essayer de contextualiser l'origine du projet. Le projet porte sur deux noyaux du thalamus. C'est une structure cérébrale dite « encéphalique », dont on a pensé pendant très longtemps qu'elle n'avait, grosso modo, qu'un rôle de relais/de transfert d'information, éventuellement de filtrage d'information. En gros, il y a de l'information qui descend du cortex et qui part à la périphérie, ou il y a de l'information qui vient de la périphérie et qui arrive au cortex. Par ses différents noyaux, le thalamus serait là pour filtrer cette information, éventuellement la réorganiser un peu et la passer plus loin. Or, on s'est rendu compte que, dans ce thalamus qui comporte une soixantaine de noyaux, certains noyaux avaient des fonctions qui dépassaient très largement le cadre de relais et qui avaient un retentissement sur le plan cognitif (« cognition » : les fonctions qui nous permettent d'avoir une connaissance du monde dans lequel nous fonctionnons, dont en particulier les fonctions mnésiques, i.e. tout ce qui touche à la mémoire). En 2009, nous avons publié un premier travail sur des noyaux qu'on appelle les noyaux intralaminaires. On a montré que, lorsque ceux-ci sont détruits, les modèles sur lesquels nous travaillons (le rat en particulier) présentent des capacités d'apprentissage tout à fait normales, dans le contexte d'une tâche où l'animal doit apprendre la localisation d'une plateforme qui lui sert de refuge. Nous avons montré que, lorsqu'on teste leur rétention de l'information cinq jours après cet apprentissage, les rats présentent des performances parfaitement normales. Mais lorsqu'on teste leur rétention vingt-cinq jours après, ces animaux ont complètement oublié la localisation de la plateforme. De ces noyaux intralaminaires, nous sommes passés à d'autres noyaux, dont la connectique nous paraissait encore plus en résonance avec ces fonctions de maintien à long terme des souvenirs. Nous avons lésé ces noyaux et nous avons constaté que, tout comme c'était le cas après une lésion des noyaux intralaminaires, les animaux apprenaient normalement, les animaux retenaient l'information pendant quelques jours et, après quelques jours, l'information avait disparu. Du coup, on s'est dit : « ce serait intéressant de comprendre ce qui se passe ». Pourquoi on

apprend correctement après une lésion de ce type-là ? Pourquoi on arrive à maintenir cette information absolument intacte pendant quelques jours seulement ? Et surtout, pourquoi, au bout de quelques jours, cette information disparaît ? Je me suis donc approché de trois autres équipes françaises : une à Bordeaux, qui manipulait des outils de transmission virale soit de traçage, soit de lésion, en relation avec les fonctions d'apprentissage et de mémorisation ; une à Marseille, qui depuis très longtemps travaille à l'enregistrement de l'activité de cellules nerveuses dans une structure qui a un rôle absolument capital dans nos fonctions mnésiques, qu'on appelle l'hippocampe, et qui s'intéresse à une catégorie très singulière de cellules, qu'on appelle des cellules de lieu (ce sont des cellules qui interviennent dans le codage de l'information spatiale) ; une équipe toulousaine, qui est connue pour ses travaux sur la neurogénèse. Il y a des données dans la littérature qui montrent que des neurones nouvellement fabriqués par notre cerveau pourraient jouer un rôle dans la longévité de la trace mnésique. Donc l'idée c'était de voir si, lorsqu'on fait ce type de lésion, l'animal intègre l'information spatiale de la même façon que le ferait un animal intact. D'où l'intérêt pour le fonctionnement des cellules de lieu, dans le contexte d'une navigation spatiale : est-ce que, quand je me déplace dans un environnement après ce type de lésion, j'intègre l'information de la même façon que le fait un animal dont le cerveau serait intact ? La raison pour laquelle nous nous sommes approchés de l'équipe toulousaine, c'est précisément cette potentielle implication des neurones nouvellement fabriqués dans le maintien d'une trace mnésique. Est-ce qu'après ce type de lésion on affecterait les régulations qui assurent la neurogénèse et, par là même, on impacterait sur cette neurogénèse – voie par laquelle on pourrait imaginer que les animaux apprennent, retiennent l'information pendant quelques jours, mais quand il s'agit pour eux de pérenniser cette information pour très longtemps, il faut qu'ils puissent mobiliser des neurones nouvellement générés et, si ce type de lésion a altéré la neurogénèse, cette mobilisation pourrait ne plus être possible ? Ici, à Strasbourg, nous avons déployé deux types de démarche. Un premier type de démarche, où nous nous sommes intéressés à la manière dont ces lésions pourraient éventuellement affecter des régulations épigénétiques. En particulier, lorsque nous apprenons et pour que nous puissions maintenir dans notre mémoire les informations que nous y avons intégrées, notre cerveau procède à la réorganisation d'un certain nombre de circuits dans des réseaux neuronaux, qui sont supposés stocker cette information. Ce stockage ne peut pas se faire s'il n'y a pas au minimum une réorganisation

fonctionnelle dans ces réseaux. Mais il se trouve qu'il y a, en plus, une réorganisation structurelle. Il y a des connexions qui disparaissent et il y a des connexions qui apparaissent dans ces réseaux. L'apparition de ces connexions, la réorganisation de ces réseaux, les modifications fonctionnelles au niveau de synapses/connexions existantes, nécessitent des mécanismes qui vont apporter par exemple un surcroît de certaines protéines permettant à cette information de se pérenniser. Or ce surcroît de protéines répond à une sollicitation particulière, qui est celle à laquelle est confronté l'animal lorsqu'on le soumet à une tâche d'apprentissage ou de rappel. Donc, pour que ces protéines apparaissent en nombre plus important, il faut qu'elles fassent l'objet d'une synthèse. Et pour faire l'objet d'une synthèse, il faut que certains gènes présents dans la chromatine soient transcrits puis traduits. Il y a des régulations qui vont faire en sorte que cette chromatine, à l'endroit où se trouvent ces gènes, soit décompactée, ce qui va permettre la lecture du gène, la synthèse d'un ARN messager. Cet ARN messager pourra alors faire l'objet d'une traduction au niveau des ribosomes par ce processus de traduction – j'entends la fabrication du surcroît de protéines dont on a besoin. Donc ce qui nous a intéressés, c'est la manière dont cette lésion allait affecter des régulations épigénétiques. Je ne vais pas entrer plus dans le détail. Je pense que c'est peut-être déjà suffisamment compliqué.

[>R1]: Oui, c'est suffisant. L'épigénétique, donc, c'est ce phénomène d'expression d'un gène ?

[>R2]: C'est tout ce qui se passe sur les gènes – tout ce qui se passe en termes de régulation et d'expression des gènes. Ça n'est pas forcément qu'une régulation permissive ; ça peut être une régulation répressive. Il peut y avoir des mécanismes épigénétiques qui vont empêcher que certains gènes soient exprimés. Ou des régulations épigénétiques qui vont favoriser l'expression de certains gènes. Ça c'est le premier aspect dans lequel on s'est engouffrés ici, à Strasbourg, en nous focalisant sur l'hippocampe – structure absolument essentielle à la mise en mémoire de l'information – et en nous focalisant aussi sur le cortex préfrontal – structure absolument essentielle à la pérennité de nos souvenirs. L'idée c'était de savoir quels types de régulation pouvaient être affectés dans l'hippocampe après ce type de lésion et quels types de régulation pouvaient être affectés dans le cortex préfrontal. Et comment, éventuellement, une altération de ces régulations pouvait être rattachée au fait que le souvenir soit intégré dans la mémoire mais n'y reste pas longtemps. Deuxième aspect : cette fois-ci on n'est plus dans le

domaine de la biologie moléculaire (l'épigénétique nécessite la mise en œuvre d'outils dans le domaine de la biologie moléculaire) mais on est dans le domaine de l'imagerie cérébrale. Qu'est-ce qu'on va essayer de visualiser ? On va visualiser de toutes petites structures qui se trouvent sur les dendrites, qu'on appelle épines dendritiques et qui sont les endroits où d'autres neurones viennent entrer en contact avec un neurone cible, auquel ils vont pouvoir transmettre de l'information. Donc, si j'ai une augmentation ou une diminution du nombre d'épines dendritiques à la suite de ce type de lésion, comparativement à des animaux témoins, ça veut dire que la lésion va affecter la capacité des circuits, dans l'hippocampe comme dans le cortex préfrontal, à se réorganiser pour permettre une pérennisation du souvenir.

[>R1]: En fait, c'est comme si le cerveau s'adaptait ? Il multiplie le nombre d'épines dendritiques ?

[>R2]: C'est tout à fait ça. Ce qui se passe après un apprentissage (toujours pour essayer de rester au niveau le plus simple de ce que je vous explique), c'est que l'hippocampe va être sollicité par la situation dans laquelle l'animal apprend – par exemple la localisation d'une plateforme. Cette sollicitation va déclencher un certain nombre de processus de réorganisation des circuits hippocampiques. Cette réorganisation va s'appuyer, entre autres, sur la genèse de nouvelles synapses. La genèse de nouvelles synapses deviendra détectable par une approche quantitative, où nous allons comptabiliser le nombre d'épines dendritiques sur des fragments standardisés de dendrites. S'il y a de nouveaux contacts, il y aura plus d'épines dendritiques ; s'il y a moins de contacts, il y aura moins d'épines dendritiques. Et nous allons avoir exactement la même approche au niveau du cortex préfrontal. Sachant que, dans le cortex préfrontal, la réorganisation des circuits neuronaux, qui vont servir de support au souvenir, est plus tardive. Elle interviendra dix à quinze jours plus tard. Du fait de ces réorganisations séquentielles, on va donc avoir un souvenir qui dépend initialement de l'hippocampe et qui, après un certain temps, tout ou partie, ne dépendra plus de l'hippocampe mais du cortex préfrontal, où il aura été reconstruit. Cette reconstruction dans le cortex préfrontal va, elle aussi, s'accompagner d'une augmentation du nombre d'épines dendritiques, qui traduit, en fait, une réorganisation des circuits : formation de nouvelles synapses ; probablement aussi disparition d'autres synapses ; mais dans tous les cas quelque chose d'extrêmement dynamique. Ce que nous cherchons à savoir c'est si, à la suite d'une lésion de ces fameux noyaux thalamiques (dont je ne vous ai peut-être pas encore donné le nom, mais qui

s'appellent « reuniens » et « rhomboïdes »), on a une augmentation moins importante du nombre d'épines dendritiques dans l'hippocampe et à quel moment, et si on a une augmentation moins importante du nombre d'épines dendritiques dans le cortex préfrontal et à quel moment. Voilà, j'espère qu'avec tout ça vous voyez un peu le cadre.

[>R1]: Oui, c'est déjà plus clair.

2 - 15:17 > 16:58 Ressources humaines

[>R1]: Combien de personnes travaillent sur le projet, ici, dans ce laboratoire ?

[>R2]: Sur ce projet, on est deux chercheurs à travailler. Trois, pardon. Plus deux doctorantes. Trois doctorants. Pardon, excusez-moi, je vais reprendre. Nous sommes quatre chercheurs statutaires à travailler sur ce projet, à quoi il faut rajouter trois doctorants et une technicienne.

[>R1]: Le sujet de thèse de chaque doctorant porte sur une partie du projet ?

[>R2]: Absolument. Il y a une doctorante qui travaille sur les régulations épigénétiques, par l'intermédiaire d'une approche en biologie moléculaire. Une autre doctorante travaille sur l'imagerie cérébrale et compte les épines dendritiques – enfin, elle ne fait pas que compter, elle prépare aussi le tissu, elle fait les tests comportementaux… elle fait tout cela. Et il y a un doctorant qui analyse l'implication de ces noyaux, par une approche utilisant un autre test de mémoire à long terme que celui dont je vous ai parlé tout à l'heure. Ce test consiste pour le rat à apprendre à localiser une plateforme, qui lui sert de refuge dans une piscine. Le doctorant fait aussi des enregistrements physiologiques.

3 - 16:58 > 23:27 Mode d'acquisition des données

[>R1]: Quelles expérimentations faites-vous exactement ? Comment ça se traduit en termes d'instruments, de résultats...?

[>R2]: Je vais d'abord parler des points communs. Le point commun aux trois approches est l'évaluation comportementale.

[>R1]: Quand vous dîtes « les trois approches », ce sont les approches des trois doctorants ?

[>R2]: Oui. Il est clair que chaque fois que nous faisons une manipulation sur le cerveau des animaux, nous devons nous assurer que les impacts attendus sur le comportement sont bel et bien vérifiés. A savoir qu'effectivement ces animaux apprennent, retiennent quelques jours puis oublient. Ça c'est le point commun entre les trois approches. Ensuite, il y a un certain nombre de variantes dans ces approches comportementales. Par exemple, lorsqu'on s'intéresse aux régulations épigénétiques, on ne va pas forcément aller au bout de l'apprentissage, on ne va pas aller au bout du délai long avant de tester le rappel. On va, par exemple, regarder ce qu'il se passe pendant que l'animal est en train d'intégrer de l'information, sur le plan des régulations qui ont lieu dans l'hippocampe ou dans le cortex préfrontal. Dans le cas de l'imagerie, il est clair qu'on ne va pas forcément tester tous les animaux, parce que ce qui nous intéresse notamment, c'est l'effet basal d'une lésion. A savoir : « est-ce que la lésion va impacter le nombre d'épines dendritiques, sans que l'animal n'ait à faire quoi que ce soit dans quelque tâche que ce soit ? » versus « qu'est-ce qui se passe quant aux conséquences de cette lésion, lorsque l'animal doit apprendre et retenir pour peu de temps ou pour un temps beaucoup plus long ? ». La dernière approche c'est une approche où on va descendre des électrodes dans certaines structures connectées à ces fameux noyaux thalamiques et où on va enregistrer les conséquences d'une lésion ou d'une inactivation sur l'activité des cellules nerveuses dans l'hippocampe ou dans le cortex préfrontal.

[>R1]: Comment mesurez-vous le nombre d'épines dendritiques ?

[>R2]: Pour compter le nombre d'épines dendritiques, on met en œuvre une technique qui consiste à colorer un certain nombre de neurones. Cette technique est très ancienne, elle date de la fin du 19ème siècle. Alors, elle a été remise au goût du jour : on n'a plus autant de manipulations à faire qu'à l'époque. A l'époque, de la mise à mort de l'animal jusqu'à l'observation des cellules, il s'écoulait entre quatre et six semaines ; aujourd'hui ça se réduit à quelques jours.

[>R1]: Il s'agit d'une injection colorée ?

[>R2]: C'est un traitement du tissu cérébral. On va récupérer le cerveau et on va le soumettre à une imprégnation argentique, qui va ensuite faire l'objet d'une révélation en différentes étapes. Par cette technique, qui a été inventée par Camillo Golgi à la fin du 19ème siècle et qu'on appelle « la réaction noire », on arrivera à colorer à peu près 2% de tous les neurones du

cerveau du rat. Vous me direz que ça n’'est pas beaucoup. Et pourquoi seulement 2% ? Personne ne le sait encore aujourd'hui. Mais les neurones sont colorés dans leur ensemble. On voit les dendrites, on voit l'axone, on voit le corps cellulaire et surtout, sur les dendrites, on voit parfaitement bien les différents types d'épines dendritiques. Donc, une fois qu'on a ce matériel, on va tout simplement le placer sous un microscope, avec un grossissement de l'ordre de cent fois – parce que les épines dendritiques sont absolument minuscules. Elles mesurent aux alentours d'un micromètre. Pour pouvoir compter ces épines, on va donc devoir faire un grossissement extrêmement important sous un microscope. Le comptage se fait ensuite selon des techniques dites « stéréologiques » non biaisées (je ne vais pas rentrer dans les détails, ça n'a pas beaucoup d'intérêt). C'est un travail qui est très long.

[>R1]: Et pour les régulations épigénétiques, comment vous y prenez-vous ?

[>R2]: Là, on ne va pas fixer le cerveau. On va récupérer le cerveau frais et on va procéder à une série de dissections et de sous-dissections des différentes structures. Là non plus je ne vais pas rentrer trop dans le détail. La dissection est assez... Disons qu’elle relève relativement de la contorsion. Mais, enfin, on y arrive. On va ensuite traiter les tissus collectés au cours de la dissection, pour pouvoir mesurer la présence de certaines protéines qui interviennent dans des régulations permettant la réorganisation des circuits neuronaux. On va aussi s'intéresser à la manière dont ces régulations épigénétiques s'opèrent, en regardant quel est l'état des sites intervenant dans ces régulations épigénétiques : est-ce qu'ils ont fait l'objet d'une réaction chimique, qui va traduire le décompactage de la chromatine et la possibilité de transcrire puis de traduire le gène ? Tout cela grâce à des outils qui relèvent du registre de la biologie moléculaire.

4 - 23:27 > 29:26 Types de données de recherche

[>R1]: Je vais maintenant passer à la partie « données de recherche ». Est-ce que vous utilisez dans vos recherches le terme de « données » ? Est-ce que c'est un terme que vous utilisez couramment ?

[>R2]: Absolument, oui.

[>R1]: Qu'est-ce que vous désignez sous ce terme, si on prend l'exemple du projet T. ?

[>R2]: Tout ce qu'on a généré, à partir de nos différentes approches méthodologiques.

[>R1]: Donc le produit de vos expérimentations ?

[>R2]: Le produit de l'expérimentation, qu'il soit qualitatif ou quantitatif, est considéré comme donnée.

[>R1]: A quoi ressemblent les résultats qualitatifs ?

[>R2]: Alors, on ne fait pas simplement qu'observer « l'animal se souvient / ne se souvient pas ». Il y a une véritable objectivation du comportement du rat, en générant un certain nombre de variables : combien de temps il met avant d'accéder à la plateforme ; quelle distance il nage avant d'accéder à la plateforme ; combien de temps il passe à se coller contre la paroi de la piscine à différents moments de son entraînement ; combien de temps il passe dans la zone où se trouvait la plateforme au cours d'un essai-test ; à quelle distance moyenne de la plateforme il effectue sa recherche ; etc. Donc on a toute une série de variables qu'on va générer et qui vont nous permettre de rendre compte de manière quantitative du comportement de l'animal. Nous pourrons ensuite comparer, sur la base de ces données quantitatives, un groupe de rats à un autre groupe de rats chez qui, par exemple, on aura effectué les lésions dont je vous ai parlé tout à l'heure. Afin de voir si, dans leur comportement, quelque chose a été modifié – soit pendant l'acquisition de la tâche (ce que j'ai appelé tout à l'heure « l'entraînement »), soit au premier test de rappel que nous allons faire quelques jours après la fin de l'acquisition de la tâche, soit au deuxième test de rappel beaucoup plus tardif, que nous ferons 25-30 jours après la fin de l'acquisition de la tâche.

[>R1]: Toutes ces données sont rentrées dans un tableur ? Sous quelle forme numérique elles se présentent ?

[>R2]: Les doctorants ou nous-mêmes (quand nous avons l'occasion de le faire) récupérons les données sur un logiciel de vidéo tracking – enfin, sur le disque dur nourri à partir d'un logiciel de vidéo tracking. Puis nous les transférons dans un tableur Excel, où nous pouvons faire nos calculs.

[>R1]: Et les autres résultats des expérimentations, sous quel type de fichier sont-ils sauvegardés ?

[>R2]: C'est quasiment toujours ce type de fichier, quand il s'agit de données quantitatives. Sous la responsabilité des étudiants. Ou alors, ce sont des données plus qualitatives. Par exemple, des photos de coupe histologique. Dans chaque groupe et pour chaque animal, on peut prendre des photographies représentatives de ce qui se passe dans ce groupe, après avoir traité le tissu selon différentes méthodes histologiques. Certaines sont des méthodes de base, qui vont simplement nous permettre de vérifier, par exemple, qu'il n'y a plus de neurones dans la zone de lésion ou qu'il y a un nombre résiduel de neurones extrêmement bas. D'autres photographies vont nous permettre de visualiser le fait que, pendant que l'animal s'est rappelé l'information, certains neurones ont été activés.

[>R1]: Comment parvenez-vous à voir l'activation de certains neurones sur un cerveau disséqué ?

[>R2]: Alors, les techniques, dont je vous ai parlé et qu'on utilise dans le cadre de ce projet, ne permettent pas de voir des neurones s'activer, pendant que l'animal est en train de faire quelque chose. Mais ce type de techniques existe. C'est par exemple l'imagerie calcique, qu'on utilise pour d'autres projets dans le laboratoire. On met à demeure, sur la tête du rat, un mini microscope fluorescent (enfin, sensible à une certaine longueur d'onde) et on va suivre l'activation des neurones en les traitant d'une certaine manière, qui va permettre de rendre le calcium présent dans ces neurones fluorescent. Quand un neurone est actif, sa concentration intracellulaire de calcium augmente. Quand il cesse d'être actif, elle diminue. Si vous rendez ce calcium fluorescent et que vous équipez l'animal d'un système de détection de fluorescence, vous pouvez suivre en temps réel non seulement l'activation d'un seul neurone mais l'activation d'une multitude de neurones, et même le pattern/le profil général d'activation de cette multiplicité de neurones. Mais ça, on ne l'utilise pas sur ce projet.

[>R1]: Ni les autres partenaires du projet ?

[>R2]: Ni les autres partenaires du projet.

5 - 29:26 > 31:04 Répartition des rôles

[>R1]: Vous disiez que les données étaient sous la responsabilité des doctorants. Qu'entendez-vous par là ? Voulez-vous dire que les doctorants conservent les données acquises sur leurs propres supports de sauvegarde ?

[>R2]: Les données acquises sont acquises sur un ordinateur qui appartient au laboratoire. Donc les données brutes sont sur l'ordinateur du laboratoire. Les étudiants récupèrent ces données. Ils peuvent évidemment les laisser sur l'ordinateur mais pas « ad vitam aeternam », parce que, sans quoi, on arrive très rapidement à l'encombrement du disque dur. Ce sont quand même des fichiers assez volumineux. Donc, les doctorants récupèrent ces données sur leur ordinateur, les traitent et les analysent statistiquement – évidemment on en discute entre personnes responsables de l'encadrement de ces étudiants. Il est clair qu'on en discute. L'étudiant n'est pas livré à lui-même pendant trois ans, à déambuler dans son sujet de thèse, dont il ne ferait peut-être rien, si tel était le cas. Non, on en discute. On regarde ensemble les résultats des expériences, on discute ensemble de ce qu'il faut faire après. On discute aussi ensemble des variables qu'il convient de retenir, de celles qu'il conviendrait peut-être de générer, en plus des variables habituelles qu'on utilise, selon ce que font les animaux dans les situations de test. Donc il y a un échange évidemment permanent entre le ou la doctorant(e) et la personne qui assure la responsabilité de cette thèse.

6 - 31:04 > 31:57 Sauvegarde des données

[>R1]: Est-ce que vous avez un serveur commun qui centralise toutes les données, aussi bien brutes qu'analysées ?

[>R2]: Il y a un serveur au sein du laboratoire, qui permet non pas seulement de rassembler toutes ces données, mais qui permet aussi de sauvegarder quotidiennement l'intégralité du travail généré à partir des différents postes informatiques du laboratoire. Tout en gardant une confidentialité, puisque, moi par exemple, je n'ai pas accès à l'espace de ma voisine de bureau sur ce serveur central.

7 - 32:00 > 33:49 Documentation des données

[>R1]: Est-ce qu'il y a une documentation des procédures ? Par exemple, est-ce que des métadonnées sont associées aux fichiers de données ? Je pense notamment aux photographies : est-ce que les fichiers sont paramétrés pour enregistrer un certain nombre d'informations de contexte – comme, par exemple, la date à laquelle la photographie a été prise, etc. ? Ou bien est-ce que c'est quelque chose qui n'est pas forcément important ?

[>R2]: Ça n'est pas forcément important. Mais je vous donne cette réponse, en précisant quand même que ça n'est pas dans l'absolu que je vous la donne. Ça peut être extrêmement important. Je sais, par exemple, que certaines personnes dans le laboratoire, font des enregistrements électro-physiologiques, en manipulant fonctionnellement les cellules de certaines structures cérébrales, et stockent ces enregistrements en les rapportant évidemment à l'animal, à la date, à l'historique de l'animal, etc. Donc, là, toutes les conditions sont précisées et, de surcroît, les enregistrements sont stockés parallèlement à une vidéo de ce qu'a fait l'animal pendant la situation de test. Ce qui fait que, à chaque moment du test qu'effectue l'animal, on arrive à la milliseconde près à lui faire correspondre une activité cellulaire ou une activité de champ, enregistrée au niveau de l'une ou l'autre des structures du cerveau qui nous intéresse.

8 - 33:49 > 36:23 Conservation des données

[>R1]: Les données seront-elles conservées au terme du projet ? Ou bien quelles données seront conservées, si toutes ne le sont pas ?

[>R2]: En général, on a une règle dans le laboratoire (mais elle est peut-être un tout petit peu « has been », je n'en sais rien, je ne me suis jamais posé la question), qui est que, tant que ça n'est pas publié, on conserve à la fois les données et le matériel brut. Par « matériel brut », j'entends par exemple les préparations histologiques : quand on a fait nos colorations, les sections de tissu sont montées sur lame et couvre-lame et on va le stocker jusqu'à ce que le travail soit complètement valorisé. Après, on a le même problème que tout espace qui n'est pas plastique, à savoir qu'au bout d'un moment il faut qu'on gère de l'encombrement. Et la meilleure façon de gérer l'encombrement, c'est de jeter. Donc, tant que ça n'est pas publié ou tant que le matériel garde un intérêt, on conserve à la fois les données générées et le matériel

brut. Quand c'est publié, en règle générale, on s'en débarrasse au bout de quelques années. Il existe aujourd'hui des systèmes de stockage, notamment pour le matériel histologique, qui consistent à photographier en série et à stocker les photos – ce qui nous permet de nous débarrasser rapidement du matériel. Par exemple, quand on utilise des marqueurs fluorescents, ces marqueurs voient leur fluorescence s'atténuer au fil du temps et, en particulier, beaucoup plus vite encore lorsqu'on les place sous un microscope à fluorescence et qu'on les excite avec une certaine longueur d'onde, qui correspond à l'émission d'une fluorescence du marqueur qu'on utilise. Donc, plus on utilise ce matériel, moins il est utilisable. Ces fameux systèmes de prises de vue ont alors un intérêt majeur, parce que sur une photographie c'est figé/stabilisé.

9 - 36:23 > 37:43 Réutilisation des données

[>R1]: Est-ce qu'il vous arrive de réutiliser des données d'une précédente publication ?

[>R2]: D'une précédente publication, non. Mais de matériel, dont nous avons pensé, à un moment donné, que nous avions épuisé la totalité de ce qu'il était capable de nous dire et que nous aurions encore au laboratoire, c'est déjà arrivé, oui.

[>R1]: Avez-vous systématiquement recours à ce procédé de photographie du matériel histologique ?

[>R2]: Non, ça n'est pas systématique. Je parlais d'un autre laboratoire. C'est un dispositif qui vaut 200 000€, donc il est clair que nous n'aurions pas les moyens d'acheter un système comme ça, pour couvrir les quelques besoins que nous aurions à l'échelle de ce laboratoire. Le système a été installé sur une plateforme et il est accessible à toute personne qui pourrait en avoir besoin.

[>R1]: Moyennant paiement ?

[>R2]: Moyennant paiement. Ce qui va de soi.

10 - 37:43 > 43:53 Échanges de données entre les partenaires du projet

[>R1]: Y a-t-il échange de données entre les différents partenaires, que ce soit pendant ou après le projet ?

[>R2]: Il n'y a pas d'échanges de données brutes dans le cadre de ce projet, dans la mesure où chacun avait sa propre ligne de travail et où ces lignes de travail étaient indépendantes pour la presque totalité de ce que nous faisions. Néanmoins, pour essayer de minimiser le nombre de rats utilisés sur l'ensemble du projet, plutôt que de faire deux fois la même chose dans deux villes différentes, on se débrouillait pour qu'une même manipulation soit faite dans une seule ville, puis que du matériel biologique soit envoyé à l'autre partenaire. C'est ce que nous avons fait avec l'équipe toulousaine. Et nous allons recevoir (probablement au cours du printemps 2018) des rats qui auront fait l'objet d'une intervention chirurgicale à Bordeaux.

[>R1]: Finalement, on peut quand même parler d'échange de données brutes ?

[>R2]: Alors, faites bien la distinction entre le matériel biologique préparé à partir de l'animal et ce que l'on va chercher sur le matériel biologique. Le matériel biologique n'est pas une donnée brute. La donné brute, c'est ce qu'on va aller chercher, ce qu'on va générer à partir de ce qu'on cherche dans le matériel biologique.

[>R1]: Donc, la donnée brute c'est déjà le fruit d'une interprétation ?

[>R2]: D'une visualisation, d'une observation. Par exemple, on aurait pu imaginer que notre partenaire toulousaine nous envoie le matériel qu'elle a généré pour visualiser les cellules nouvellement formées dans le système nerveux, et en particulier dans l'hippocampe ; mais on ne va pas avoir besoin de le faire, puisqu'elle s'occupe aussi de la quantification. Si vous voulez, on n'envoie pas de données brutes, sauf sur demande. On échange du matériel biologique. On échange de l'information. Par exemple, en juin l'année dernière, on a organisé une réunion de travail à Marseille, où les quatre partenaires se sont retrouvés et où nous avons passé une journée à exposer ce que nous avions collecté comme données. Mais pas comme données brutes ; comme données moyennées, c'est-à-dire des données déjà passées par la moulinette des analyses statistiques descriptives et analytiques. Nous avons échangé : où en est chaque partenaire par rapport à sa partie du projet ; quels sont les résultats qui ont été dégagés ; il y a des publications qui ont déjà été faites. Par exemple, le partenaire marseillais a fini sa partie du projet, alors que le projet court encore jusque fin 2019. Une grosse partie de

ce projet marseillais a déjà été publiée (le papier est sorti il y a quelques semaines) ; l'autre partie reste à publier. Je vous donne un exemple d'échange. A Marseille, ils ont regardé les lésions qu'ils faisaient avec une technique d'IRM morphologique. C'est-à-dire qu'on prend l'animal (il est vivant, il est anesthésié), on le met dans une IRM, puis on va acquérir les données qui permettent de visualiser en 3 dimensions l'organisation de son cerveau et on va procéder à cette visualisation. Cette acquisition s'est faite selon des séquences particulières. Ici, à Strasbourg, nous avons eu besoin de nous assurer que les animaux que nous allions traités avec un produit étaient lésés correctement. Pourquoi ? Parce que : (1) nous ne disposons pas de ce produit en quantité illimitée ; (2) sa synthèse est formidablement chère. Donc, traiter des animaux qui n'auraient pas de bonnes lésions, ça voudrait dire gaspiller du produit. On a donc demandé à notre partenaire marseillais de nous communiquer les séquences, avec lesquelles il avait acquis les données en IRM pour visualiser les lésions de ces animaux. Et on a donné ces séquences à un collègue strasbourgeois, qui a pu acquérir l'information sur nos propres rats avec une IRM ici, à Strasbourg. Donc, voyez, il y a des échanges de bons procédés, des échanges de données analysées, des discussions aussi sur l'interprétation qu'il convient d'apporter à ce qu'on observe ; mais ce n'est pas dans notre habitude d'échanger des données brutes.

[>R1]: Parce que ça n'a pas d'intérêt ?

[>R2]: Oui, parce qu'il n'y a pas d'intérêt de le faire dans la structure précise du projet, dont nous sommes en train de parler.

[>R1]: Oui, chaque partenaire du projet a son but de recherche.

[>R2]: Absolument.

11 - 43:53 > 45:54 Valeur économique des données

[>R1]: Quelle valeur accordez-vous aux données ? Par exemple aux données de ce projet (je parle ici des données et pas du matériel biologique) ? Est-ce qu'elles ont une valeur économique ? Est-ce qu'il peut y avoir des dépôts de brevets ?

[>R2]: Je ne dirais pas que la valeur économique est immédiate. Je m'explique : je ne vais pas générer de l'argent avec ces données. Évidemment, je ne peux pas exclure que, dans le cadre

du travail que nous entreprenons au sein de ce projet, nous ne tombions pas sur quelque chose qui pourrait nous conduire à proposer un brevet. Ça, on ne peut pas l'exclure a priori. Mais, d'un autre côté, vous ne faites pas non plus une recherche toujours dans l'optique de structurer un brevet d'invention dans la foulée – notamment lorsque c'est du fondamental. Maintenant, la valeur indirecte, oui, il y en a une. Parce que si vous tombez sur un résultat majeur, vous allez chercher à le publier dans un journal de forte réputation et, par ce biais-là, vous allez crédibiliser de futures demandes de financement. Donc, valeur économique indirecte. Inévitablement.

12 - 45:54 > 53:17 Politiques des éditeurs en matière de données scientifiques

[>R1]: Est-ce que les éditeurs scientifiques de votre domaine vous demandent ou vous proposent de déposer des fichiers de données, à côté de la publication ?

[>R2]: Ça n’est pas dans les habitudes de la plupart des journaux. Je pense que des journaux comme Nature et Science demandent les données brutes ou, au moins, demandent aux auteurs de tenir à la disponibilité de leur lectorat les données brutes. La plupart des journaux, dans lesquels nous avons l'habitude de publier, ne le demandent pas. Mais je me pose très sérieusement la question depuis quelque temps de savoir si ça n'aurait pas un intérêt de disposer d'un support en libre accès, sur lequel nous puissions déposer à la fois nos données brutes et un texte qui consignerait l'interprétation que nous en faisons. Je vous donne un exemple. Ça fait deux ans que j'essaie de publier des données. Ce sont des données qui ne vont pas tout à fait dans le sens, dans lequel le prédirait la théorie en vigueur actuellement. Ce sont des données qui me conduisent donc à ramer en contresens et à secouer un peu le cocotier théorique, dans le contexte d'une vision que nous avons de l'interaction entre plusieurs systèmes de mémoire. Ça fait deux ans que j'essaie de publier ces données et je me heurte à une succession de refus – avec parfois des rapports qui suintent la mauvaise foi. C'est un travail de trois ans. C'est un doctorant qui l’a fait. C'est un boulot immense. Il est synthétisé en une trentaine de figures. Alors, je me dis que s'il y a deux ans, j'avais pu déposer ces données sur un site en « open access », avec un texte qui explique pourquoi on a fait la manip, comment on l’a faite, quels sont les résultats princeps et comment nous les

interprétons, je pourrais continuer encore 5 ou 6 ans à essayer de publier ce travail, dans tous les cas j'aurais fait date.

[>R1]: Il y a une plateforme nationale – HAL – qui permet de déposer en libre accès. Vous pouvez aussi le faire sur la plateforme de Strasbourg – UnivOAK.

[>R2]: Oui, je sais. Je ne suis pas familiarisé du tout avec ça. J'entends bien la discussion depuis quelques années à Strasbourg. J'entends bien aussi la discussion au niveau national et même international, par rapport à la main mise des grandes maisons d'édition sur les données scientifiques. Puisqu'à partir du moment où elles les publient, elles vous exproprient littéralement de votre copyright –en contrepartie de la publication (gratuite, la plupart du temps encore) de vos travaux. Je réfléchis très sérieusement à l'intérêt qu'aurait pour nous (et en particulier dans ce cas de figure) le fait de pouvoir déposer ça. Simplement, toutes les maisons d'édition ne sont pas en accord avec cela. Je sais bien qu'on y va. On y va doucement.

[>R1]: Effectivement, si vous déposez le « preprint » dans une archive ouverte, il est possible qu'un éditeur refuse ensuite de le publier.

[>R2]: Il y en a qui acceptent. Et je pense que ça va probablement faire tache d'huile, parce qu'on est quand même en train d'assister ces dix dernières années à une réforme en profondeur du système de l'édition scientifique. Il s'est passé quelque chose, notamment le succès absolument dingue et très rapide de Plos ONE. C'est un journal qui a été créé en 2006 (si mes souvenirs sont bons) et qui très rapidement s'est mis à publier annuellement plusieurs milliers d'articles. Le modèle économique est complètement différent de ce qu'on connaissait avant. Avant c'était : j'envoie un manuscrit ; s'il est accepté après évaluation par les pairs, il est publié, sous condition que je transfère mes droits d'auteur à l'éditeur. Aujourd'hui, dans le système « open access », je conserve mes droits sur le manuscrit, mais je paie la publication. [...] Dans le premier modèle économique, l'auteur du manuscrit transfère ses droits à l'éditeur et l'éditeur, en contrepartie de cela, peut demander à l'université de payer. Le comble c'est que, quand j'utilise une figure d'un de mes articles dans un travail de recension, je suis obligé, pour utiliser ma propre figure, de demander l'autorisation à l'éditeur du journal, dans lequel cette figure a été publiée. Cette figure, je l'ai générée (pas seulement en générant les données mais aussi en assurant l'iconographie) ; et elle ne m'appartient pas.

[>R1]: Oui, simplement parce que l'éditeur dispose des droits exclusifs de diffusion… C'est un peu aberrant.

[>R2]: C'est un peu limite, on va dire. Donc, en général, ce que je fais, c'est que je la redessine – différemment, bien sûr.

13 - 53:17 > 1:01:21 Ouverture des données

[>R1]: Qu'est-ce que vous pensez de ce que propose la Commission européenne, à savoir mettre en libre accès les données brutes ?

[>R2]: Je n'y vois aucun inconvénient. A partir du moment où j'accepte l'idée de publier des données moyennées/traitées. Finalement je pars de quoi ? Je pars du matériel brut et je le traite d'une certaine manière, qui m'est peut-être propre. Il y a peut-être d'autres manières de traiter ces données ; il y a peut-être d'autres façons d'analyser statistiquement ces données. Je ne vois pas pourquoi je ne pourrais pas les rendre librement accessibles, dans leur forme la plus brute. Alors, il faut savoir aussi ce qu'on entend par « forme la plus brute ». Est-ce que je vais mettre en ligne les vidéos de chaque essai de chacun de mes rats, enregistrées par le système de vidéo tracking, et démerdez-vous pour générer des variables comme la distance d'accès à la plateforme ou la latence d'accès à la plateforme ? Ou bien, est-ce que je mets en ligne ces variables quantitatives, générées par le logiciel à partir du système de vidéo tracking ? Si je fais le boulot de générer ces variables, je trouve ça un peu débile de demander à quelqu'un d'autre de faire le même boulot. Après, très rapidement, quand on commence à réfléchir à ce genre de questions, on ne peut pas faire autrement que d'évoquer aussi la fraude. Est-ce que j'ai cherché à frauder ? Est-ce que les variables quantitatives que je propose sont une transposition fidèle de ce que j'observe ? Mais là, vous avez bien conscience d'une chose : c'est qu'on finit par rentrer dans un cercle infernal, qui va finir par générer beaucoup de folie. Moi, j'ai plutôt tendance à faire confiance. Maintenant, on voit bien aujourd'hui (je vous l'ai dit indirectement tout à l'heure, quand j'ai parlé d'intérêt économique indirect) que la qualité du support, dans lequel je publie mes résultats, a un impact inévitable sur la crédibilité avec laquelle je vais faire une demande de financement, et donc sur la probabilité d'obtenir ce financement. Ça veut dire quoi ? Ça veut dire que (et les multiples cas de fraude qu'on dénonce ces derniers temps, ici ou là à travers le monde, mais aussi en France et à Strasbourg,

le montrent), dans le milieu scientifique, si je ne suis pas quelqu'un de rigoureux, si je ne suis pas quelqu'un de totalement honnête, si je finis par craquer, parce que je n'arrive pas à financer mes recherches, je vais peut-être être tenté de donner un petit coup de main à mes résultats, de manière à les faire passer d'une publication dans un journal lambda à une publication dans un journal dont le retentissement sera plus fort – ce qui, inévitablement, contribuera à augmenter ma probabilité d'obtenir un financement pour des projets futurs. On est entré dans un système très vicieux aujourd'hui. Le fait de pouvoir déposer des données dans une archive ouverte contribuera probablement à réduire cela. Mais je n'en suis même pas totalement convaincu, parce que, tant qu'on ne me demandera pas de déposer le matériel brut dans sa forme la plus brute possible, je pourrai bricoler ce que je veux.

[>R1]: Donc, demander la publication des données brutes permettrait de réduire la fraude ?

[>R2]: Je ne suis pas sûr que ça permette de la réduire. Ce qui permettrait très probablement de réduire la fraude, c'est la mise en archive ouverte non pas des articles, non pas des données dérivées, mais du matériel brut lui-même.

[>R1]: Pourquoi en archive ouverte et pas chez un éditeur ?

[>R2]: Parce que ça n’est pas gérable, je pense.

[>R1]: D'un point de vue économique ?

[>R2]: Du point de vue des espaces de stockage déjà : l'éditeur n'a pas pour vocation de stocker l'information brute ; l'éditeur a pour vocation de diffuser de l'information. Ça n'est pas la même chose.

[>R1]: Mais un éditeur dispose de moyens financiers bien plus importants qu'une archive ouverte.

[>R2]: Oui et non. Dans l'absolu, vous avez raison. Après, il faut se poser la question de savoir quel est l'objectif de l'éditeur : faire du fric – on est d'accord. On peut imaginer qu'il le fasse en proposant des services d'archive ouverte contre paiement. C'est à ça que vous pensiez ?

[>R1]: Oui. Je pense que si les éditeurs décident de proposer des plateformes pour déposer les données brutes, ils mettront en place un modèle économique.

[>R2]: Ce sera forcément un modèle économique. Autrement le système n'est pas vivable. Viable, pardon. Il est vivable mais il n'est pas forcément viable économiquement.

[>R1]: Je pense que les plateformes d'archives ouvertes seront, elles aussi, obligées d'instaurer un modèle économique, si elles veulent disposer d'espaces de stockage suffisants.

[>R2]: Il faut qu'elles s'appuient sur un système de financement, c'est inévitable. A moins que le financement soit public. Mais je n'y crois pas.

14 - 1:01:21 > 1:07:04 Parcours professionnel

[>R1]: J'aurais une dernière question, un peu plus personnelle. Dans quel établissement avez-vous effectué votre formation universitaire ?

[>R2]: L'établissement s'appelait Université Louis Pasteur.

[>R1]: Ici, donc.

[>R2]: A Strasbourg, oui.

[>R1]: Avez-vous remarqué une évolution dans la gestion des données ? Par rapport à leur sauvegarde, par exemple ?

[>R2]: L'évolution a suivi, grosso modo, celle des capacités de stockage et celle de la convivialité des logiciels.

[>R1]: Quelle a été la conséquence du fait que les logiciels soient devenus plus ergonomiques ?

[>R2]: Je vous donne un exemple. Moi, quand j'ai fait ce qu'on appelait à l'époque une maîtrise – l'équivalent du M1 –, j'ai tapé mon mémoire à la machine à écrire. L'arrivée des premiers ordinateurs dans le laboratoire, ça date de 1985-1986. C'était des Apple, je crois. Vous aviez une disquette avec un logiciel. Vous mettiez le logiciel dans le lecteur de disque 3,5 pouces. Puis, une fois que le logiciel était chargé, vous sortiez la disquette et vous remettiez une disquette de stockage. Vous pouviez passer une journée à travailler ; si vous oubliiez le soir de sauvegarder sur la disquette de stockage, avant de quitter le laboratoire, vous aviez passé une journée pour rien. Moi ça m'est arrivé une fois, quand j'écrivais ma thèse. J'étais venu un dimanche, parce que c'était un moment tranquille. J'ai passé mon

dimanche à bosser sur un chapitre de thèse et le soir j'ai éteint l'ordinateur. J'avais oublié de sauvegarder. Catastrophe ! Bon, ça ne m'est arrivé qu'une fois. Mais, c'est simplement pour vous dire qu'à cette époque-là on n'avait pas de système de vidéo tracking. Quand on faisait des tests comportementaux, on le faisait avec feuille, crayon, gomme. On notait des trucs. Avec, bien sûr, un nombre de variables extrêmement limité, parce que vous ne pouviez pas à la fois prendre des notes, contrôler douze chronomètres et regarder ce que faisait l'animal par rapport à chacune des variables. On stockait nos feuilles dans des classeurs. Aujourd'hui, tout est numérique.

[>R1]: Vous n'avez pas de cahiers de laboratoire ?

[>R2]: Si, bien sûr.

[>R1]: Ce sont des cahiers papier ?

[>R2]: Oui, papier.

[>R1]: Oui, en fait, ça a permis de gagner en précision et de collecter plus de données ?

[>R2]: Collecter des données, en attendant d'essayer d'en faire quelque chose. Les systèmes de vidéo tracking vous permettent de collecter 70, 80, 100 voire 150 variables en même temps. Qui peut le plus, peut le moins. Donc, autant collecter un maximum et stocker. Aujourd'hui on peut filmer des bestioles, on peut stocker les vidéos numériques, etc. Avant ça n'était pas possible. Moi, je me rappelle des premiers PC à disque dur qu'on avait au laboratoire : on était super content quand on avait 30 Mo. Super content ! 30 Mo de stockage ! [...] C'était un truc inimaginable. Alors, regardez aujourd'hui ! Donc, la manière dont nous avons traité nos données, dont nous avons stocké nos données, dont nous avons généré nos données, a été évidemment très fortement conditionnée par l'évolution des capacités de stockage et de calcul des outils informatiques. Avant tout autre chose.

Entretien avec le chercheur 26 (astronomie)

1 - 00:00 > 11:01 Thématique scientifique

[>R1]: Quel est votre domaine de recherche ?

[>R2]: Je m'intéresse aux galaxies, notamment pour essayer de comprendre comment les galaxies se sont formées, à partir du big bang, qui est l'origine de l'univers. Il y a eu des espèces de petites fluctuations primordiales de densité de matière, qui ont évolué jusqu'aux galaxies d'aujourd'hui, qui ont différentes formes, différentes tailles... Il y a des petites galaxies, des galaxies naines, des galaxies spirales, des galaxies elliptiques... Donc moi j'essaie de voir et de comprendre comment les galaxies se sont formées. Vous savez déjà ce qu'est une galaxie ?

[>R1]: Je ne saurais pas définir ce qu'est une galaxie. C'est un ensemble d'étoiles ?

[>R2]: Voilà, exactement. On pourrait dire un grand nombre d'étoiles. Notre propre galaxie, c'est la voie lactée. On est à l'intérieur de la voie lactée. C'est pour ça qu'on voit, quand il fait un petit peu noir, cette espèce de grande traînée dans le ciel, qui représente les étoiles de notre galaxie. Il y a d'autres galaxies comme notre voie lactée : la galaxie d'Andromède, etc. Il y en a des milliards. Elles ont différentes formes. Pour savoir comment il y a eu ces regroupements d'étoiles, il y a plusieurs méthodes. Il y a une méthode, qu'on utilise en astronomie, parce qu'on a la chance de pouvoir disposer de machines à remonter le temps. C'est-à-dire que, quand on fait une observation avec un télescope et qu'on voit des astres loin, comme la lumière met un temps limité, on va remonter dans le temps. Donc on va pouvoir voir les galaxies bébés, telles qu'elles étaient il y a dix milliards d'années par exemple. C'est une première méthode : remonter dans le temps. Simplement, ces galaxies qu'on voit il y a dix milliards d'années, on ne sait pas comment elles vont évoluer, parce qu'on ne peut pas les suivre. On a des instantanés de l'univers à différentes époques. On utilise de gros télescopes pour essayer de remonter dans le temps. La deuxième méthode, c'est celle qu'on pratique ici à l'observatoire. C'est ce qu'on appelle l'archéologie galactique. On va prendre les galaxies d'aujourd'hui, telles qu'elles sont. On va prendre notre voie lactée, on va prendre les galaxies les plus proches, et on va essayer de comprendre, de gratter, de sonder ces galaxies, pour voir

si elles auraient dans leur environnement des reliquats de leur histoire passée – par exemple des restes de collisions. L'idée standard c'est qu'une galaxie se forme par collisions successives de petites galaxies et ces collisions entre galaxies créent des débris. Donc on essaie de chercher ces débris, qu'on devrait encore voir actuellement. Cette image par exemple [il montre une photographie au mur], c'est une galaxie elliptique qu'on voit au centre. Avec une imagerie classique on verrait une elliptique. Avec des techniques d'observation un peu particulières, on voit tout autour : toutes ces espèces de structures qu'on voit, ces anneaux, qu'on ne connaissait pas avant, sont les restes de collisions passées. En étudiant les galaxies proches comme ça, en regardant ces structures, on peut en déduire que cette galaxie a subi une collision il y a quelques milliards d'années. Donc on part des objets d'aujourd'hui et on va remonter dans le temps en confrontant les observations avec des modèles/des simulations numériques de collisions de galaxies, pour essayer de voir à quoi elles ressemblaient dans le passé.

[>R1]: Quand il y a collision, il y a création d'une troisième galaxie ?

[>R2]: Exactement. Quand il y a une collision entre galaxies, on part de deux galaxies spirales par exemple. Dans les galaxies spirales, les étoiles tournent autour du centre de la galaxie. Elles tournent gentiment. Quand deux galaxies spirales se rencontrent, les étoiles elles-mêmes ne se rencontrent pas – une galaxie c'est plein d'étoiles, mais les distances entre les étoiles sont très importantes. Par contre, les orbites de ces étoiles vont être modifiées. Les étoiles vont sentir, disons, la masse de l'autre galaxie. Donc il y a ce qu'on appelle des effets de marée, comme on a par exemple sur Terre – les océans se soulèvent à cause de l'influence de la lune et du soleil. Au niveau des galaxies, il se passe un petit peu la même chose. Donc ces effets de marée vont perturber les orbites des étoiles. Une partie des étoiles va être arrachée de la galaxie. Elles vont former toutes les structures qu'on voit autour des galaxies. Et une autre partie des étoiles va prendre une espèce de mouvement complètement aléatoire. Les étoiles ne vont plus être en rotation. On va avoir une espèce de gaz d'étoile, si vous voulez. Du coup, la morphologie/la forme de la galaxie va changer et, plutôt que d'avoir un disque, la galaxie va s'épaissir et va former une espèce de bulbe, comme une pomme de terre, et on va former comme ça une galaxie elliptique. On s'est longtemps posé la question : pourquoi il y a des galaxies spirales et pourquoi il y a des galaxies elliptiques. L'un des scénarios c'est que les

galaxies elliptiques étaient des spirales auparavant, mais sont le résultat de la fusion de galaxies spirales.

[>R1]: Comment élaborez-vous un modèle ? Est-ce que vous utilisez des données de référence ? De quoi partez-vous ?

[>R2]: Au départ on est parti d'observations. Il y a notamment eu dans les années 1950 des grands sondages du ciel. Ce qu'on connaissait avant les années 1950, ce sont effectivement les galaxies elliptiques, les galaxies spirales, les galaxies naines. Et, dans ces sondages du ciel, on a vu qu'il y avait des objets ayant des formes bizarres, qui ne ressemblaient ni à une spirale, ni à une ellipse. On avait l'impression de voir deux galaxies côte à côte. Par exemple l'image là-bas [il montre une autre photographie au mur] : cette espèce de galaxie en forme de cœur. Donc on voyait des objets un peu bizarres comme ça et on se disait "mais qu'est-ce que c'est ?". Il a fallu un peu de temps pour comprendre que ce sont en fait des galaxies en cours de collision. Là ça paraît un peu évident, mais il y a des cas où, quand la fusion a eu lieu, on voit une galaxie avec des espèces de grandes coquilles, de queues qui s'en échappent, d'antennes. Et on ne savait pas du tout ce que c'était. Au centre de chaque galaxie, il y a un trou noir et, pendant longtemps, on pensait que c'était ce qui s'appelait à l'époque un monstre qui, au centre des galaxies, éjectait de la matière. Les formes bizarres étaient dues à un jet qui venait du trou noir au centre. Il a fallu attendre les années 1970 pour qu'on s'aperçoive que ces formes bizarres, en fait, n'étaient pas du tout liées à un monstre au centre des galaxies, mais étaient le résultat de simples effets de marées. On est habitué aux marées terrestres. Dans le cas des galaxies, on n'imaginait pas que les marées puissent former des espèces de grands tentacules autour des galaxies. Il a fallu des simulations sur ordinateur. Dans les années 1970, les premiers ordinateurs ont été capables de reproduire de façon assez réaliste ces structures-là. Ce qu'on faisait, c'était vraiment des simulations très simples : on représentait une galaxie avec des particules massives qui réagissaient simplement aux lois de la gravité et on les montait ensemble. On avait quelques centaines de particules qui réagissaient ensemble. Et on reproduisait comme ça, sur ordinateur, en utilisant les lois de la physique/de la gravitation, les formes observées. Et ça, ça a convaincu les astronomes. Ils ont été convaincus que ces formes qu'on voit ne sont pas dues au monstre mais à de simples effets de gravitation. Qui n'étaient pas intuitifs. On ne s'attendait pas à voir ce genre de formes. Du coup, les observations et les simulations sont vraiment très importantes. Tous les modèles qu'on a actuellement reposent à

la fois sur des données d'observations – qui d'une certaine façon sont la réalité –, qu'on cherche à reproduire, et des modèles théoriques couplés à des simulations sur ordinateur, qui vont voir si les simulations sont en accord avec les observations. Dans une simulation, il y a des conditions initiales que l'on ne connaît pas bien, donc on émet des hypothèses de départ. Si une simulation ne reproduit pas la réalité, elle sera vraisemblablement fausse. Soit la simulation n'est pas suffisamment précise, soit le modèle théorique derrière n'est pas bon. Donc on essaie de faire caler les simulations et les observations. Notamment on a besoin de simulations, parce que les échelles de temps sont très grandes. L'échelle de temps d'une collision entre galaxies c'est plusieurs centaines de millions d'années. Généralement on n'a pas la patience d'attendre cent millions d'années, pour voir comment va évoluer un système. Donc on va faire accélérer le processus sur un ordinateur.

2 - 11:01 > 18:33 Types de données et pratiques de partage

[>R1]: Utilisez-vous le terme "données" dans vos recherches ?

[>R2]: Oui. On l'utilise tout le temps. On a à la fois des données d'observation et des données de simulation. Donc on a à la fois des données d'un univers réel et des données d'un univers virtuel -mais qui sont aussi importantes. Maintenant, je pense, la moitié de la communauté travaille sur des données de simulation. A tel point parfois que, dans une conférence, quand on nous présente une image, on est incapable de savoir si c'est une vraie observation ou si c'est le résultat de simulations sur ordinateur, tellement elles deviennent réalistes.

[>R1]: Où récupérez-vous les données d'observation ?

[>R2]: Les données d'observation c'est quelque chose qui est unique. En tout cas, une chose pour laquelle l'astronomie a été précurseur, ça a été de partager les données. Pour acquérir des données d'observation, on fait une demande de temps de télescope avec un comité qui attribue ou non le temps de télescope. C'est quelque chose de très compétitif. Pour les plus gros télescopes, on a une chance sur cinq d'avoir le temps d'observation. Et à partir du moment où on a du temps d'observation, on peut faire son observation. Le plus souvent ce sont des techniciens et des ingénieurs qui vont faire des observations pour nous. On va récupérer les données. De ces données-là on est propriétaire pendant un an simplement. On a un an pour les exploiter. Après elles passent dans le domaine public – enfin, elles sont accessibles à tous.

Elles sont toutes archivées dans les bases de données qui sont associées à chaque télescope. Si j'ai observé au Chili, dans cet observatoire qui s'appelle l'ESO, c'est l'ESO qui dispose de la liste des données. Donc je sais à n'importe quel moment quelles observations sont prises. Même mes concurrents, je sais ce qu'ils sont en train de faire. Par contre, je ne peux récupérer les données qu'au bout d'un an.

[>R1]: Sous quelle forme ça se présente ?

[>R2]: Les données sont des fichiers informatiques, dans un format qui permet le partage, qui est compréhensible par tout le monde. Pour une image du ciel par exemple, on va avoir ce qu'on appelle l'astrométrie – il y a des descripteurs de l'image qui vont dire dans quelle direction on a pointé, à quel moment on a pris l'image, avec quel filtre on a pris cette image, etc. Donc le fichier va donner toutes sortes de conditions, qui font qu'il y a toutes les informations nécessaires pour pouvoir exploiter les données, même si ça n'est pas nous qui en sommes à l'origine. Ce format de fichier est le même quel que soit l'observatoire. Ça veut dire qu'on peut utiliser des logiciels et, peu importe si l'image vient d'un télescope de Hawaï ou du Chili, on aura la même manière d'exploiter ces données. Depuis très longtemps, depuis les années 1960, il y a un format qui est bien défini et qui permet d'échanger ces données.

[>R1]: A votre avis, pourquoi toutes ces pratiques de partage ont été mises en place ?

[>R2]: Je pense qu'il y a une raison. Strasbourg est très précurseur là-dessus, parce qu'il est à l'origine des formats d'échange des données. Le CDS est une base de référence mondiale pour l'échange des données. La raison pour laquelle il y a eu très tôt ce partage, c'est qu'on s'est rapidement rendu compte que construire des observatoires coûte cher. Un pays individuel n'a pas vraiment les moyens de construire un observatoire lui-même. Donc très rapidement, dès les années 1960, il y a eu cette idée de regrouper certains pays comme l'ESO, pour dire "okay, on va mettre ensemble de l'argent afin de construire de gros télescopes, qui seront concurrentiels par rapport à ceux des Américains par exemple". Donc il y a déjà cette idée de collaboration entre les laboratoires. Ce qui est très différent de ce qui se passe dans d'autres domaines de la recherche, où la science est faite dans le laboratoire et où on va avoir tendance à protéger ses données par rapport à d'autres. Là, on était obligé de travailler ensemble, pour pouvoir accéder à de gros moyens d'observation. Du coup, il y a très rapidement eu des collaborations internationales. En astronomie, il n'y a pas un papier qui soit purement français.

C'est un petit monde, donc ce ne sont que des collaborations internationales de toute façon. On travaille depuis longtemps ensemble, du coup l'idée d'échanger les données était naturelle. La deuxième raison, je pense, c'est que c'est facile. C'est facile de partager une image du ciel. Quand on a, dans d'autres domaines, des données d'archéologie par exemple, la base des recherches – un vase, etc. – est physiquement présente. Donc c'est difficile à partager. Pour nous, c'était facile de définir des protocoles d'échange. Parce que, globalement, pour l'observation d'un point dans le ciel, ce qui compte, ce sont ses coordonnées. Les coordonnées sont les mêmes partout dans le monde. De même, les filtres, c'est facile à définir. Donc il est relativement facile de décrire une donnée astronomique. Relativement, car il a quand même fallu mettre en place des protocoles, etc. Ça a mis un peu de temps. Mais c'était peut-être plus facile que dans d'autres domaines de la recherche. Comment décrire un vase de façon complètement objective ? Ça n'est pas facile. Tandis qu'une image, d'abord on peut la télécharger sur ordinateur –surtout avec Internet –, ensuite il suffit de quelques descripteurs pour pouvoir exploiter scientifiquement cette donnée. Donc je pense que ça a aussi contribué. En dehors du fait qu'il y avait de toute façon des collaborations internationales. Voilà, c'est peut-être un domaine de recherche où il est plus facile de partager les données.

3 - 18:33 > 35:09 Traitement des données et compétences

[>R1]: Dans vos recherches, vous partez d'observations qui existent déjà ? Ou bien vous arrive-t-il de demander du temps d'observation ?

[>R2]: Pendant longtemps, j'ai demandé du temps d'observation. Je partais très régulièrement observer. Je continue à le faire un petit peu, parce que le développement des instruments va très vite. On a des instruments de plus en plus performants. Et c'est vrai qu'on a tendance à vouloir utiliser le dernier instrument "à la mode", parce qu'on a un flot d'informations qui est beaucoup plus important. Auparavant on faisait une image. Maintenant on a des instruments qui permettent d'avoir une espèce de cube de données, donnant l'image mais aussi comment cette image varie en fonction de la longueur d'onde – donc on a des informations spectroscopiques. Ça nous permet de pouvoir interpréter beaucoup mieux les images. Donc on a tendance à utiliser les derniers instruments. Maintenant, si on regarde globalement, si on prend l'exemple du télescope spatial Hubble, la plupart des articles scientifiques ne sont pas

écrits par les gens qui ont fait la demande de temps de télescope et qui l'ont obtenue. Ils sont faits par des gens qui ont trouvé ces images-là dans les archives. On croule sous les données. Il y a les données des télescopes individuels. Il y a aussi de plus en plus ce qu'on appelle des grands sondages du ciel. C'est un consortium de chercheurs qui va dire : "Okay, on ne va pas pointer juste un seul objet. On va couvrir tout le ciel dans une certaine longueur d'onde". Donc là ça crée de grandes cartes du ciel, avec un flot d'informations énorme, que les chercheurs à l'origine de la demande ne pourraient pas exploiter à eux seuls. Donc ces données-là sont publiques. Et il y a énormément d'études qui sont faites, notamment des études statistiques. Par exemple, pour étudier quels sont les types de morphologie des galaxies, on a besoin de récolter des informations sur des millions de galaxies. Et bien, on va pouvoir le faire grâce à ces grands sondages. Dans ce cas-là, la plupart des papiers reposent sur ces grands sondages, donc sur des données qui sont présentes dans les archives. Maintenant c'est vraiment le mode standard. Il y a encore des demandes de temps de télescope – heureusement. Mais ça, c'est quelque chose qui est très bien, d'abord pour les étudiants. La plupart de mes données de thèse, je les ai obtenues quelques mois avant ma soutenance. Donc gros stress à la fin. Parce que durant toute ma thèse je partais à Hawaï, au Chili, les conditions météo étaient mauvaises... Donc c'était très frustrant. A l'époque, il n'y avait pas encore trop ces grands sondages. Même si on partageait déjà les données. Les données existaient, mais comme c'était le tout début d'Internet, il n'y avait pas de moyen de savoir facilement qu'elles existaient. Maintenant c'est très facile de récupérer les données. Ça n'était pas le cas avant. Donc on se basait essentiellement sur ses propres observations. Maintenant ça a complètement changé : quand on démarre une thèse, on est sûr qu'on aura les données qu'il faudra. Peut-être que ce ne sera pas exactement celles auxquelles on avait pensé au départ, parce qu'on n'obtient pas forcément le temps de télescope – on a une chance sur cinq, donc ça n'est pas évident. Mais on sait qu'il existe plein, plein de données. Par contre, la difficulté maintenant pour un étudiant ou pour un chercheur, ça n'est pas tellement d'acquérir les données, mais c'est de savoir créer des catalogues homogènes, consulter des bases pour savoir ce qui est disponible – puisque chaque observatoire a sa propre base de données. C'est pour ça que Strasbourg intervient aussi. Les services du CDS c'est aussi pour pouvoir mettre ensemble des données qui sont issues de plein de télescopes différents. Plutôt que d'aller voir dans chacun des

télescopes ce qui est disponible, les outils développés par le CDS permettent de savoir depuis une interface unique tout ce qui est disponible, quelle qu'en soit l'origine.

[>R1]: Les étudiants sont-ils formés à la recherche de données ?

[>R2]: Pas trop. Enfin, un petit peu. On a un parcours astrophysique ici. Parce qu'on est à Strasbourg, où il y a le CDS. Il y a une certaine sensibilisation. Mais on voudrait que ça aille plus loin. C'est vrai que les formations traditionnelles qu'on a en master sont très théoriques. Elles sont centrées autour de la physique et du traitement de données en tant que tels – i.e. quand on obtient une image brute du ciel, elle est inexploitable, donc il y a toutes sortes de traitements à faire. Donc on est habitué à faire ça et pas encore trop à exploiter ce flot de données. C'est une des raisons pour lesquelles on veut faire évoluer le master. Il y a cette réflexion à l'université. Parce que ça ne concerne pas seulement l'astrophysique. On est en train de vivre maintenant dans une autre époque, où on est submergé par les données. Il faut pouvoir les exploiter et de façon intelligente. C'est là qu'intervient l'intelligence artificielle, qui est un nouvel outil. Pour reprendre un domaine qui me concerne bien, la classification morphologique des galaxies, pendant très, très longtemps, c'est une classification qui se faisait à l'œil. Un petit groupe de personnes disait : "telle galaxie, on la caractérise à l'œil". On était expert. C'était facile. Là il commence à y avoir des limites. On n'est pas suffisamment nombreux, donc on va faire appel à ce qu'on appelle une contribution du public. Il existe des projets comme Galaxy Zoo, où on demande au public non expert – mais qu'on essaie de former – de classer. Ce sont des projets de citizens-scientists. A l'œil on pouvait classer des milliers de galaxies. Là on va pouvoir passer à quelques dizaines de milliers voire à un million de galaxies. Maintenant je travaille sur un sondage du ciel, qui va pouvoir générer les images d'un milliard de galaxies. Donc ça n'est plus possible. Là on est obligé de faire appel à des techniques de plus en plus utilisées d'apprentissage profond (deep learning). On utilise ce genre de méthodes comme des boîtes noires, parce que derrière on ne sait pas ce qu'il y a, on ne sait pas comment ça fonctionne. Donc là il y a un risque. C'est devenu tellement populaire et ça marche tellement bien que plutôt que de classer à l'œil, on va fournir aux algorithmes des millions d'images de galaxies spirales elliptiques. Soit on va lui dire ce qu'est une spirale et ce qu'est une elliptique sur le plus petit échantillon et il va pouvoir ensuite déterminer automatiquement sur un plus grand échantillon si ce sont des galaxies spirales ou elliptiques. Soit on va aller au-delà avec un apprentissage non supervisé, où c'est directement la machine

qui va dire "ça je vais appeler ça une spirale, ça je vais appeler ça une elliptique", sans lui avoir dit avant que c'était une galaxie spirale ou une galaxie elliptique. C'est comme si on lui présentait une image classique : on fournit des millions d'images de chiens et de chats, la machine est capable de dire si c'est un chien ou si c'est un chat, sans qu'on lui ait appris avant. Du coup, ça devient un petit peu inquiétant, parce que derrière il y a quand même des biais. Il y a des phases d'apprentissage certes, mais qui peuvent être complètement biaisées. Si pendant l'apprentissage on fournit à la machine des millions d'images de chats et trois images de chiens, forcément la machine va complètement ignorer les images de chiens et elle va dire que tout ce qu'on va lui présenter sont uniquement des chats. Donc il faut comprendre comment la machine fonctionne. Et ça, ça demande un minimum de formation. Et c'est cette formation-là qu'on n'a pas pour l'instant. On essaie d'interagir avec des spécialistes de l'intelligence artificielle, qui viennent d'autres domaines. Mais, à un moment, ces nouveaux outils sont tellement utilisés qu'il faudra absolument avoir été formé autour de l'exploitation de ces grandes quantités de données.

[>R1]: Quelles sont les personnes qui conçoivent les algorithmes ?

[>R2]: Pour l'instant, ce sont des ingénieurs ou des spécialistes de l'intelligence artificielle.

[>R1]: Qui ne sont pas astronomes ?

[>R2]: Qui ne sont pas astronomes. C'est-à-dire qu'on essaie de créer des collaborations. Il y a eu quelque chose d'assez symptomatique par exemple. Je parlais tout à l'heure du projet Galaxy Zoo, qui permet de classer un très grand nombre de galaxies par le "simple" citoyen. Il y a eu un challenge qui a été fait et qui a été de dire : on va essayer de voir quels sont les algorithmes capables de reproduire la classification qui a été faite à l'œil par des dizaines de milliers de personnes. L'algorithme qui a gagné n'était pas du tout conçu par des astronomes. Il a été conçu par un groupe qui avait développé des algorithmes pour reconnaître les numéros sur les plaques d'immatriculation – qui est un exercice encore plus complexe que de faire reconnaître une spirale d'une galaxie elliptique. Pour eux, c'était quelque chose de très facile et ils ont gagné haut la main avec leurs algorithmes, qui n'étaient pas du tout faits pour distinguer les spirales des elliptiques.

[>R1]: En ce qui concerne la partie modélisation, que faites-vous des données en sortie ? Est-ce qu'elles aussi sont mises en commun ?

[>R2]: C'est une très bonne question. Elles sont moins mises en commun. Il y a des efforts pour les mettre en commun. Notamment les simulations qu'on appelle cosmologiques, où l'on essaie de reproduire tout un pan d'univers sur de très gros calculateurs et qui génèrent des cubes de données énormes. Le chercheur qui génère ces cubes de données n'est pas capable de les exploiter complètement, parce qu'il y a énormément d'informations dedans. Idéalement, il faudrait pouvoir partager ça, mais c'est un peu plus difficile que pour les données d'observation. Il est facile de décrire les observations. Il suffit de donner les coordonnées alpha et delta, le temps de pose, etc. Ça suffit. Une simulation numérique, d'abord c'est un univers virtuel, donc il n'y a pas de coordonnées – ça n'a pas de sens de définir des coordonnées, puisque c'est virtuel. Après, pour une simulation numérique, le nombre de descripteurs est beaucoup plus important, puisqu'il faudrait idéalement, pour comprendre la simulation numérique, connaître toutes les conditions initiales, c'est-à-dire qu'il faudrait qu'il y ait un descriptif de toutes les recettes qui ont été mises dans la simulation numérique pour pouvoir la faire tourner. Il y a plein de petits trucs/d'astuces/de recettes de cuisine, d'une certaine façon, dans les simulations numériques. Il peut même y avoir des choses qui peuvent nous paraître aberrantes. Par exemple, vous savez que la vitesse de la lumière est fixe. Elle est de 300 000 km/s. Elle ne peut pas bouger. Mais dans les simulations numériques on fait varier la vitesse de la lumière pour des questions purement numériques. Ça n'a aucun sens physique, mais ça permet d'améliorer certains algorithmes, etc. Donc on met dans les simulations numériques plein de trucs contre-intuitifs, qu'il faudrait expliquer dans un descriptif. Et là ça devient plus compliqué. Pour l'instant, il y a des efforts qui sont faits et qui tentent de définir des protocoles d'échange des simulations numériques, mais ça reste un petit peu compliqué. Donc il vaut mieux discuter avec l'équipe et essayer de comprendre. Mais c'est un vrai enjeu. Parce que, comme je disais, ces simulations sont de plus en plus grandes. Donc la personne qui va exploiter ces simulations numériques n'est pas forcément celle qui les a faites tourner. Cette personne-là va prendre la simulation, pensant que c'est le vrai univers, d'une certaine façon. Et c'est tellement réaliste qu'on a vraiment l'impression d'avoir un vrai pan d'univers. Le risque c'est qu'on interprète mal certains phénomènes observés dans ces simulations numériques. Celui qui a développé le code sait que ce sont des artefacts purement numériques. Maintenant, celui qui les exploite et qui n'a pas toutes les informations risque de l'oublier. Donc il y a là un vrai danger, qui vient du fait que les simulations sont quelque

chose de compliqué, qu'il y a beaucoup de paramètres derrière une simulation et que ces paramètres ne sont pas forcément connus de la personne qui va l'utiliser.

[>R1]: J'ai du mal à comprendre ce qu'il y a à analyser dans les données en sortie de simulation.

[>R2]: En sortie de simulation, c'est comme si on avait des images du ciel à différentes longueurs d'onde et variant avec le temps. La simulation va évoluer avec le temps. Pour prendre un exemple, dans la simulation, on va représenter les étoiles par des particules. Dans les galaxies il y a de nouvelles étoiles qui se forment. Donc on va rajouter des particules pour les étoiles qui se forment. Les étoiles se forment à partir de nuages de gaz. Donc, dans la simulation, on va utiliser ce qu'on appelle des codes hydrodynamiques, qui vont représenter les nuages de gaz, qui plus tard formeront les étoiles. A un moment donné, dans la simulation, on va avoir ce qu'on appelle un cube de données, avec la distribution à un moment donné de toutes les étoiles – les étoiles vieilles, les étoiles jeunes –, la distribution du gaz autour de ces étoiles, la distribution de ce qu'on appelle la matière noire, qui est une des composantes importantes de la galaxie, dont on ignore l'origine mais dont on a besoin dans les simulations. Les données ça va être un instantané de l'univers simulé. On va demander : "génère-moi une carte avec l'ensemble des étoiles" ou "génère-moi une carte avec l'ensemble du gaz". Et on va vous dire ça à différents moments. On dit : "génère-moi la même carte au bout d'un million d'années simulées". Et après on va pouvoir comparer. Une fois qu'on a fait ça, on ne va pas pouvoir tout analyser. On va dire : "Tiens, il y a cet objet-là dans la simulation qui me paraît intéressant, cette galaxie qui est en train de se former. Je vais la suivre en fonction du temps". Donc on fait ensuite une espèce de zoom sur une région particulière et on va voir comment évolue ce système. Ou on va plutôt faire des statistiques : je vais compter dans cette simulation le nombre de galaxies qui ont des formes très irrégulières et je vais voir comment elles évoluent en fonction du temps.

[>R1]: D'accord. On part du postulat que le modèle est vrai.

[>R2]: Exactement. Et, à partir du modèle, on va générer une image qui ressemble à une vraie image du ciel. Comme ça, on va pouvoir l'exploiter comme s'il s'agissait de vraies images. On va même parfois pouvoir utiliser les mêmes algorithmes d'analyse d'images, que celles-ci viennent d'une simulation ou d'une vraie observation.

4 - 35:09 > 42:04 Conservation et citation des données de simulation

[>R1]: Est-ce qu'il y a un système de citation ? Quand vous analysez les données de simulation d'un autre chercheur, puis que vous rédigez un article, citez-vous son travail ?

[>R2]: Oui, bien sûr. On doit donner des références. Donc si on utilise une simulation du millénium – c'est une des simulations qu'on peut utiliser –, on va citer l'article qui présente cette simulation-là.

[>R1]: Actuellement existe-t-il une banque de données pour les données de simulation ?

[>R2]: Oui, je comprends. Même pour les observations, actuellement, il y a une tendance qui consiste à ne pas forcément citer l'auteur qui présente la simulation mais à citer un DOI, qui correspondra à l'observation elle-même. On n'est qu'au tout début de ça. Je pense qu'il y a une sorte de réticence aussi, parce que du coup on perd... La façon dont on est évalué en tant que chercheur, c'est en fonction du nombre de citations. Donc si on cite uniquement les données et plus le travail de l'observateur, même si je disais tout à l'heure que les observations sont publiques, souvent, si on sait que les observations ont été prises par telle personne et que ce sondage a été décrit par telle personne, même si on n'est pas obligé de le faire, généralement on cite un article précédent. Maintenant, effectivement, si la tendance est de citer uniquement la référence/le DOI, qui est éternel, on risque de perdre... Donc je pense qu'il y aura une réticence dans la communauté. Pour les simulations, ce sera un peu pareil. Sauf qu'il n'y a pas encore vraiment de bases de données de simulation qui soient unifiées. Ce qui se passe généralement, c'est qu'une simulation est faite. Il y a une forte incitation à ce que cette simulation devienne publique. Mais il n'y a pas l'obligation qui existe pour les observations. Pour les observations, au bout d'un an, quel que soit le télescope, on perd la propriété. Donc, pour les simulations, c'est un peu ambigu. De plus en plus, effectivement, quand on fait des demandes de calcul sur les très grands calculateurs – de même qu'on fait des demandes d'observation, on fait des demandes de calcul, qui sont compétitives –, il y a une incitation à ce que ces données-là deviennent publiques. Ça n'est pas encore obligatoire. Ça le devient au niveau de tous les projets européens – là ça va le devenir. Ce qui se passe, c'est que cette simulation est mise sur un site web du laboratoire, mais pas dans des conditions de conservation pérenne sur le très long terme. Donc, après, si la personne quitte le laboratoire, s'il n'y a plus de support... Tandis que, pour les observations, ce que fait le CDS permet que

dans vingt ou trente les données soient encore là. Ce qui se passe aussi dans les grandes bases de chacun des observatoires, c'est que c'est quelque chose qui a vocation à être pérenne. Pour les simulations, ça n'est pas encore le cas. Il y a des efforts pour aller dans ce sens-là, pour avoir des plateformes communes de mises à disposition des simulations, mais on n'y est pas encore.

[>R1]: Est-ce pensé comme des banques de données qui seraient annexes aux centres de calcul ?

[>R2]: Oui. Pas forcément pour le stockage et l'archivage à long terme. Pour les simulations sur ordinateur, on a besoin des grands centres de calcul. Mais ça génère des flots de données immenses. L'une des difficultés actuellement pour les simulations numériques, c'est de pouvoir rapatrier ces données. Parce que ça prend trop de temps. On n'est pas capable de le faire par Internet. Donc les données sont rapatriées en allant les chercher dans le centre de calcul et en les ramenant, parce que ce sont des flots de données trop importants. Pour éviter ça, l'idée c'est d'associer aux centres de calcul un centre de données, qui va permettre d'exploiter ces données-là à distance – sans avoir à les rapatrier chez soi, parce qu'on n'y arrive pas. Les débits ne sont pas suffisamment importants. Il y a une idée actuellement de rapprocher les centres de données des supercalculateurs. La raison première c'est l'exploitation des données. La deuxième raison, qui est peut-être encore plus importante, c'est que la donnée est au centre de tout ce qui est intelligence artificielle -puisque celle-ci repose sur l'analyse d'un très grand flot de données. Donc là il n'y a pas le choix. Le calculateur ne va pas aller chercher des données qui sont très loin, parce que ça prendrait trop de temps. Donc il y a aussi cette tendance actuelle d'essayer de rapprocher les supercalculateurs des centres de données, qui sont pour l'instant plutôt séparés.

[>R1]: Est-ce qu'il s'agit de centres de données qui conservent les données de manière pérenne ?

[>R2]: Pas pour l'instant. Là on est plutôt dans des centres qui répondent à des besoins de calcul et d'analyse. Ça n'est pas encore la vocation de ces centres de conserver les données de façon pérenne. Il y a d'autres grands centres de données qui se développent. Il y a notamment un plan gouvernemental pour créer des centres de données régionaux labellisés, pour conserver les données sur le plus long terme.

[>R1]: Ce sont les datacenters ?

[>R2]: Oui, ce sont les datacenters, comme il y en a un ici – où il n'y a pas vraiment de supercalculateur. Pour l'instant, nous, on a notre propre salle de serveurs dans les locaux de l'observatoire. Mais il y a une forte pression de l'université pour qu'on mette nos données dans le datacenter qui est en train d'être construit. Donc ça se fera à terme. On va commencer petit à petit. Donc il y a effectivement une idée d'essayer de regrouper les centres de données pour d'abord avoir un niveau de fiabilité qui est plus important que ce qu'on peut avoir ici – les salles sont plus grandes, la protection pour les incendies répond à un standard plus élevé... Mais ça a aussi un coût plus élevé. Donc on peut imaginer qu'avec le développement des datacenters, les données seront archivées de façon plus pérenne.

5 - 42:04 > 46:48 Standards d'échange de données

[>R1]: En quoi consiste l'observatoire virtuel ?

[>R2]: Je pense que vous aurez vraiment intérêt à aller discuter avec les spécialistes du CDS. Globalement, l'observatoire virtuel c'est un ensemble de protocoles d'échange des données. Je disais tout à l'heure qu'en astronomie il y a des descripteurs des données, qui existent depuis longtemps. L'observatoire virtuel c'est une version un peu améliorée de ces descripteurs, qui sont encore plus universels que ceux qu'on utilise actuellement. Ils sont potentiellement utilisables par toutes sortes de logiciels, par toutes les bases de données. Pour l'instant, on avait un format d'échange qui était le format FITS. C'était un format qui était un petit peu propriétaire. L'observatoire virtuel a vraiment pour but de définir des protocoles mais qui sont définis à un niveau organisé complètement international. Les protocoles qu'on a utilisés jusqu'à présent sont des protocoles un petit peu pragmatiques, utilisés par tous. Mais il n'y a pas un organisme derrière qui les certifie. Dans l'observatoire virtuel, il y a vraiment des commissions : un protocole est défini et validé de façon internationale, et il fait force de loi. C'est ça l'observatoire virtuel. Globalement, l'idée c'est de pouvoir récupérer des données d'un peu partout, de n'importe quel télescope, de les mettre ensemble, sans qu'on sache leur origine d'une certaine façon. En pratique, derrière l'observatoire virtuel, il y a ces protocoles d'échange, très normatifs. C'est un ensemble d'outils, comme ceux développés par le CDS, avec des interfaces comme Aladin. Mais ça n'est pas un observatoire physique. Ce sont

vraiment des protocoles d'échange de données. La nouveauté c'est que ce sont des protocoles conçus au niveau international.

[>R1]: Y a-t-il également une réflexion sur les données de simulation dans l'observatoire virtuel ?

[>R2]: Il y a eu des réflexions de faites. Il y a des groupes de travail à l'université qui ont essayé de réfléchir là-dessus. Il y a A. par exemple, chez nous, qui développe des simulations numériques. Il a participé à un groupe de travail sur ces questions. Il y a aussi B., qui s'occupe un petit peu de ça. Il y a des tentatives, mais – je le disais tout à l'heure – c'est plus compliqué, parce qu'il n'y a pas de descripteurs évidents. Donc on y va petit à petit. Il y a un premier protocole qui avait été défini, mais qui était une usine à gaz et qui a peu été utilisé. On en est encore aux balbutiements, disons.

[>R1]: Etes-vous plutôt du côté de la production de modèles ou de la réutilisation de données de simulation ?

[>R2]: Moi je suis observateur de formation. Donc j'utilise des données du ciel. Mais j'ai toujours travaillé, depuis très longtemps, avec des numériciens. Je suis incapable de faire tourner des codes numériques. Par contre, je peux les interpréter. Mon travail c'est vraiment de comparer observations et simulations. Mais mon expertise technique est plus autour des observations.

[>R1]: "Numéricien", qu'est-ce que ça veut dire ?

[>R2]: Numéricien c'est celui qui fait les simulations numériques sur ordinateur.

6 - 46:48 > 53:25 Financement des observatoires

[>R1]: Quand vous demandez du temps d'observation, une fois sélectionné, accédez-vous gratuitement aux données ?

[>R2]: Oui, c'est l'avantage qu'on a, en fait. En astronomie, c'est gratuit depuis très longtemps, parce qu'on a partagé les moyens d'observation. Le modèle standard c'est que c'est complètement gratuit. Même quand on obtient du temps de télescope au Chili, si on est amené à se déplacer, le voyage est pris en charge par l'organisme européen. Ça ne coûte rien au laboratoire. Donc c'est un gros avantage. Si on n'est pas amené à se déplacer – de plus en plus

on a des observations qui sont faites en mode service, donc on ne se déplace plus pour observer –, là on reçoit les données de manière totalement gratuite. Après, ça n'est pas complètement gratuit : c'est au niveau gouvernemental que la France contribue chaque année au financement de cet organisme européen. Ça n'est pas au niveau du laboratoire mais au niveau de l'État français, du Ministère des affaires étrangères. C'est ce qu'on appelle les grandes infrastructures de recherche. Du coup, ça n'est plus aux chercheurs de payer. C'est payé au niveau au-dessus. Après, ça va même au-delà. La tendance c'est aussi de dire : "on peut utiliser même des observatoires américains, c'est gratuit". Parce qu'il y a une espèce de réciprocité. Moi j'utilise un radiotélescope au Mexique. Je n'ai pas à payer pour ça. Ce sont des comités internationaux. Il peut y avoir des restrictions d'accès. C'est-à-dire qu'il y a parfois du temps international et du temps national. Mais on collabore avec tout le monde, donc généralement on arrive à passer à travers. Mais normalement, le temps de télescope des grands observatoires européens est réservé aux pays qui contribuent. Ce qui est normal. Ce qui va se passer, c'est que, quand il va falloir programmer le temps d'observation, s'il y a deux demandes, une qui vient de l'Europe et une qui est internationale ou qui vient d'un pays qui n'a pas contribué, il y aura une préférence pour le pays qui a contribué. Mais ça n'est pas systématiquement refusé. Ça dépend un petit peu de la compétition. Il s'agit de trouver un équilibre. Et, globalement il y a des facilités un peu partout, des espèces d'accords tacites. Il y a quelques exceptions : par exemple, des grands sondages qui se mettent en place, pour lesquels, comme ça coûte très cher, on demande d'avoir des tickets d'entrée. Un laboratoire va investir de l'argent et donner quelques centaines de milliers d'euros pour participer à un grand projet. Du coup, les chercheurs de ce laboratoire auront accès aux données. Donc ça n'est pas au cas par cas. On ne va pas "acheter "les données. Mais le laboratoire va disposer de tickets d'entrée. C'est quelque chose qui était très peu courant dans le passé, qui commence un peu. Il y a le LSST par exemple, qui est un grand télescope, qui va faire un sondage panoramique du ciel d'ici quelques années. Là c'est ce mode de financement. Les données vont être publiques au bout d'un an. Donc on peut patienter. Mais ce seront les données brutes, donc elles seront très peu exploitables. Si on veut profiter de données qui soient avancées, là il faut avoir des tickets d'entrée. Il y a certaines missions spatiales qui fonctionnent sur le même principe. Pendant un an ou deux, les données ne sont pas complètement publiques ou bien ce sont

seulement les données brutes qui sont publiques et on n'a pas les moyens pour les traiter. Là il peut y avoir des enjeux financiers.

[>R1]: Comment fonctionne l'allocation de temps d'observation entre l'ESA et la NASA par exemple ?

[>R2]: Il y a des accords. Généralement ce ne sont pas des accords financiers en tant que tels. C'est une contribution à la construction d'une partie de l'instrument. Par exemple, on associe le télescope spatial Hubble essentiellement aux Américains, mais il y a 20% de Hubble qui a été construit par les Européens. Donc il y a des contributions qu'on appelle [inaudible] : c'est du travail qui a été fait par les Européens sur des missions américaines. Et c'est pratiquement le cas chaque fois. Et vice versa : il y a des missions qui sont essentiellement européennes – Euclid par exemple, qui va être lancé d'ici quelques années –, mais les Américains apportent par exemple une caméra infrarouge et donc un certain nombre de chercheurs américains font officiellement partie du consortium et ont le droit d'accéder aux données, à cause de cette contribution-là. L'ESA ne va pas dire : "je verse tant de millions d'euros juste comme ça", parce que de toute façon l'ESA n'a pas vocation à faire ça. L'argent qu'on donne à l'ESA, c'est pour qu'il soit réinvesti dans des boîtes, dans des laboratoires. Donc ce que va dire l'ESA c'est : "pour cette mission-là, je vais apporter tel instrument". Après, le temps est globalisé et donc les Européens n'auront pas uniquement accès à cet instrument mais ils auront accès à tout l'observatoire. Et, de même, les Américains auront accès à l'observatoire européen.

Entretien avec le chercheur 37 (géologie)

1 - 00:00 > 04:23 Visite de Bruno Latour

[…]

2 - 04:25 > 09:00 Ressources humaines

[>R1]: Pour faciliter la discussion, je me focaliserai sur le projet H. Je diviserai l'entretien en deux parties : une partie sur le contexte du projet de recherche ; puis une partie sur les données du projet et la façon dont elles sont gérées. J'ai d'abord une toute première question : vous avez un statut un peu spécial – CNAP, c'est ça ?

[>R2]: Ah, vous êtes bien renseignée !

[>R1]: Est-ce que vous pourriez me dire en quoi ça consiste ?

[>R2]: Le CNAP c'est le Corps National des Astrophysiciens et des Physiciens. C'est un corps assez ancien, qui a été créé pour que des chercheurs puissent dédier une partie de leur temps à la notion d'observatoire et d'observation. Un chercheur, par exemple, va passer 100% de son temps sur de la recherche scientifique. Un enseignant-chercheur va consacrer la moitié de son temps à l'enseignement et l'autre moitié à la recherche. Quelqu'un qui est CNAP, lui, va passer un tiers de son temps en recherche, un tiers en enseignement et un tiers avec une charge d'observation. Par exemple, quelqu'un qui est dans un observatoire de sismique va faire de la surveillance sismique. Il va passer du temps à regarder les capteurs, à faire de la veille, à faire du risque, à regarder s'il y a des enregistrements qui sont suspects et, si oui, comment cela s'explique, etc. Et surtout, dans notre statut, nous avons la charge de rendre un certain nombre de nos données publiques via des bases de données qui, grâce aux outils numériques, sont maintenant disponibles via internet. Si, par exemple, vous allez sur le site internet de notre observatoire, vous trouverez le lien vers notre base de données. Nous, on fait de la surveillance environnementale. Donc on va regarder les pluies et les ruisseaux. Dans la base de données, vous pouvez rechercher par exemple la composition chimique de toutes les pluies entre le 1er janvier et le 31 décembre 2003 (nous, on surveille depuis 1986). Ou bien,

tiens, le ruisseau, qu'est-ce que ça a donné ? Donc vous rentrez le ruisseau et vous avez toutes les données.

[>R1]: Est-ce que vous encadrez des doctorants ?

[>R2]: Bien sûr. Comme on est chercheur, on encadre des doctorants. Dans le cadre du projet H., par exemple, il y a trois bourses de thèse, qui sont associées à trois grandes thématiques de travail.

[>R1]: Ces doctorants sont tous rattachés à l'EOST ?

[>R2]: Oui, ils sont tous ici, à Strasbourg. Il y en a un en géochimie avec moi. Les autres sont dans d'autres unités, en géophysique et en hydrologie.

3 - 09:00 > 16:00 Objectifs scientifiques

[>R1]: Est-ce que vous pourriez me résumer, de manière assez simple, les objectifs scientifiques du projet H. ?

[>R2]: Tout à fait. Ce projet porte sur la question de la ressource en eau dans les zones de montagne. Dans les zones de montagne, l'eau (pour l'eau potable, l'agriculture, l'industrie) provient des sources ou des petits ruisseaux de montagne. Dans la plaine, vous pouvez prendre l'eau de la nappe phréatique. La nappe n'est pas un réservoir infini mais il est énorme. Donc, dans la plaine d'Alsace, il n'y aura jamais de problème de pénurie d'eau, parce qu'il y a cette nappe qui est une immense ressource d'eau. Dans les zones de montagne, le « château d'eau » c'est la montagne elle-même. La question qu'on se pose actuellement est liée au changement climatique, notamment au changement du régime pluviométrique des pluies. Par exemple, on sait que dans le Nord-Est de la France la couverture neigeuse va tendre à disparaître. Or la couverture neigeuse, lorsqu'elle fond après l'hiver, va s'infiltrer dans le sous-sol, aller dans les fractures de la roche et rester dans cette porosité de la roche. Ce qui va faire que, même pendant l'été, même s'il ne pleut pas pendant un ou deux mois, il y aura toujours de l'eau dans les sources. Donc la question qu'on se pose c'est quel impact va avoir le changement climatique sur la ressource en eau. Je vous ai parlé de la neige, mais on pourrait aussi parler de la forme du régime des pluies. S'il pleut tous les mois de l'année, de manière assez douce, l'eau va avoir le temps de rentrer dans le sol et de s'infiltrer. Par contre, si vous

avez des pluies violentes, ce sont de grosses quantités d'eau qui vont arriver en une seule fois et qui, au lieu de s'infiltrer, vont ruisseler. Alors, ça n'est pas perdu pour les zones en aval (les plaine) ; par contre, pour les zones en amont (les montagnes) c'est perdu. Voilà. Donc, nous, on travaille là-dessus. Et pour comprendre cette ressource en eau, il faut qu'on comprenne la structure de la montagne. Ça c'est le premier volet de la question : quelle est la structure du réservoir/château d'eau ? Donc est-ce que la montagne est poreuse, est-ce qu'il y a des poches, quelle est la porosité de la roche, c'est-à-dire quelle est la capacité de stockage en eau, où est-ce que l'eau se recharge, etc. ? Donc on va combiner un certain nombre d'outils géophysiques pour caractériser la structure de la montagne. Ensuite, le deuxième volet consiste à modéliser. Donc on va faire un suivi hydrologique spatio-temporel pour essayer de modéliser le flux d'eau. Et puis il y a aussi la question de la composition donc de la qualité de l'eau – c'est le troisième volet. Est-ce que la qualité de l'eau varie ? En fonction de la durée de résidence de l'eau dans un système, les processus d'interaction avec les minéraux ne sont pas les mêmes. Plus le temps de résidence est long, plus il y a d'interactions, plus l'eau va être chargée. On travaille aussi sur l'impact des travaux forestiers et de la sylviculture sur le sol, l'eau, etc. L'eau transporte des matières dissoutes, mais elle transporte aussi des sédiments et des matières en suspension (des particules d'argile ou de matières organiques), qui interagissent fortement avec un certain nombre de contaminants. Il n'y a pas forcément de contaminants dans notre système, mais il y a tous les contaminants atmosphériques qui arrivent. L'atmosphère, comme on la pollue malheureusement beaucoup, amène du plomb, des acides, des composés organiques et toutes sortes de choses. Et cela va avoir une affinité particulière avec les argiles et les matières organiques qui sont transportés par le ruisseau sous forme de matières en suspension. Donc, pour résumer, on travaille sur (1) la géométrie du site, (2) l'hydrologie et (3) la géochimie (c'est-à-dire la composition/la qualité de l'eau).

[>R1]: Pouvez-vous me réexpliquer quel est le but de l'axe « modélisation » ?

[>R2]: Oui, tout à fait. La modélisation va nous permettre de comprendre comment fonctionne le système. Un bassin versant ou une fosse de château d'eau, c'est quoi ? C'est comme une boîte noire : vous avez les entrées (ce sont les pluies) et vous avez ce qui sort (c'est ce qui sort des sources ou du ruisseau). Nous, on connaît ce qui rentre, on connaît ce qui sort. Donc on va essayer de comprendre ce qui se passe à l'intérieur. Pour cela, on va faire des modèles, c'est-à-dire des fonctions numériques : y=f (x). Le x c'est là [elle montre le haut, ce

qui rentre] et le y c'est là [elle montre le bas, ce qui sort]. On va essayer de trouver les mécanismes qui contrôlent la signature. Et si on arrive à modéliser, on dira « voilà, quand il y a une pluie qui a telle forme, on a telle sortie ». Ça c'est déjà un gros, gros travail. Une fois qu'on y est parvenu, on se dit : « Maintenant, j'imagine que mon changement climatique fait, par exemple, qu'il ne pleut pas pendant un mois. Qu'est-ce qui va se passer sur ma sortie ? Quel signal je vais avoir ? ». Donc ça c'est nécessaire. Modéliser, ça permet (1) de comprendre ce qui se passe dans le « château d'eau » et (2) de pouvoir prédire le futur. On utilise les scénarios fournis par le GIEC : on entre des scénarios de pluies ou de températures dans notre modèle et, comme ça, on voit quel impact ça a.

[>R1]: Quel est le quatrième axe qui est mentionné sur le site de l'ANR ?

[>R2]: C'est l'impact du changement climatique.

[>R1]: C'est un peu la conclusion ?

[>R2]: Oui, c'est ça, c'est la conclusion.

4 - 16:00 > 19:52 Partenaires du projet

[>R1]: Quels sont exactement les partenaires impliqués dans le projet ? Sur le site de l'ANR, j'en ai vus trois ; mais dans le résumé du projet il est dit qu'il y en a neuf.

[>R2]: Il y a bien 9 partenaires. En fait, dans les projets ANR, il y a les partenaires à qui on va donner une somme d'argent. Mais il ne peut pas y avoir neuf partenaires (ça n'est pas possible). C'est un peu technique, c'est administratif. Dans le projet H., il y a quatre partenaires qui ont reçu une somme directement : le BRGM, l'INRA, l'Institut de Physique du Globe et le L. Après, dans le volet 1 par exemple, l'IPGES va travailler avec des sismologues venant d'un autre centre de recherche. Ces personnes ne sont donc pas partenaires du projet, mais elles sont associées. On va faire appel à elles pour certaines missions. Elles seront payées, etc. On va « sous-traiter ». Mais l'ANR ne leur donnera pas directement une somme. C'est la différence entre les partenaires et les laboratoires associés.

[>R1]: Est-ce que, vous, vous rémunérez ces chercheurs pour ce qu'ils font ou bien est-ce qu'il s'agit d'une « collaboration », où vous co-publiez...?

[>R2]: Oui, c'est une collaboration. On les rémunère dans le sens où, si par exemple ils ont besoin d'une sonde, on va acheter la sonde ensemble. Quand ils vont venir faire des mesures sur le terrain, on va prendre en charge leurs frais de mission. S'il y a du fonctionnement, on va prendre en charge le fonctionnement. Voilà comment ça se passe.

[>R1]: Et ensuite les résultats seront publiés en co-publication ?

[>R2]: Oui. Dans le projet ANR, on a aussi un partenaire privé, qui est un cabinet d'étude de sols. Là c'est une prestation : pour 10 000€ on travaille ensemble, mais ils nous rendent une carte pédologique du bassin. C'est une prestation de service. On a prévu un certain budget pour un certain rendu. Donc ça n'est pas un partenaire.

[>R1]: Comment ça se passe au niveau des données ? Au terme de leur mission, est-ce qu'ils vous donnent les données et en sont dépossédés ?

[>R2]: Ça dépend. Quand c'est une prestation de service, on finance et, à la fin, on a un rapport et on devient les « dépositaires » de la donnée. Parce que c'est un cabinet d'études et non un scientifique. Si on fait une publication, le cabinet sera associé à la publication et, bien sûr, on discutera scientifiquement du contenu de l'article. Mais, les données, elles sont pour nous. Pour les collaborations scientifiques, en revanche, ça fonctionne comme dans toutes les collaborations scientifiques. C'est-à-dire que, si l'équipe 1 et l'équipe 2 travaillent ensemble, elles vont s'échanger les données et essayer de publier ensemble.

[>R1]: Oui, c'est une copropriété des données.

[>R2]: Voilà.

5 - 19:52 > 36:27 Répartition des rôles

[>R1]: Comment sont répartis les rôles entre les différents partenaires (je parle des principaux partenaires – ceux qui ont reçu une somme d'argent de l'ANR) ?

[>R2]: En fait, on travaille sous forme de workpackage (WP). On a 4 workpackages, mais on peut dire qu'il y en a trois principaux (le dernier c'est la modélisation finale avec les changements climatiques).

[>R1]: Chaque workpackage correspond à un des objectifs que vous m'avez décrits tout à l'heure ?

[>R2]: Oui, exactement. Chaque partenaire a son financement, qui correspond soit à des frais d'analyse, soit à des frais d'équipement ou de matériel, soit au financement d'un doctorant. Ici, chaque workpackage a un doctorant. Le projet a été écrit comme ça. Pour remettre un peu le projet dans son contexte, il y a eu en 2014 ou 2015 un EquipEx, CRITEX, qui a été financé pour l'équipement de la zone critique. Les EquipEx sont des projets sur 8 ans. CRITEX regroupe des sites d'observation comme le bassin versant du Strengbach (i.e. le site de l'observatoire) un peu partout dans le monde (en Bretagne, au Laos, en Inde, etc.). L'idée c'est de développer des équipements nouveaux pour essayer de mieux comprendre et de mieux caractériser la zone critique. Donc de mieux caractériser le sol et le sous-sol, de mieux caractériser les échanges d'énergie, de gaz et d'eau entre l'atmosphère et la canopée. Il y a tout un groupe de travail qui étudie les transferts entre la végétation, le sol et l'atmosphère. D'autres groupes essaient de caractériser la structure du sous-sol, avec des outils de sismique par exemple. L'idée de l'EquipEx CRITEX c'était d'associer des outils avec des sites. Au total, en France et dans le monde, il y a 22 sites naturels surveillés. L'idée c'était d'en choisir trois sur lesquels on allait mettre le maximum d'équipements. Sur les trois sites qui ont été choisis, il y a le bassin versant du Strengbach, il y a un site en région parisienne à Orgeval et il y a un site en Bretagne, Agrhys. Grâce à ce financement, on a donc pu bénéficier d'un certain nombre d'équipements. Après, les équipements ne financent pas les analyses, les doctorants, etc. Donc l'objet de cet ANR ça a vraiment été de venir en support de ces équipements.

[>R1]: Il y avait donc déjà une collaboration autour de cet EquipEx ?

[>R2]: Oui, il y avait déjà, si vous voulez, une espèce de communauté. Et puis, le bassin versant du Strengbach est équipé depuis 1986, donc c'est un site qui est maintenant bien caractérisé, qui a des historiques. Il y a cinq ans, on a commencé à travailler avec des géophysiciens. Les géophysiciens n'ont pas forcément l'habitude de travailler sur des sites en montagne avec de la pente, des arbres, etc. (ils travaillent souvent la terre plus profonde), donc il a fallu adapter, faire des développements méthodologiques, etc. Je pense que le projet H. a été élu par l'ANR également parce que les reviewers ont vu que c'était un projet qui ne sortait pas de nulle part, qu'il y avait déjà des gens qui travaillaient dessus. Ils savaient qu'il

n'y avait pas beaucoup de risques, qu'il y avait du support, qu'il y avait des équipements, qu'il y avait déjà une communauté et des collaborations effectives, donc qu'il y aurait des résultats. Je pense que ce qui a fait qu'on a été financé, c'est le fait que c'est un site bien caractérisé, sur lequel travaille déjà une communauté. Bon, et puis on n'a pas été sélectionné la première année ; on a été sélectionné la deuxième année. On n'a pas réussi tout de suite.

[>R1]: Est-ce que le fait que vous soyez plusieurs partenaires a joué?

[>R2]: Oui, oui, bien sûr. C'est nécessaire.

[>R1]: C'est obligatoire ?

[>R2]: Oui, oui, c'est obligatoire.

[>R1]: Donc vous me disiez que l'Institut de Physique du Globe avait en charge le premier axe du projet.

[>R2]: C'est ça, exactement. Le deuxième axe c'est ici, au L., mais c'est l'équipe de modélisation qui s'en charge. En fait, depuis janvier, le L. a une nouvelle structure. On est maintenant deux équipes : la première travaille plus sur les processus (i.e. les interactions entre l'eau, les minéraux et les contaminants dans le milieu vivant) ; la deuxième équipe travaille sur le transfert d'eau et donc la modélisation hydrologique. Donc le deuxième volet du projet est traité par l'équipe modélisation et le troisième volet par l'autre équipe du L. C'est dans ce workpackage 3 qu'interviennent l'INRA et le BRGM.

[>R1]: Vous, vous êtes dans cette équipe ?

[>R2]: Oui, moi je travaille et je porte cet axe-là. Alors, je ne vous en ai pas parlé, pour ne pas trop compliquer, mais dans cet axe 3, on travaille à la fois sur la ressource en eau et sur la ressource en sol.

[>R1]: Dans le sens « qualité du sol » ?

[>R2]: Voilà. Et même dans le sens « fertilité du sol ». Souvent on se dit « les forêts elles poussent et voilà ». Alors qu'en réalité les forêts en France sont exploitées. Ce sont des forêts qui sont plantées et coupées. Et ce depuis plusieurs siècles. Or, un sol forestier c'est comme un sol en plaine : il peut s'appauvrir dans certains contextes. Pourquoi ? Parce que les arbres ont besoin de nutriments, quand ils poussent (principalement de phosphore, d'azote, de carbone, mais aussi d'autres nutriments comme le calcium et le magnésium). Par exemple, ils

trouvent le calcium et le magnésium, dont ils ont besoin, dans les sols. Et ce calcium et ce magnésium, ils viennent de la roche. Parce qu'en fait un sol c'est une roche qui s'est désagrégée, puis il y a eu des plantes, puis de la matière organique, et puis ça a formé un sol. Si, par exemple, on regarde cette image-là [elle va chercher une feuille accrochée au panneau d'affichage], on voit bien ce qu'est un sol. C'est un granite qui commence à se fracturer, à se déstructurer, jusqu'à devenir de plus en plus petit. Et puis, au-dessus, on va trouver le sol. Voilà. Donc, si dans la roche vous n'avez pas beaucoup de calcium et de magnésium, vous n'en aurez pas beaucoup non plus dans le sol. Dans les roches calcaires, par exemple, il y a beaucoup de calcium et de magnésium ; donc il y en a aussi beaucoup dans le sol. Pour les forêts, il n'y a donc pas de problème. C'est le cas des forêts du Jura, qui sont en bonne santé. Dans les Vosges, où on trouve soit du granite soit du grès, les sols qui se forment ne sont pas très riches en calcium et en magnésium. Si, en plus, sur ce sol, vous plantez des arbres, si vous les coupez, les exportez, puis en replantez d'autres, les arbres vont pomper et épuiser le sol. Or un sol met entre 10 000 et 100 000 ans pour se former. En réalité, dans la nature, qu'est-ce qui se passe ? Vous avez une forêt. Au bout d'un moment, les vieux arbres tombent et ce qu'ils ont pompé repart dans le sol. C'est pour cette raison que les forêts primaires fonctionnent très bien. Parce qu'elles ne sont pas exploitées. Quand les vieux arbres tombent au sol, tout ce qu'ils ont pris va se dégrader ; les nouveaux arbres vont les prendre, ils vont mourir à leur tour, et puis voilà, ça va tourner. C'est ce qu'on appelle le recyclage biologique. Mais si vous plantez, coupez, exportez, replantez, etc. à une vitesse plus rapide que la formation du sol, forcément, au bout d'un moment, le sol s'épuise.

[>R1]: Ça veut dire qu'il y a des zones où il vaut mieux faire de la sylviculture que dans d'autres ?

[>R2]: Les forêts sont naturelles, mais après il faut voir quels arbres on plante. Si vous plantez des résineux (des épicéas par exemple), leurs épines vont tomber au sol et former une litière. Or les épines sont acides. Donc, quand elles se dégradent, elles sécrètent des acides. Et, en fait, où se trouvent les nutriments des sols ? Ils se trouvent sur la surface des minéraux, des argiles ou de la matière organique. Et là où c'est intéressant... Je ne sais pas, vous avez fait un peu de bio ?

[>R1]: Je me suis arrêtée au bac S.

[>R2]: Ah bon, ben, quand même alors ! Donc vous savez ce que c'est qu'un proton. En fait, voyez [elle dessine un schéma], ça, ce sont des cations (ils sont chargés). De l'autre côté, vous avez des protons. Les protons viennent, par exemple, de la dégradation de la litière, qui est acide, ou de ce qu'on appelait les pluies acides. En fait, quand on brûle des énergies fossiles, on va émettre des oxydes de soufre et d'azote. Quand ces oxydes de soufre et d'azote se retrouvent dans l'atmosphère, ils se transforment en acide sulfurique et en acide nitrique. Et l'acide sulfurique H2SO4, dès qu'il est dans l'eau, va donner SO2, plus des protons et des sulfates. Voilà, et c'est pareil avec les nitrates. Tout ça pour dire que, quand vous brûlez des énergies fossiles, vous vous retrouvez avec des protons, qui sont positifs. Donc ils vont faire partir les cations et être emportés dans l'eau. Certains vont être pris par la végétation ; les autres vont être drainés dans les [inaudible] et ensuite on va les retrouver dans les eaux de rivière. Il y a toujours eu des forêts dans les Vosges, ça n'est pas un problème. Le problème c'est qu'avec les pluies acides et la sylviculture on a appauvri les sols. Donc, nous, on travaille sur la question de la fertilité de ces sols. La question qu'il y a derrière, c'est justement la culture du bois. C'est-à-dire que, dans une commune de montagne comme le village d'Aubure, les principales ressources sont le bois et la chasse. Donc, si les forêts sont en mauvaise santé, c'est très embêtant pour la population. C'est pour ça qu'il faut adapter la sylviculture et c'est pour ça que, dans nos partenaires, il y a aussi l'ONF. On travaille avec eux. Donc voilà, on travaille sur la question de la fertilité des sols : est-ce que les sols sont encore fertiles et pendant combien de temps ? Là, par exemple, j'ai un doctorant qui fait des expérimentations sur les sols forestiers, pour essayer justement de déterminer où se trouvent ces nutriments (le calcium et le magnésium), s'ils sont accessibles, quels processus vont permettre de les récupérer et pendant combien de temps on va encore en avoir.

[>R1]: En quoi consiste le partenariat avec l'ONF ?

[>R2]: Par exemple, l'ONF va pouvoir nous fournir des données chiffrées sur les quantités de bois qui sont exportées par parcelle. Parce que qui dit sylviculture, dit dynamique. Donc il y a des arbres qui sont plantés puis coupés. Nous, ça nous permet d'avoir une idée de l'évolution de la forêt au cours du temps.

[>R1]: En échange, est-ce qu'ils vous demandent quelque chose ? Ou est-ce que leur partenariat est motivé par le seul intérêt de faire avancer la recherche ?

[>R2]: La collaboration est multiple et elle dure depuis un moment. Par exemple, là, il y a un projet d'amendement calco-magnésien. Ça se fait beaucoup en Allemagne. L'idée c'est que, s'il manque des nutriments à la forêt et qu'on veut continuer à avoir de la forêt, on va lui amener des nutriments (comme quand on amène de l'azote dans les champs de maïs). Donc on va récupérer du calcaire et de la dolomite (avec du calcium et du magnésium) dans les carrières de calcaire, on va les broyer finement, puis, par hélicoptère ou par avion, on va faire un spray et déposer le calcaire sur les sols, pour que ceux-ci retrouvent des nutriments et que les forêts aillent mieux. Tout ça pour dire que l'ONF et les communes environnantes autour d'Aubure avaient ce projet d'amendement calco-magnésien. Le problème c'est que ça coûte très cher, notamment à cause de l'application par hélicoptère. Pour l'instant, ils n'ont pas encore eu les financements. Ça c'est un projet sur lequel on a travaillé aussi. Donc ils ont récupéré nos données de sol, comme ça, ils n'ont pas eu besoin de les mesurer. Ça leur a permis d'économiser ça. Donc ce sont de petites choses comme ça : ils nous donnent des données ; nous, on leur donne des données. Voilà.

6 - 36:27 > 38:33 Ressources humaines

[>R1]: Dans votre équipe BISE, combien de personnes travaillent sur le projet ?

[>R2]: On est pas mal en tout dans le projet, mais de l'équipe on doit être trois permanents et un doctorant. Dans l'autre équipe, ils sont plus : ils sont quatre permanents et un doctorant. A l'IPGS, ils sont plus, parce qu'il y a beaucoup d'outils. Après, c'est compliqué : on peut avoir dix personnes, mais qui travaillent à 10% sur le projet. Donc ça va dépendre. Et quand je vous dis « on est trois permanents dans notre équipe », je ne compte pas les personnes qui font les analyses chimiques, etc.

[>R1]: Ce sont des techniciens, des ingénieurs…?

[>R2]: Oui. Donc il y a quand même plus de monde que ça.

[>R1]: Et quand vous dîtes « permanents », ce sont des chercheurs ?

[>R2]: Oui. Donc ça n'est pas tout à fait vrai. Dans notre équipe, on va dire qu'il y a trois chercheurs, quatre ou cinq techniciens ingénieurs et un doctorant. Par contre, ce que je n'ai pas dit et qu'il est important de souligner, c'est que, pour faire fonctionner un observatoire ou

un projet, il ne faut pas que des chercheurs. Il faut aussi des techniciens, des ingénieurs, des doctorants, des étudiants. Sinon, ça ne fonctionne pas. Il faut que tout le monde participe.

7 - 38:33 > 43:41 Partage des données

[>R1]: Je vais passer à la partie « Données ». Est-ce que vous-mêmes utilisez spontanément le terme de « donnée » ? Si oui, qu'est-ce que vous désignez sous ce terme ?

[>R2]: Oui, le mot « donnée », on l'utilise beaucoup. Et, effectivement, il y a plein de définitions à la donnée. D'autant plus que, nous, avec l'observatoire, on a les données librement accessibles. Ces données-là, c'est très simple : elles sont accessibles. On les met sur notre site internet et tout le monde peut les télécharger. Maintenant, c'est vrai qu'avec les normes Inspire, les Creative Commons, les DOI, il faut qu'on revoit notre politique. Parce qu'on se rend compte qu'il faut associer des données à des métadonnées, voire protéger les gens. Protéger la donnée et protéger le fournisseur de données. Parce que le problème des données, c'est que quelqu'un peut les utiliser et les vendre. Donc il faut associer la donnée avec une norme Creative Commons, en disant « c'est une donnée non lucrative, qui ne peut pas être utilisée pour être vendue ».

[>R1]: Mais ça, vous le mettez déjà sur votre site internet ?

[>R2]: Non, pour l'instant, on ne l'a pas encore fait. Parce que ce sont des choses qui sont en train de se mettre en place. Et puis c'est assez lourd. Et moi je n'ai pas fait cette formation. Donc il faut que je me forme. C'est compliqué – je suis déjà débordée.

[>R1]: Parce que ce serait vous qui prendriez tout ça en charge ?

[>R2]: Ben oui. Parce que, malheureusement, dans la recherche, ce qu'on manque vraiment le plus, c'est de personnel. On est de moins en moins et on doit tout faire. C'est moi qui ai fait le site internet, c'est moi qui m'occupe des données, de la validation des données, qui vais faire visiter le site aux gens quand ils veulent le visiter… Voilà. Donc ça c'est quand même un gros problème. C'est, à mon avis, le problème n°1 de la recherche. Le problème ce ne sont pas forcément les dotations. Parce que, par exemple nous en sciences, on trouve des projets. Ce qu'il faut ce sont des personnes permanentes. Et ça, c'est l'État qui doit nous les attribuer.

[>R1]: Oui, c'est bien de mettre à disposition les données, mais ça implique des métadonnées, etc. et donc des personnes pour s'en occuper.

[>R2]: Tout à fait. Donc, sous le terme de données, il y a la donnée librement consultable. Ensuite, il y a la donnée scientifique. Nous, par exemple, dans le cadre de l'ANR ou d'autres projets, – bon, ça n'est pas tout à fait en place – on partage des ordinateurs et ensuite on dépose nos fichiers. Comme ça, chacun peut les récupérer, etc.

[>R1]: Vous le faites aussi dans le cadre du projet H. ?

[>R2]: Oui, oui. Alors, ça n’est pas encore tout à fait au point, parce que, comme on est dans deux bâtiments différents, ça n’est pas si simple. Mais, en tout cas, ici en interne, on a un PC qui est partagé.

[>R1]: C’est un PC qui est accessible à distance ?

[>R2]: Oui, il est accessible à distance. Enfin, en intranet ; pas depuis chez moi. Ici, dans mon bureau, je peux récupérer des données. Après, on a des données qui sont communes. Là, par exemple, je vais vous montrer, on a quelqu'un qui fait du géo-référencement. On a ce qu'on appelle un MNT : un Modèle Numérique de Terrain. Ce sont des cartes. Et donc, ça, ce sont des choses qui sont communes à toutes les personnes qui travaillent sur l'observatoire. Voilà. Alors, évidemment, sur le papier, ça vous semble juste un papier. Mais, en fait, tout est géoréférencé avec des coordonnées GPS. C'est dans un logiciel. Donc, par exemple, quand quelqu’un a fait une mesure à tel endroit, il a les coordonnées GPS, donc il va aller dans le modèle et ça va lui positionner exactement les points. Et ensuite ça pourra être échangé. Si quelqu'un a fait de l'électrique (c'est un outil géophysique) à tel endroit et si, moi, ensuite, je veux faire de la sismique ou de la gravimétrie, je me dis : « ah ben, c'est quand même bête, je ne vais pas aller ailleurs, puisqu'on a déjà des données là ». Donc je vais récupérer les données, et puis, sur le terrain, avec le GPS, je vais me repositionner au même endroit. Voilà, ça peut servir à ça par exemple.

[>R1]: Donc c’est un logiciel commun ?

[>R2]: Alors, le logiciel, c'est un logiciel qu’on trouve partout. Ce qui est commun, c'est le un modèle numérique de terrain (MNT). Après, ça peut être des choses toutes bêtes comme

mettre des photos en commun. Comme ça, quand quelqu'un a besoin de décrire ou de faire une présentation du site, il a une bibliothèque de photos disponible.

8 - 43:41 > 50:46 Types de données

[>R1]: Comment différenciez-vous la donnée librement consultable via le site de l'observatoire de la donnée scientifique ?

[>R2]: C'est extrêmement simple, parce que la liste des données librement consultables c'est la même. C'est par exemple le ruisseau, la composition en calcium du ruisseau tous les 15 jours... Ça c'est fourni, c'est dans la base de données.

[>R1]: C'est une donnée acquise par un instrument ?

[>R2]: Non, non, non. C'est un long processus : on va sur le terrain, on échantillonne, on ramène le flacon, on le filtre et ensuite on l'analyse. Ah non, non, ça n'est pas du tout simple. Enfin, ça n'est pas instantané. Mais, par contre, ça fait partie des données qu'on fournit. Donc c'est très simple : les données qui sont librement consultables, on en a la liste. Donc celles-ci, ça ne change pas. Après, les données un peu complexes, ça ne peut pas être des données librement consultables. Déjà parce qu'elles sont très complexes. Par exemple, faire un profil sismique, c'est beaucoup de travail : une fois qu'on a le signal, il faut l'inverser, travailler dessus… En plus, le problème de la donnée, c'est que… Si on mesure par exemple l'acidité. Je mesure le pH (c'est une valeur brute, simple). Je la mets sur le site internet. Quelqu'un peut l'utiliser. J'ai expliqué que « voilà, ça a été mesuré avec un pH-mètre, la valeur vaut 6,52, l'incertitude c'est plus ou moins 2 ». Donc, ça, c'est fourni, c'est clair, c'est protocolisé, c'est simple. Maintenant, si quelqu'un fait un profil sismique par exemple : il envoie des ondes, il enregistre des vitesses de propagation d'ondes… Bref, une fois qu'il a ça, il n'a rien. Donc il va interpréter la vitesse de diffusion dans le milieu, en disant : « Ici la vitesse s'atténue. C'est parce que l'onde rencontre autre chose. Elle est passée du sol à la roche ». Donc là j'interprète, je déduis, je fais des calculs, j'ai des logiciels, et j'en déduis par exemple qu'ici j'ai 1,50m de sol puis je suis dans la roche. Mais, ça, ce n'est pas la donnée brute. C'est la donnée interprétée. C'est très délicat de mettre à disposition des données interprétées. Parce qu'à valider ça n'est pas simple. Donc, là aussi, dans la donnée scientifique, il y a la donnée protocolisée. Celle-là, je peux la rendre accessible, puisque moi je ne l'ai pas interprétée, je ne

l'ai pas transformée. C'est une mesure et je donne une incertitude. Donc je suis sûre de ma donnée. Maintenant, la donnée interprétée, si ça se trouve, dans dix ans on va avoir d'autres logiciels. A partir de la donnée de terrain, je vais donner une autre interprétation. Ça aussi c'est compliqué, parce que, éthiquement, rendre disponible une donnée interprétée, ça n'est pas simple. Alors que, quand on la publie, on passe par un comité de review. Les reviewers disent « oui, je suis d'accord » ou « non, je ne suis pas d'accord ». Notre donnée a été publiée. Donc on peut la réutiliser.

[>R1]: Ça donne une garantie.

[>R2]: Oui. Donc, là aussi, il faut faire attention à ce qu'on rend public. Il faut que ce soit des données qui soient éthiques. C'est très important.

[>R1]: C'est intéressant. C'est un point de vue que je n'avais pas encore entendu.

[>R2]: Par exemple, on mesure la température extérieure. Bon, voilà : hier il a fait 12°C – plus ou moins, parce que, sur ma mesure, j'ai 1°C ou 0,1°C d'incertitude. Mais je ne me suis pas trompée. En plus, mes capteurs, je les vérifie. Quand il y a une dérive, je la vois. Donc c'est une donnée qui est sûre. Maintenant, si j'écris : « on a regardé les scénarios climatiques, sur lesquels il y a une incertitude, et puis je peux vous dire qu'en 2100... ». Ça, je ne peux pas le rendre comme ça. Donc je vais le publier, l'expliquer, en disant « attention, j'ai utilisé un scénario, c'est une possibilité… ». Et là, il faut lire tout l'article. Alors on comprend. C'est comme quand des gens citent une citation hors contexte. Admettons qu'on ait fait un article, dans lequel il y a une figure qui dit « voilà, avec ce scénario-là, en 2100 il n'y aura plus d'épicéas sur le site mais des hêtres ». Nous, dans l'article, on va expliquer que c'est un scénario possible, etc. Maintenant, si quelqu'un sort cette citation de son contexte, voyez… Donc c'est pour ça qu'on ne peut pas mettre sur notre site internet les données disant « voilà, en 2100, il n'y aura plus d'épicéas ». On ne peut pas faire ça. En plus, nous, dans notre communauté, on va différencier la donnée brute de la donnée traitée, de la donnée validée...

[>R1]: Qu'est-ce que vous appelez « donnée validée » ?

[>R2]: Par exemple, quand je vous dis « le calcium, c'est simple, je regarde et après je le publie », il y a quand même une phase de validation. Parce que le calcium a été mesuré au labo, mais parfois il peut y avoir un problème. Ou alors il a été échantillonné sur le terrain, mais, je ne sais pas, peut-être que la personne qui l'a échantillonné a confondu deux choses. Je

veux dire, ça arrive. Donc, nous, avant de prendre la donnée et de la mettre en ligne, on va faire un certain nombre de procédures de validation, pour être sûr que c'est le bon échantillon, que l'appareil qui l'a mesuré n'avait pas de problème, etc.

[>R1]: Ces procédures sont systématiques ?

[>R2]: Systématiques. Je pense que c'est important de le faire. C'est aussi pour ça qu'on ne peut pas mettre en ligne les données en temps réel. Moi je fais la validation une fois par an – en décembre ou en janvier. Je valide les données de l'année précédente. Parce que ça me prend un mois.

[>R1]: Oui, c'est pour ça, j'ai fait un test : j'ai cherché les données de mars 2018 et je n'ai obtenu aucun résultat. C'est parce qu'elles n'ont pas encore été mises en ligne ?

[>R2]: C'est même plus compliqué que ça. Techniquement, comme je l'ai dit, on a des problèmes de personnel. Et la personne qui s'occupait de la base de données, elle... voilà. Donc on n'a plus personne qui s'occupe de la base de données. Donc, de toute façon, elle est arrêtée. Bon, là, on est en train de mettre en place des solutions, donc ça va revenir. Voilà, il y a aussi ce problème. Mais ce sont des problèmes humains, c'est autre chose.

9 - 50:46 > 59:31 Mode d'acquisition des données

[>R1]: Quelles données acquérez-vous ? Comment les traitez-vous ? Est-ce que vous documentez les procédures ? On peut prendre l'exemple des données qui sont acquises dans le cadre du troisième axe.

[>R2]: Le troisième axe est un peu différent, parce qu'il y a de l'expérimentation. L'expérimentation c'est encore autre chose.

[>R1]: Ce sont des expérimentations en laboratoire ?

[>R2]: Voilà, on fait des expériences en laboratoire. On prend des sols, on fait des petits tests, etc. Donc ça, évidemment, ça ne peut pas être de la donnée libre. Parce que ce ne sont pas des données de terrain. En plus, c'est dans le cadre d'une thèse. On retombe vraiment dans les schémas classiques, c'est-à-dire on fait de l'expérimentation, on réfléchit dessus, on essaie d'interpréter et on publie.

[>R1]: Est-ce que vous acquérez aussi des données de terrain ?

[>R2]: Oui. Les données de terrain c'est typiquement la carte des sols. Une fois qu'elle sera validée, etc., je la mettrai sur le site internet et tout le monde pourra la récupérer.

[>R1]: Vous parlez de la carte qui est réalisée par le bureau d'études ?

[>R2]: Oui, voilà.

[>R1]: Est-ce qu'il y a aussi des relevés ?

[>R2]: Ça on le fait avec l'INRA. Il y a des relevés de biomasse, etc. C'est un peu compliqué. Ces relevés nous aident à comprendre quel est l'impact de la végétation sur les transferts de nutriments. Par exemple, on va quantifier les chutes de litière. On a mis des grands bacs, et puis tous les Δt on va mesurer la quantité de litière qui est tombée au sol, pour faire des bilans, etc. C'est ce qu'on appelle des quantifications de biomasse. Ou bien on va prendre une certaine surface et compter le nombre d'arbres, la circonférence des troncs, la hauteur, etc. sur cette surface. Là aussi ce sont des quantifications de biomasse. Ça c'est pareil : une fois que les données seront validées, etc., on pourra les mettre à disposition. Parce que ce sont des données de terrain.

[>R1]: Et le BRGM, quel est son rôle ?

[>R2]: Le BRGM, c'est tout à fait autre chose. Avec eux, on va travailler sur la caractérisation des phases minérales. Un sol est constitué de matières organiques, d'humus, mais aussi d'argile et de morceaux de minéraux. Donc, avec le BRGM, on caractérise toutes ces phases qui sont dans le sol : quels minéraux il y a (du quartz, des feldspaths, etc.), quelle taille ont les grains de quartz par exemple, quelle est leur granulométrie, de combien est leur surface réactive, quelle est leur capacité d'échange (i.e. leur capacité à avoir des cations sur leur surface)… Ces quantifications, on les fait à la fois avec l'INRA, avec certaines méthodologies, et avec le BRGM, avec d'autres méthodologies. Le BRGM va plus travailler sur la caractérisation des phases minérales. L'INRA, lui, va plus travailler sur la pédologie pure et la biomasse (tout ce qui est écologie forestière, en fait).

[>R1]: Et vous, ici, vous travaillez plus sur les expérimentations ?

[>R2]: Oui. Et, dans le cadre de la thèse du doctorant, on va aussi utiliser des traceurs qui sont les isotopes. Nous, notre spécificité dans notre équipe c'est d'utiliser certains traceurs comme

les rapports isotopiques, pour essayer de mieux comprendre les processus. Vous voulez que je vous explique rapidement ce que c'est qu'un isotope ? On va prendre un exemple (ça n'est pas forcément celui qu'on regarde, mais c'est vraiment le plus simple à comprendre) : l'oxygène. Vous avez de l'oxygène 16 et de l'oxygène 18. Ça correspond au nombre de protons et d'électrons. Donc il y a deux isotopes. L'un est plus lourd que l'autre. Dans la plupart des molécules d'H2O, c'est de l'oxygène 16. Mais certaines ont de l'oxygène 18. On va prendre l'exemple le plus simple : vous avez une casserole avec de l'eau. Dans votre eau, vous avez beaucoup d'oxygène 16 et un peu d'oxygène 18. Vous avez donc un rapport entre les deux, qu'on appelle Δ16O/18O et qui est égal à x. Vous mettez l'eau à chauffer. Quand elle atteint 100°C, une partie va partir en phase vapeur. La molécule 18O est un peu plus lourde, donc elle va avoir tendance à rester plus longtemps. Par contre, les molécules 16O vont partir plus vite. Au bout d'un certain temps, le x final va donc être enrichi en oxygène 18 : x final va être inférieur à x initial. Dans la vapeur, en revanche, vous allez avoir un rapport inverse. L'eau restant dans la casserole va avoir tendance à s'alourdir, tandis que l'eau en phase vapeur va avoir tendance à s'alléger. Les eaux peuvent donc avoir des rapports différents. Parce qu'on pourrait dire : « dans la nature, c'est partout pareil ». En fait, dans la nature, ça n'est pas partout pareil. Dans la nature, Δ16O/18O varie. Donc, suivant les masses d'eau, en fonction d'où elles viennent (au-dessus des océans, etc.), vous allez avoir des Δ16O/18O différents. Et puis, lorsque ces masses d'eau arrivent sur le site, il va y avoir des processus d'évaporation, d'évapotranspiration, etc. Donc ça va aussi changer. En suivant les rapports isotopiques de la pluie, de l'eau dans tous les petits ruisseaux, de l'eau dans la sève, de l'eau dans les solutions du sol, on va pouvoir reconnaître les processus. Alors, nous, on ne regarde pas forcément l'eau, pour différentes raisons. On va plutôt regarder le strontium, le calcium, le néodyme, le plomb, l'uranium, le thorium, le lithium, le bore. On regarde les isotopes de chacun de ces éléments, parce qu'ils n'ont pas tous la même histoire. Par exemple pour le plomb, là où ça a été très intéressant, c'est qu'autrefois il y avait du plomb dans les essences. Et ce plomb avait une certaine composition isotopique. Là où c'est incroyable, c'est qu'on peut tracer, grâce au plomb retrouvé dans les lacs ou les arbres, l'arrivée de l'essence avec du plomb (ce qu'on appelait le plomb anthropique). On en retrouve en Antarctique, on en retrouve partout. Et, comme ça, on peut même tracer les masses d'air. Or, typiquement, quand on fait une mesure

isotopique, on ne va pas la mettre sur internet le lendemain. Parce qu'obtenir un rapport isotopique, ça représente des mois de travail.

10 - 59:31 > 1:02:35 Sauvegarde des données

[>R1]: Ces données, vous les rentrez dans des fichiers Excel ?

[>R2]: Oui. Alors, vous soulevez un autre gros, gros enjeu : celui de la conservation des données. Parce que chaque doctorant va travailler sur un projet. Mais, quand le doctorant part, il ne faut pas que les données partent avec lui. Donc il faut mettre en place une politique de conservation et d'archivage des données.

[>R1]: Il y en a une dans le laboratoire ? Dans certaines unités de recherche, les doctorants déclarent par écrit que les données resteront la propriété du laboratoire.

[>R2]: Oui, ici aussi, ils signent une charte. Mais, en même temps, il faut l'anticiper. Parce que, d'accord, le doctorant vous remet une clé USB. Mais il faut que, vous, vous y compreniez quelque chose dans ses fichiers. Donc il faut protocoliser un peu les choses. Après, c'est à chaque encadrant de mettre en place des choses. Mais, nous, dans le cadre du projet ANR, c'est vrai qu'on a quand même une politique. On y travaille là.

[>R1]: C'est un travail en cours ?

[>R2]: Oui, c'est en cours.

[>R1]: Et comment vous imaginez cette politique ?

[>R2]: Pour le moment, ce qu'on a imaginé, – ça n'est pas très inventif mais...- c'est une architecture de fichiers sur un ordinateur dédié du réseau.

[>R1]: Une arborescence ?

[>R2]: Une arborescence, voilà. C'est vraiment comme ça qu'on le voit pour le moment.

[>R1]: Est-ce qu'il y aura des règles de nommage de fichiers ou des métadonnées à rentrer ?

[>R2]: Ça, il faut qu'on y vienne aussi. Moi, en général, je demande aux doctorants de mettre la date en fin de fichier. Parce que là c'est pareil : le doctorant va faire des modifications, donc il va mettre « version 1 », « version 2 », « version 2 bis »… Et ça, ça ne va pas. Donc il faut

essayer d'avoir effectivement une politique de nomination des fichiers. Alors, nous, on a déjà fait ce travail sur les échantillons. C'est-à-dire que, nous, puisqu'on prélève des eaux, des roches, des litières, on a mis en place un système d'archivage. Alors, ça n'est pas pour les données mais c'est pour les échantillons. On a réussi à récupérer un lieu, où on stocke tout. Parce qu'avant, dans le bureau d'untel, il y avait les échantillons de sa thèse, dans le bureau d'untel... Ça n'était pas possible.

11 - 1:02:35 > 1:06:55 Valeur économique des données

[>R1]: Quelle valeur – scientifique, économique… – accordez-vous aux données ?

[>R2]: Alors là… Économique, je n'ai pas envie de me poser la question. Je ne sais pas quelle est la valeur économique ou financière de la donnée. Par contre, je sais ce qu'elle a coûté. Je ne sais pas ce qu'elle vaut, mais je sais ce qu'elle a coûté. Ça n’est pas pareil.

[>R1]: Et en considérant combien elle a coûté, est-ce que vous pourriez dire si elle a une grande valeur ?

[>R2]: Ben, ça va dépendre des données… Ça n'est pas simple, cette question. Je ne sais pas, je n'y ai jamais vraiment pensé, donc j'improvise. Le projet, il a un coût. Les gens qui vont sur le terrain pour récupérer des échantillons, ça coûte. L'échantillon, quand on le ramène au laboratoire et qu'on le traite, ça coûte. Le doctorant qui va travailler dessus, ça coûte. Il y a aussi mon salaire. Voilà. Alors, on ne fait pas vraiment ce travail de coût réel. On l'a fait pour l'observatoire. On a calculé son coût réel et son coût consolidé, parce qu’on est en train de monter une infrastructure de recherche de la zone critique, dans le but de monter ensuite un ESFRI européen. C'est compliqué, parce qu'on inclut les salaires mais aussi tous les équipements qui sont sur le site depuis 1986 (sachant que certains équipements s'amortissent). Donc, voilà, déjà ça, c'était un travail intéressant mais lourd et pas évident. Est-ce que les choses qu'on a mises en place ont pris de la valeur, parce que justement elles ont permis d'acquérir beaucoup de données ? Ou bien est-ce qu'au contraire elles en ont perdu, parce que, chaque année, depuis qu'on les a installées, leur coût diminue ? Je n'en sais rien moi, voyez.

[>R1]: Est-ce que ça justifierait que les données soient payantes ou qu'il y ait une forme de gratification (pas forcément financière, mais ne serait-ce que par la citation) ?

[>R2]: Ça n'est pas pareil : la citation, quelque part, permet de reconnaître le travail. Maintenant, je pense que les universités ou les centres de recherche n'ont pas vocation à fournir de la donnée payante. Nous, on est payé par l'argent public. Et puis, attention, délivrer une donnée payante, ça veut dire aussi que... Si je reprends l'exemple de mon calcium. Mon calcium, je le mesure ici. On fait évidemment toutes nos vérifications. Mais on n'est pas norme ISO 9002. Or quelqu'un qui paie va dire « attendez, moi je veux bien vous acheter votre donnée, mais je veux un gage de qualité, je veux que ce soit agréé Y ». Mais, nous, si on rentre là-dedans, on est foutu. Parce que, pour le coup, il faudrait doubler les effectifs (déjà qu'on a peu de personnel). Donc je ne vois pas trop ni l'intérêt que L'État va y avoir, ni celui que nous allons y avoir. Après, on va dire « ah ben, maintenant que vous êtes agréés ISO 9002, vous pouvez faire laboratoire ouvert ». Et ensuite n'importe qui va dire : « Attendez, j'aimerais bien avoir une mesure. Je vais la faire chez vous, et puis je vais vous demander de l'argent ». Dans ce cas, on n'est plus un organisme de recherche, mais un organisme de prestation de services. Or, une prestation de service, c'est pareil, il faut des gens dédiés, qui ne font plus de la recherche mais de la prestation de services. C'est un autre métier. Donc je ne suis pas sûre que ce soit tout à fait judicieux.

12 - 1:06:55 > 1:09:14 Réutilisation des données

[>R1]: Qui réutilise les données ? Avec l'observatoire, avez-vous un peu une idée de qui réutilise les données et dans quel but ?

[>R1]: C'est une bonne question. Malheureusement, comme on est un peu en panne d'ingénieur pour notre base de données... Je sais que maintenant il existe des outils pour savoir qui a consulté les données, qu'est-ce qu'ils en ont fait, etc. Malheureusement, nous, on ne l'a pas. Donc je ne peux pas vraiment vous dire. De temps en temps, je suis sollicitée par des gens, soit parce qu'ils n'ont pas vraiment compris qu'ils pouvaient récupérer directement les données, soit parce qu'ils veulent quand même passer par moi.

[>R1]: Ce sont des chercheurs ?

[>R2]: Non, pas forcément. Les chercheurs savent comment faire. Ils savent récupérer les données sans passer par les gens. Ce sont plutôt des personnes qui veulent avoir des données d'hydrologie ou des données de composition. Ça peut être des particuliers, ça peut être des

étudiants pour des projets de recherche, ça peut être des associations, ça peut être des maisons de parc, ça peut être aussi la commune… Voilà.

Entretien avec le chercheur 43 (géographie)

1 - 00:37 > 06:33 Objectifs scientifiques du projet de recherche

[>R1]: Est-ce que vous pourriez me résumer un peu le projet T. ?

[>R2]: T. est un projet de recherche mené en collaboration entre notre laboratoire de géographie et des laboratoires en informatique. Il a pour objectif de proposer des méthodes de traitement de données à haute fréquence temporelle. Ce sont des données particulières, puisqu'on s'intéresse à des observations issues de capteurs. Ça peut être des capteurs spatiaux, aériens ou terrestres. Les capteurs spatiaux nous donnent des images satellites ; les capteurs aériens des photographies aériennes ou des images issues de drones ; les capteurs terrestres sont des instruments d'acquisition d'images 3D. Le projet est donc consacré à l'observation multi-capteurs à haute fréquence temporelle sur des sites liés à l'occupation du sol ou à des problématiques environnementales. On a par exemple des sites d'étude, qui sont dans des contextes montagnards, où on va plutôt s'intéresser aux données acquises à haute fréquence sur un glissement de terrain, afin de suivre l'évolution de celui-ci. En fait, on n'a pas une seule et unique thématique. On essaie de définir plusieurs thématiques, pour lesquelles les données à haute fréquence temporelle peuvent avoir un intérêt, i.e. peuvent apporter une connaissance nouvelle à nos objets d'étude.

[>R1]: Qu'est-ce que vous appelez « occupation du sol » ?

[>R2]: L'occupation du sol c'est le fait d'avoir une cartographie précise des modes d'occupation du sol. C'est dire par exemple : « ici ce sont des prairies », « là c'est du maïs », « là c'est de l'urbain », dans l'urbain ça peut être du tissu individuel (un lotissement individuel), des zones d'activités, etc. C'est ça l'occupation et l'utilisation du sol : c'est la manière dont l'espace est occupé par différentes fonctions urbaines.

[>R1]: Les trois types de capteurs sont-ils complémentaires ?

[>R2]: Oui. Ils nous permettent d'avoir à la fois la vision du haut et la vision terrestre (la vision terrestre permet d'avoir des informations encore plus précises sur un objet, notamment sur ses dimensions 3D).

[>R1]: Les thématiques se rapportent-elles toutes au domaine de la géographie ?

[>R2]: On va dire que oui. On peut généraliser en disant que ce sont des thématiques environnementales, pour lesquelles il y a des enjeux et des besoins. On travaille en effet beaucoup en interaction avec les collectivités locales. On identifie les phénomènes ou objets d'étude, sur lesquels il y a le besoin d'avoir une connaissance supplémentaire. Pour donner un exemple très concret : la région Grand Est travaille sur une cartographie à grande échelle des modes d'occupation du sol. Or, parmi les thèmes d'occupation du sol qui la préoccupent, avoir une cartographie des prairies dites « permanentes » est quelque chose qui n'est pas du tout facile à obtenir. Ça se fait beaucoup par photo-interprétation manuelle mais ça n'est pas très fiable. La région aimerait bien savoir ce que peut apporter l'imagerie aérienne et spatiale pour cartographier ce thème particulier d'occupation du sol. Nous, dans le cadre du projet T., on essaie d'extraire des informations des images à très haute fréquence temporelle. Qu'est-ce qu'on entend par « très haute fréquence temporelle » ? Il y a quelques années, si on arrivait à avoir deux ou trois images satellitaires par an, on était content. Une image satellite c'est une zone de 60 x 60 km avec des conditions atmosphériques correctes, pour pouvoir voir quelque chose (dès qu'il y a des nuages, on ne voit plus rien). Aujourd'hui, avec les nouveaux capteurs (i.e. depuis 2015-2016), on peut avoir des images optiques et radars (deux signaux complémentaires) tous les 5 ou 10 jours. Ce qui change énormément le volume de données à traiter et les méthodes qu'on va devoir appliquer pour traiter ce volume. Pour le moment, on sait traiter une image. Mais avoir un flux d'images plus fréquent, ça apporte une autre information, qu'on n'avait pas jusqu'à maintenant. C'est ça l'objectif du projet T. C'est assez méthodologique, mais c'est appliqué à des chantiers opérationnels ou issus des besoins des collectivités notamment.

2 - 06:33 > 10:59 Répartition des rôles entre les partenaires du projet

[>R1]: Comment sont répartis les rôles entre les différents partenaires du projet ?

[>R2]: Nous, on est un laboratoire de géographie, mais on ne s'intéresse pas qu'à la thématique. On a aussi une petite équipe, qui travaille sur le développement de nouvelles méthodes de traitement d'images. On ne code pas, on ne développe pas directement, parce qu'on n'est pas des codeurs informatiques, mais on met en place des chaînes de traitement, qui

peuvent être innovantes, à partir de bouts de code existants. Ça se fait en collaboration avec des informaticiens. Donc on n'est pas juste là pour la thématique. On essaie vraiment d'intégrer les laboratoires et leurs compétences. Les partenaires en informatique ne connaissent pas forcément l'imagerie satellitaire. Pour eux, une image satellite c'est une image comme une autre. Ils ont l'habitude de traiter de l'imagerie médicale ou de faire de la fouille de données issues du bancaire par exemple. L'image de télédétection est un nouveau type de donnée pour eux. Donc on essaie de voir quelles méthodes ils appliquent dans l'imagerie médicale par exemple. Et inversement.

[>R1]: J'ai vu que, parmi les partenaires, il y avait aussi une université australienne et l'Institut de Physique du Globe de Strasbourg.

[>R2]: En fait, quand on dit « le partenaire de Strasbourg », ça regroupe plusieurs laboratoires : le L. et le I. Parce que les applications ne sont pas forcément des applications dont nous, le laboratoire L., sommes spécialistes. L'application glissements de terrain, par exemple. On travaille en interaction avec le I. là-dessus, car un glissement de terrain est une surface naturelle évolutive, qui comprend aussi des enjeux humains.

[>R1]: Et l'université australienne – la M. University ?

[>R2]: Dans le cadre des projets ANR, on ne peut pas faire apparaître un laboratoire étranger. Donc le laboratoire australien est partenaire du projet, mais non financé. Il a ses propres financements. C’est quelqu'un qui est spécialisé dans la fouille de données, le deep learning, etc. C'est pour ça qu'il participe au projet.

[>R1]: C'est un chercheur en particulier ?

[>R2]: C'est une petite équipe, dont un chercheur en particulier. Enfin, non, ça n'est pas une petite équipe, mais ils sont au moins deux ou trois à participer au projet. [...]

[>R1]: Comment faites-vous pour travailler ensemble ? Comment sont réparties les tâches entre les partenaires ?

[>R2]: On fait des réunions très régulièrement, au terme desquelles on se fixe des objectifs. Le travail en commun passe beaucoup par l'encadrement d'étudiants entre deux laboratoires. Actuellement, on est dans la phase où des sujets de thèse en cotutelle vont commencer à être lancés.

[>R1]: Il y aura plusieurs doctorants ?

[>R2]: Il y aura au moins deux doctorants, encadrés en multi-laboratoire. C'est de cette manière qu'on arrive à échanger.

3 - 10:59 > 13:30 Mode d'acquisition des données

[>R1]: Est-ce que vous travaillez avec un ensemble d'images test ?

[>R2]: Une fois qu'on a défini les enjeux et les applications, oui, on travaille avec des données test. L'avantage c'est que les images à haute fréquence temporelle qu'on va utiliser sont aujourd'hui devenues complètement accessibles de manière gratuite.

[>R1]: Ça n'est pas vous qui acquérez ces images ?

[>R2]: Non – en tout cas, pas celles-là. On va les chercher sur des plateformes de distribution au niveau européen ou national. Au niveau national par exemple, on va chercher des images via le pôle Theia. C'est un pôle de distribution, qui structure l'animation dans le domaine de la télédétection en France et qui centralise l'accès aux données. L'accès à l'imagerie satellitaire se fait beaucoup par cet organisme. Ils ont mis en place une plateforme de distribution, sur laquelle les chercheurs peuvent faire des demandes et acquérir des données. Ça représente une partie des données qu'on utilise. Après, il y a d'autres catégories de données, qu'on ne va plus chercher sur des plateformes à l'extérieur, mais qu'on va acquérir nous-mêmes. Quand ce sont des données terrestres, par exemple, on utilise un instrument qui s'appelle un Lidar. Les données générées appartiennent au laboratoire qui dispose du matériel. Pour le Lidar, on a construit une plateforme commune au sein de l'Université : elle s'appelle la plateforme d'observation de la Terre ; elle est hébergée sur le site de l'EOST. En fait, il y a à la fois du matériel issu de la géographie et du matériel issu de l'EOST. On essaie de mutualiser les achats de gros instruments comme ça, parce que c'est du matériel pour l'acquisition de données qui a un certain coût.

4 - 13:30 > 19:12 Mode de diffusion des données

[>R2]: La réflexion sur les données, nous, en géographie, on est dedans – que ce soit par des enquêtes, par la collecte avec différents instruments... On baigne dedans, on est vraiment

concernés. Au cours d'un projet, on essaie toujours de penser à la création de fiches de métadonnées, afin d'avoir une trace et de pouvoir ensuite mettre à disposition des données-résultats. Pas forcément la donnée acquise par le capteur (i.e. la donnée brute ou même prétraitée) – ça ne va pas servir à grand monde, puisqu'elle est déjà disponible sur certaines plateformes. Ce sera plutôt la donnée à valeur ajoutée, i.e. un résultat ou une carte (par exemple, la cartographie des prairies, la cartographie de l'occupation du sol ou celle des tissus urbains). On travaille à la mise en place d'un portail de cartographie en ligne, qui permettrait de valoriser ces données.

[>R1]: Les cartes seraient en libre accès ?

[>R2]: Oui. On peut visualiser la donnée, en superposition sur un fond. Vous avez par exemple le portail GéoGrandEst. C'est la plateforme régionale de mutualisation de la donnée géographique. Ça a d'abord été une plateforme de la région Alsace, sous le nom de Cigal, avant de s'ouvrir à la région Grand Est. Chaque partenaire, dont notre laboratoire, peut mettre les données ou les produits à valeur ajoutée en ligne. Généralement, on met assez peu de données brutes.

[>R1]: Y a-t-il une politique consistant à déposer systématiquement les métadonnées de tous les jeux de données en ligne ?

[>R2]: Oui. Ça c'est plus une réflexion labo. Le laboratoire est dans cette démarche d'essayer de récupérer les métadonnées auprès des chercheurs. Moi j'en ai déjà fourni plusieurs (une cartographie de la végétation, par exemple). Mais c'est compliqué. Je pourrai vous donner le contact de notre ingénieur d'étude, qui s'occupe des données. C'est lui qui pourra vous parler plus précisément de ces problématiques. Parce que c'est vrai que c'est compliqué d'obtenir ça des gens. Ça reste encore une contrainte. Ça va venir, on sera obligé.

[>R1]: Il s'agit d'une politique en lien avec la directive Inspire ?

[>R2]: Oui, c'est ça. Et moi je suis complètement pour. Il faut. On fait des recherches qui sont financées par le secteur public, donc il est tout à fait logique que les données ou les traitements/résultats soient accessibles. Savoir que tel labo possède telles données, rien que ça, c'est déjà... Ça évite de refaire des relevés multiples, voire ça permet de travailler ensemble sur certaines choses.

[>R1]: Y a-t-il d'autres enjeux qui rentrent en ligne de compte, lorsque vous souhaitez mettre une carte à disposition, par exemple ? Je pense notamment aux problématiques de publication.

[>R2]: Oui, c'est ça un peu la difficulté. Généralement, le chercheur va plutôt avoir tendance à publier d'abord (c'est ce qu'on a tous tendance à faire). De toute façon, de plus en plus d'éditeurs demandent de fournir les lots de données avec la publication.

[>R1]: Dans votre domaine aussi ?

[>R2]: Oui. Ça se fait de plus en plus.

[>R1]: Est-ce que vous auriez des exemples d'éditeurs qui le demandent ?

[>R2]: Oui. Ce sont plutôt des revues en informatique. Ça se fait un peu plus dans ce domaine et un peu moins, on va dire, dans la communauté géo ou dans les publications en géographie. Comme moi je travaille beaucoup avec des laboratoires d'informatique, il m'est arrivé de fournir les données avec la publication. Vous avez par exemple la revue Remote Sensing, qui est en accès gratuit (donc qui oblige les chercheurs à payer la publication). Je crois que, dans cette revue, on peut aussi déposer les données. On « peut », on ne doit pas forcément.

[>R1]: Oui, c'est seulement une recommandation.

[>R2]: Voilà, oui. Mais il y a certaines revues, où ils le demandent vraiment. Je n'ai pas de nom en tête, il faudrait que je regarde.

5 - 19:12 > 20:52 Définition d'une donnée scientifique

[>R1]: Est-ce que vous utilisez spontanément le terme « donnée » ?

[>R2]: Oui. Parce que c'est mon cœur de métier. En géographie, la donnée c'est le fondement. Parce que c'est le cœur du système d'information géographique. C'est à partir de là qu'on va pouvoir suivre des phénomènes, faire des analyses et apporter des connaissances nouvelles sur ces phénomènes. Donc la donnée c'est le cœur du métier du géographe – la donnée géographique, dès qu'elle est localisée.

[>R1]: En géographie, la donnée serait donc une information géo-localisée ?

[>R2]: Alors, attention au vocabulaire : « donnée », « information », « connaissance ». Quelque chose qui est issu d'une collecte, c'est brut. C'est une donnée. Dès qu'on la transforme, ça devient une information.

[>R1]: D'accord. A partir du moment où il y a une analyse,...

[>R2]: C'est ça. A partir du moment où il y a une valeur ajoutée, ça devient une information. Donc, pour moi, une donnée reste quelque chose de brut. Une image satellite, par exemple, c'est brut. Je vais la chercher sur la plateforme.

[>R1]: La donnée c'est l'image issue d'un capteur ?

[>R2]: Oui, c'est ça.

6 - 20:52 > 24:30 Stockage des données

[>R1]: Que faites-vous des données que vous acquérez ? Est-ce que vous les conservez ? Est-ce que vous les mettez en ligne ?

[>R2]: On essaie de les stocker dans des bases de données à différents niveaux. Ça peut être juste une suite de fichiers, stockés dans des répertoires – ça c'est ce qui est fait au niveau du laboratoire.

[>R1]: Le laboratoire dispose d'un espace de stockage partagé ?

[>R2]: Oui. C'est un espace qui est géré par notre ingénieur d'étude. Il essaie de récupérer et de mettre à notre disposition des données (les données de l'IGN, par exemple). Et on y stocke aussi nos données. Le problème, c'est que ça représente des volumes de données, qui sont vraiment très importants – particulièrement dans mon équipe et dans mon domaine. C'est pourquoi on essaie d'avoir des serveurs, qui nous permettent de stocker la donnée. Ça se fait en relation avec l'université, forcément, étant donné la politique du datacenter qui arrive. On est très consommateur d'espace, à la fois pour le stockage qu'on appelle « à froid » et pour les traitements de la donnée. Au sein de l'Université de Strasbourg, on travaille en collaboration avec d'autres laboratoires : C. et I. Une plateforme de traitement massif de données est en train d'être mise en place. Elle se base sur l'infrastructure de calcul du pôle HPC de l'université. Dans le cadre de ce projet-là, qui s'appelle A2S et auquel le projet T. contribue particulièrement, en lien également avec un projet de CPER obtenu entre 2015 et 2020, il y a

eu tout un investissement pour acheter des serveurs de calcul, qui sont stockés à l'HPC. Ces serveurs permettent de faire du stockage et du traitement massif de données. Il y a donc un lien très fort entre les projets T. et A2S pour le traitement de la donnée massive.

[>R1]: Quelle distinction faites-vous entre stockage à froid et stockage à chaud ?

[>R2]: A chaud, c'est quand on récupère les données pour faire des calculs. A froid, c'est la sauvegarde des données.

7 - 24:30 > 31:48 Sources de données

[>R1]: Concernant les banques d'images satellitaires, quels sont les droits d'utilisation de ces données ? Est-ce que leur utilisation est libre ?

[>R2]: Non, on signe toujours une sorte de licence d'utilisation, qui limite l'utilisation des données dans le cadre public de la recherche ou de l'enseignement. C'est comme pour le Géoportail de l'IGN. C'est une licence « recherche et enseignement ».

[>R1]: Les données sont-elles gratuites ?

[>R2]: Ces données ne sont jamais gratuites, parce qu'il y a toujours quelqu'un qui paie derrière. Mais elles sont en accès libre : l'utilisateur ne paie pas l'accès aux données. Voilà, il faut simplement bien distinguer la gratuité : quand on a accès gratuitement à des images, c'est quand même derrière le CNES qui paie Airbus pour la distribution des images. Donc il y a toujours un payeur. Quand on a accès aux images Sentinel par Theia ou par le Sky Hub de l'ESA, c'est l'Europe derrière qui paie les images ; mais les chercheurs et la communauté ont un accès libre à cette donnée. Il y a un satellite qui a été lancé, c'est l'Europe qui paie. ll y a des stations de réception, c'est l'Europe qui paie. Et après il y a une mise à disposition libre de ces données-là.

[>R1]: Ça m'intéresse justement, parce que j'essaie de comprendre quelle valeur ont les données.

[>R2]: Moi j'ai un tout petit peu travaillé là-dessus. Sur le spatial en particulier : j'ai essayé, dans le cadre de mon HDR, de quantifier le coût de l'image. Il y a des études qui ont été faites. Aux États-Unis, l'imagerie Landsat a été mise à disposition en accès libre depuis bien plus longtemps qu'en Europe. Donc il existe des études montrant que les utilisations sont de plus

en plus importantes. Ce sont des études sur le chiffrage. Ça peut peut-être vous intéresser. Même au niveau de l'ESA, il y a des publications sur la valeur de la donnée. C'est le nerf de la guerre.

[>R1]: Qu'est-ce que l'ESA ?

[>R2]: L'ESA c'est l'agence spatiale européenne. L'ESA c'est au niveau de l'Europe et le CNES c'est au niveau de la France. L'ESA regroupe les représentants des différentes agences spatiales nationales. Tous les programmes de données liés à l'imagerie satellitaire sont en ligne via l'ESA. Copernicus, par exemple – c'est le programme de l'ESA pour diffuser de l'imagerie satellitaire et des produits dérivés à l'échelle de l'Europe. Donc ça regroupe tous les programmes d'occupation du sol, ainsi que tout ce qui est lié aux catastrophes naturelles (puisque, quand il y a des évènements particuliers, une charte est activée pour qu'il y ait des accès gratuits à de l'imagerie, afin de produire de l'information pour les services de protection civile par exemple). Il y a plein, plein de points d'entrée. C'est assez complexe, le domaine de l'imagerie. C'est un sujet d'actualité, quoi.

[>R1]: Est-ce que les géographes apprennent à utiliser ces ressources pendant leur formation ?

[>R2]: Oui. Les géographes sont formés dès les premières années. Je donne moi-même un cours Collecte, sources et acquisition des données en deuxième année de Licence. Manipuler de la donnée c'est le cœur de métier du géographe.

[>R1]: Est-ce que les étudiants ou doctorants sont sensibilisés à la question des métadonnées, de l'ouverture des données... ?

[>R2]: Oui, de plus en plus on les forme à ça. Il y a des formations en géographie, où l'on peut se spécialiser dans le domaine de la géomatique et dans le traitement de l'information géographique numérique. Nous, on a un cours spécifique sur l'information géographique, dans lequel on parle des métadonnées, des serveurs de cartographie en ligne, etc. Oui, on les sensibilise à ça. On voit les normes Inspire, les ISO 19115…

[>R1]: Qu'est-ce que c'est ISO 19115 ?

[>R2]: Ce sont des normes de formalisation d'échange de données géographiques. Il y a des normes qui disent : « voilà, il faut tel et tel descripteur dans les fiches de métadonnées ». Inspire utilise des normes ISO. Ce sont des normes conseillées, sans être obligatoires. Donc ça

n'apporte rien à un laboratoire d'être labellisé ISO. Pour un laboratoire de recherche, ça n'a pas d'intérêt ; mais pour certains organismes, qui sont à l'interface entre privé et public, respecter certaines normes dans la production d'informations géographiques peut avoir son importance (afin d'être reconnu par l'ESA au niveau européen, par exemple). En tout cas, c'est une norme qui est censée se généraliser. Dès qu'on manipule et qu'on échange des données géographiques, le meilleur moyen c'est d'avoir une certaine formalisation.

[>R1]: Vous utilisez ces normes dans GéoGrandEst ?

[>R2]: Oui, les métadonnées de GéoGrandEst sont fondées sur Inspire.

8 - 31:48 > 35:37 Échanges de données entre collaborateurs

[>R1]: J'avais une question sur l'échange de données entre les partenaires.

[>R2]: Généralement, on met en place une sorte de convention entre les laboratoires, pour que les données mises à disposition par l'un ou l'autre restent utilisées dans le cadre du projet, avec la mention liée au projet. C'est plutôt comme ça qu'on formalise. Parce qu'entre laboratoires de recherche on n'est pas obligé de faire un accord de consortium stipulant toutes ces affaires-là, comme quand on travaille avec un partenaire privé. On a quand même, dans nos laboratoires, une charte de confidentialité : on n'est pas censé diffuser les données comme ça.

[>R1]: Est-ce le laboratoire qui est propriétaire des données terrestres que vous acquérez ?

[>R2]: Oui, c'est plutôt le laboratoire qui est propriétaire. Après, il peut y avoir des labels différents, puisqu’on est des laboratoires CNRS. On va avoir des données qui sont rattachées au labo, mais sous l'étiquette CNRS.

[>R1]: Concrètement, comment échangez-vous ces données entre partenaires ? J'ai vu que vous aviez un site internet pour le projet T. Est-ce que ça passe par là ?

[>R2]: Non, parce qu'il faudrait créer un espace de stockage trop important. Souvent, le plus simple c'est un site FTP, qu'on met en place pour permettre un accès aux données en lecture et/ou en écriture. Après, ce sont des échanges classiques sur disque dur. Mais le plus commun c'est un site FTP.

[>R1]: Est-ce qu'un site FTP induit forcément la notion de cryptage ?

[>R2]: Oui, il y a une notion de sécurité derrière. C’est un accès sécurisé à un serveur de données à distance, dans le sens où il faut un login et un mot de passe pour s'y connecter. Il n'y a pas de cryptage, mais c'est quand même sécurisé. On ne crypte pas les données, quand on se les échange. Le cryptage c'est très différent.

[>R1]: A quoi vous sert le site internet du projet T. ?

[>R2]: Il a d'abord une fonction de communication auprès du public et de l'ANR. Et il a également une fonction de communication entre nous, puisqu'on a aussi un intranet, où l’on met les comptes rendus de réunion, des articles qu'on voudrait s'échanger, etc.

9 - 35:37 > 42:26 Résultats de recherche

[>R1]: Est-ce que le projet aboutira à créer des applications ?

[>R2]: Ce ne sont pas des applications. Ce sont plutôt des thèmes d'étude, sur lesquels on travaille. On ne crée pas une application. On traite de la donnée et on essaie de trouver des méthodes. L'idée c'est de travailler avec de l'open source et de proposer du code, qui soit en libre accès.

[>R1]: Donc vous produisez plutôt une méthode qu'un outil abouti ?

[>R2]: Ça ne peut pas être un outil abouti, parce que ça n'est pas notre travail. En recherche, on ne crée pas un outil ou une plateforme ; on développe des codes, qui peuvent être réutilisés et appliqués pour d'autres choses. Donc la production est plutôt au niveau développements, méthodes voire rendus sous forme de cartographie (mais ce sera souvent validé sur de petits sites d'étude, afin de prouver que la méthode fonctionne et donne tel résultat).

[>R1]: C'est de la recherche fondamentale ?

[>R2]: Oui, mais qui est quand même appliquée à des thèmes demandés. Ça reste de la recherche appliquée.

[>R1]: Avez-vous établi les thématiques en réponse à une demande des collectivités ?

[>R2]: Oui.

[>R1]: Allez-vous leur fournir quelque chose au terme du projet ?

[>R2]: Non. On va leur montrer ce qu'on peut arriver à faire avec des données multi-sources et à haute fréquence temporelle. Ils pourront télécharger le code source, si c'est un code en libre accès. Après, si ça intéresse une collectivité, elle a plusieurs choix : soit elle a quelqu'un qui peut prendre en main un code et traiter de la donnée ; soit elle fait appel à des prestataires, pour faire du traitement, etc. L'idée d'avoir cette plateforme A2S, qui fait du traitement massif, c'est que des gens extérieurs comme des collectivités puissent demander une prestation de type : « voilà, je voudrais avoir une cartographie mensuelle de tel ou tel mode d'occupation du sol ».

[>R1]: La plateforme A2S peut faire de la prestation de service ?

[>R2]: Oui, puisque ce sera une plateforme. Ça veut dire qu'elle peut faire un certain type de prestation de service liée à du traitement massif.

[>R1]: Ça veut dire que l'échange avec les collectivités territoriales est indirect ?

[>R2]: Oui, tout à fait. On ne répond pas directement à un appel d'offre, qu'ils auraient émis. Parce qu'on n'est pas bureau d'étude. La recherche n'a pas vocation à répondre à une commande publique.

[>R1]: Qu'est-ce qui figure dans les publications ? Qu'est-ce que vous y exposez ?

[>R2]: Les publications peuvent avoir deux vocations. Soit elles sont à vocation méthodologique et ce seront donc plutôt les codes qui seront mis en avant ; soit elles ont un aspect plus applicatif et là ce seront les données utilisées, les données en sortie et la manière dont celles-ci sont traitées, qui seront abordées.

[>R1]: Pour vous, les codes sources sont-ils des données ?

[>R2]: [Hésitation]. Ça dépend ce qu'on admet derrière le terme « donnée ». Si c'est quelque chose de brut qu'on produit, non, moi je n'appellerais pas ça une donnée.

[>R1]: Ce serait plutôt une forme de connaissance ?

[>R2]: Non... Je ne sais pas, quand on fait de la propriété intellectuelle, c'est une sorte de savoir-faire.

[>R1]: C'est brevetable ?

[>R2]: Ça n'est pas brevetable mais – comment dire – c'est protégeable.

[>R1]: On pourrait le mettre sous droit d'auteur ?

[>R2]: Oui, c'est ça. C'est ce qui se fait maintenant avec les Creative Commons et autres licences du libre. On labellise le code source.

[>R1]: Un code source serait davantage comparable à une publication, dans le sens où il demande une certaine créativité.

[>R2]: Oui, une production. C'est pour ça que, pour moi, la donnée reste quelque chose de brut, quelque chose qui est en entrée. Le code, ce n'est pas une entrée, c'est une sortie.

10 - 42:26 > 45:17 Ouverture des données

[>R1]: Est-ce qu'on peut revenir rapidement sur toutes ces recommandations : la directive Inspire, le mandat de la Commission européenne pour l'ouverture des données financées par le programme H2020…? Qu'est-ce que vous pensez de tout ça ?

[>R2]: Moi je trouve normal que toute production issue de l'argent public puisse être accessible. On doit pouvoir voir les résultats et y avoir accès. Je trouve que c'est plutôt une bonne chose. C'est l'open data, quoi !

[>R1]: Quelle utilité y voyez-vous ?

[>R2]: L'utilité c'est d'éviter de refaire x fois ce qui a déjà été fait, de capitaliser et d'améliorer les connaissances, puisqu'on aura une profondeur historique sur un certain nombre de choses, grâce à des données acquises par d'autres. Ça va permettre aussi de valider nos propres méthodologies, si on sait qu'il y a un jeu de données disponible avec une vérité terrain ou des références. Ça ne peut qu'être bénéfique, si tout le monde joue le jeu.

[>R1]: Ces données peuvent-elles être utilisées dans un but commercial ?

[>R2]: Non, justement. C'est vraiment la limitation. Ça reste dans le cadre scientifique, voire ça peut percoler dans le milieu privé, si l'organisme privé s'approprie un des codes et propose autre chose/une valeur ajoutée. Il ne va pas vendre le code, il va vendre un produit, qui a une valeur ajoutée, directement utilisable par la collectivité. La collectivité, elle s'en fiche d'avoir le code. Elle n'aura dans le service ni les moyens, ni les compétences pour traiter tout ça.

Donc ce qui va l'intéresser, c'est la carte ou la donnée transformée – i.e. avec une valeur ajoutée.

11 - 45:17 > 49:28 Valeur économique des données

[>R1]: Vous m'avez dit que vous aviez essayé d'évaluer la valeur des données géographiques…

[>R2]: Même en termes économiques, c'est toujours difficile d'estimer et d'attribuer des coûts.

[>R1]: Comment vous y êtes-vous prise ?

[>R2]: J'ai regardé un peu ce qui existait, mais je n'ai pas essayé de quantifier quoi que ce soit. Je vous donnerai le lien vers les quelques articles sur ce sujet. Je sais qu'il y a une doctorante à l'université, qui travaille sur la valeur de l'information géographique. Elle est en sciences économiques, au B. Elle travaille avec la région.

[>R1]: C'est intéressant. Je n'ai pas l'impression que cette question se pose encore dans les autres domaines.

[>R2]: Alors que dans le domaine géographique, effectivement, il y a beaucoup de réflexions là-dessus.

[>R1]: Comment vous l'expliquez ?

[>R2]: La réflexion est plus ancienne, parce qu'en géographie la donnée est le cœur de métier. Mais il est certain que la question va se poser pour tout type de donnée, produite dans tout type de laboratoire. Beaucoup de questions se posent sur la valeur économique de la donnée et de son cheminement. En particulier avec l'ouverture des données. Ça a beaucoup évolué depuis l'ouverture des données par l'IGN, notamment.

[>R1]: Je me pose la question : si les données coûtent si cher, pourquoi les rend-on librement accessibles ? Ça signifie qu'un industriel va pouvoir les utiliser, créer une valeur ajoutée et faire du profit.

[>R2]: C'est une vaste question. Ça reste une politique de vision à long terme. Les États-Unis ont tablé là-dessus depuis plus trente ans. Les images Landsat, par exemple, sont mises à

disposition gratuitement par l'USGS depuis le début (l'USGS c'est l'équivalent de l'IGN). Ils jouent sur le fait que ça va apporter de la valeur ajoutée ailleurs.

[>R1]: Vous savez s'ils ont d'ores et déjà pu mesurer le retour sur investissement ?

[>R2]: A mon avis, ils ont dû travailler là-dessus. S'ils l'ont fait, c'est qu'ils ont eu un certain retour sur investissement. Sinon, ils auraient arrêté.

12 - 49:28 > 55:27 Parcours professionnel

[>R1]: Dans quel(s) établissement(s) avez-vous effectué votre formation universitaire ?

[>R2]: Je suis d'origine belge. J'ai fait mes études de géographie à Namur les deux premières années, puis à Louvain-la-Neuve les deux années suivantes. Ensuite j'ai fait une cinquième année à l'Université de Strasbourg en géographie. Puis une thèse. J'ai eu mon premier poste à Caen pendant trois ans, puis je suis revenue ici à Strasbourg.

[>R1]: Est-ce que vous avez constaté des différences dans la gestion des données ?

[>R2]: Oui, de grandes différences entre les laboratoires. Il faut avoir les moyens. Stocker les données demande un investissement humain et financier. Or tous les laboratoires ne peuvent pas le faire. En géographie, on est un peu plus sensibilisé à ça. Mais il peut y avoir des différences assez énormes entre un laboratoire comme le nôtre et un laboratoire qui va davantage être orienté vers les sciences sociales. Stocker des données d'enquête, c'est effectivement un autre problème. Il y a des chercheurs dans le laboratoire, qui pourraient vous parler des données liées à des enquêtes. Il y a également l'interaction géographie/santé. C'est une problématique particulière, très en lien avec les individus. On va collecter des données sur des individus, afin d'étudier par exemple l'obésité, les allergies ou tout ce qui a trait à l'environnement. Dans le laboratoire, il y a différents axes : certains travaillent plutôt sur la mobilité/accidentologie, d'autres sur la mobilité/déplacement, d'autres sur la climatologie urbaine... C'est pour ça qu'on ne dit plus qu'on est un laboratoire de géographie. On est un laboratoire interdisciplinaire. Il y a des géographes, mais pas seulement. Il y a par exemple une physicienne de l'atmosphère, des hydro-écologues... C'est très diversifié. Ça répond à de multiples problématiques environnementales.

[>R1]: Au niveau du volume de données, ce sont surtout les images satellitaires qui sont gourmandes en stockage ?

[>R2]: Oui, je pense qu'on est un gros consommateur. Une image à très haute résolution peut par exemple atteindre 1 Go. On est vraiment dans des problématiques de volume. C'est pour ça aussi qu'on travaille avec des informaticiens. Parce qu'on ne peut plus se permettre de stocker ça sur des petits disques durs, avec lesquels on se balade. C'est fini.

Entretien avec le chercheur 51 (sciences politiques)

1 - 01:48 > 10:52 Objectifs scientifiques du projet de recherche

[>R1]: Dans un premier temps, pourriez-vous me résumer les objectifs scientifiques du projet ?

[>R2]: C'est un projet qui essaie d'analyser la manière, dont la question du conflit d'intérêt a émergé et a été prise en charge par les autorités publiques dans le secteur du médicament. Les autorités publiques de régulation, mais aussi les acteurs privés concernés par l'évaluation des médicaments et la publication de données de recherche cliniques. Nous ne sommes pas ceux qui allons traquer le conflit d'intérêt. Je veux dire, nous ne sommes pas là pour dire : « regardez, celui-ci a un conflit d'intérêt grave » ou « non, là ce n'est pas un cas de conflit d'intérêt ». Ce n'est pas notre problème. Notre problème c'est de regarder comment ont émergées différentes règles et modalités visant à prévenir ce qui a été identifié comme problème, à savoir les conflits d'intérêt. Tout ça vous le trouvez sur le site Internet. Il y a le site officiel de l'ANR, et puis on a fait un blog, que vous avez dû trouver.

[>R1]: Un carnet de recherche ?

[>R2]: Oui. Il y a une partie, qui est centrée sur les revues médicales, donc sur les publications scientifiques. Là, les rédacteurs en chef et les comités de lecture se sont saisi de ces questions et ont essayé de proposer des modalités pour identifier un conflit d'intérêt. Par exemple, tous les auteurs doivent dire, en marge de l'article qui présente les résultats d'essais cliniques, par qui leur recherche a été financée, pour qui ils ont travaillé... Voilà, il y a différents journaux internationaux. Évidemment, les journaux les plus centraux et à la pointe – le *New England Journal*, le *Journal of American Medical Association* – ont réfléchi à la question et ont fait des propositions/des préconisations, qui ont été reprises et qui sont peu à peu rentrées dans les pratiques des chercheurs. Donc il y a une partie comme ça dans le projet. C'est un peu l'histoire et la sociologie de ces outils que sont les revues, avec aussi des allers et retours, des controverses... Il y a des numéros qui traitent de la question. Nous, ça nous informe sur comment le problème est posé. Ça nous donne une certaine vision de ce qu'est un conflit d'intérêt. Ce qu'on essaie de voir également dans la manière dont ce problème est

construit, c'est la confusion/le flou qui existe entre lien d'intérêt et conflit d'intérêt. Ce n'est pas parce qu'on a des liens qu'il y a des conflits. La deuxième partie du projet est consacrée aux dispositifs qu'ont mis en place les autorités publiques françaises, grosso modo, après le scandale du Médiator : à savoir la loi Bertrand de 2011, où on a dit « il faut que les industriels rendent publics les liens qu'ils ont avec la recherche et les résultats de la recherche ». Des liens financiers. Ça n'est pas seulement « je finance une recherche », c'est aussi « j'invite tel médecin à tel évènement ». Il s'agit d'un transfert financier, donc je dois le savoir. On étudie la base Transparence Santé, qui a été mise en place à ce moment-là. C'est une base de données publique, dans laquelle vous pouvez rechercher le nom de votre médecin traitant et voir tout ce que celui-ci a pu recevoir de la part d'industries ou de producteurs de médicaments. Ça va de 20€ (un repas lié à une visite dans un hôpital – il était là et il a signé la feuille d'émargement de la présentation du nouveau médicament) à des conventions pour recherches. Ce sont les industriels qui sont obligés de faire ces déclarations et de remplir le tableau. Ce ne sont pas les médecins qui le font. La question que, nous, on se pose, c'est de savoir ce que cette publicisation des liens financiers entre des professionnels de santé et des industriels a pu changer dans les comportements/les pratiques ordinaires de transferts financiers. Pour l'instant, les autorités de régulation n'en font pas grand chose – pour ce qu'on en sait. C'est toujours l'idée de dire : « voilà, on vous informe, vous le public », « on ne vous cache rien ». Ça ne veut pas dire que, s'il y a un lien ou un transfert financier, celui qui signe une étude sur les essais d'un médicament travaille forcément pour l'industriel. Donc on travaille sur ce sujet, pour savoir si ça a changé quelque chose. Et puis, à côté de ça, il y a d'autres dispositifs qui sont mis en place, toujours dans le sens de la transparence. Il y a toutes les déclarations d'intérêt que doivent faire les experts qui vont évaluer le médicament, lorsqu'ils sont dans les comités comme la Haute Autorité de Santé, le Comité d'Évaluation de Médicaments ou le fameux comité qui doit décider si un médicament est remboursé ou non par la Sécurité sociale. Ce sont des spécialistes du domaine professionnel de santé, qui analysent, discutent et qui doivent publier une déclaration d'intérêt, en disant qu'ils ont travaillé pour untel sur tel essai. Nous, la question qu'on se pose, c'est : qu'est-ce que cette déclaration d'intérêt ou cette obligation de déclaration d'intérêt fait à leurs pratiques ? Est-ce que ça change quelque chose ou pas ? Qu'en font les gens qui nomment ces experts dans les conseils/autorités/etc. ? C'est-à-dire : est-ce qu'ils consultent ces déclarations d'intérêt ? Est-ce qu'ils se disent : « Ah ben

non, je ne vais pas nommer untel, parce qu'il a travaillé pour Servier. On ne voudrait pas que dans le comité il y ait des gens ayant travaillé pour Servier ». Voilà, c'est ça l'enquête. Et puis il y a un troisième volet, dont je ne m'occupe pas tellement, qui porte sur certains médicaments : les médicaments Alzheimer, le Benfluorex… Là, en suivant le médicament, on essaie de saisir les liens entre, d'un côté, les experts de l'administration et les professionnels de santé et, de l'autre, les producteurs industriels.

2 - 10:52 > 12:32 Répartition des rôles

[>R1]: Comment sont répartis les rôles entre les différents axes ?

[>R2]: Dans le projet, on est 6 ou 7 permanents. Certains sont à mi-temps. Ça dépend, c'est un peu en fonction de nos compétences – parce qu'il y a des juristes, des historiens des sciences, des sociologues, des politistes. Et c'est un peu en fonction de nos questionnements par ailleurs. Il y a H., qui a pas mal travaillé sur la question de l'évaluation de la réglementation du médicament. Il y a B., qui a lui aussi travaillé sur la réglementation de produits. Il y a mon collègue C., qui travaille plus sur l'histoire du médicament. Voilà, on aborde le projet avec ces regards différents.

[>R1]: Chacun mène ses recherches de son côté ?

[>R2]: On se voit tous les mois ou tous les deux mois. On dit chacun où on en est, on discute de ce qu'on a vu et puis on redivise à chaque fois le travail.

3 - 12:32 > 18:10 Mode d'acquisition des données

[>R1]: Quels types d'informations collectez-vous ? Comment vous y prenez-vous ?

[>R2]: Moi je collecte deux types d'informations. La première, ce sont des informations type déclarations d'intérêt, qui sont publiées, donc il s'agit de les collecter. On voudrait arriver à les traiter en essayant de faire une base de données à partir de la liste des experts, qui sont dans différentes commissions. On prendrait leur déclaration d'intérêt et regarderait si ces experts ont des choses en commun. On va comparer aussi entre l'avant et l'après la loi. Ça c'est un premier travail, qui est à la fois un travail de prosopographie (c'est-à-dire sur les propriétés sociales de ces individus) et de traitement de l'information. L'idée c'est aussi de travailler sur

les usages de ces déclarations d'intérêt : rencontrer des gens qui ont fait ces déclarations d'intérêt et leur demander si l'obligation de faire une déclaration d'intérêt a changé quelque chose pour eux (si c'est très compliqué ou non, etc.). Ça c'est le deuxième type de matériau, qui se présente plutôt sous forme d'entretiens.

[>R1]: Ce sont des entretiens semi-directifs ?

[>R2]: Oui, c'est ça.

[>R1]: Pour ce qui est de la base de données, avec quel outil travaillez-vous ?

[>R2]: Pour la prosopographie ?

[>R1]: Oui.

[>R2]: Ce n'est pas moi qui vais faire le traitement quanti, ce sont les ingénieurs d'études. Nous, on va collecter sous Excel. Ensuite, je pense que les ingénieurs vont importer ces données sous SPAD. SPAD est un logiciel d'analyse géométrique des données. On n'est pas dans une logique, où on veut saisir les effets de certaines variables sur un comportement (on ne fait pas des régressions ou des choses comme ça). On met en évidence des variables qui vont ensemble. C'est une vision d'un espace social, avec un groupe typique de gens, qui ont des propriétés allant ensemble, par rapport à un autre groupe de gens, qui ont d'autres propriétés. Je ne sais pas si vous connaissez ce type d'analyse...

[>R1]: Pas vraiment. Je ne connais pas du tout SPAD.

[>R2]: C'est un logiciel qui a été inventé par des gens pour faire des analyses de correspondances ou des analyses factorielles de correspondances, dans la perspective de l'espace social ou du « champ » de Pierre Bourdieu. C'est Brigitte Leroux qui a mis au point cet outil informatique. Aujourd'hui la plupart des laboratoires de sociologie l'utilisent. On le trouve notamment sur la logithèque de l'Université.

[>R1]: C'est un logiciel open source ?

[>R2]: Ça n'est pas open source, c'est un logiciel qui s'achète. Mais on a des licences collectives. On forme nos étudiants, etc. Moi je ne sais pas le faire fonctionner en tant que tel. Je le comprends et je sais l'interpréter. Mais la manip en tant que telle, je ne sais pas la réaliser.

[>R1]: Qu’obtenez-vous en sortie ?

[>R2]: En sortie, on peut caractériser des populations. C'est plutôt descriptif dans un premier temps : on peut caractériser ces populations, on peut montrer des oppositions internes ou des choses qui vont ensemble... Après, on peut mettre en évidence des éléments typiques, qui vont être la contribution principale sur une population.

4 - 18:10 > 31:21 Ouverture des données

[>R1]: A votre avis, est-ce que cette base de données pourrait avoir une valeur pour des personnes externes au projet ou pour vous, mais dans le cadre d'un autre projet ?

[>R2]: C'est toute la question, dont on a essayé de débattre lundi et mardi à ce fameux colloque. C'est tout l'enjeu. C'est-à-dire que, les uns ou les autres, nous produisons des données pour les besoins de notre recherche, avec une question bien particulière. Mais qu’en fait-on ensuite ? Est-ce qu'on les pérennise pour des enquêtes futures ? C’est ce que font M. et d'autres avec la base de données sur les députés européens. Pour chaque législature, il s'agit de mettre à jour la base, en complétant les nouveaux entrants, ceux qui s'en vont et ceux qui restent au Parlement européen, et de pouvoir décrire et expliquer une évolution ou l’absence d'évolution. Ensuite, qu'est-ce qu'on fait de ça, quand les gens ont fini de travailler là-dessus ? Est-ce qu'on peut partager avec d'autres ? A quelles conditions ? Tout le monde n'est pas d'accord là-dessus.

[>R1]: Vous, vous en pensez quoi ?

[>R2]: C'est ambigu, parce que moi j'adorerais travailler sur des données qui ont été collectées dans les années 1960 ou 1970 par exemple – si tant est que ça soit possible... Je ne sais pas sous quelle forme on aurait ces données. Sous forme de cartes perforées ? En même temps, je sais bien que quand on fait une base de données prosopographique, les variables qu'on identifie sont des variables qui correspondent à la question qu'on s'est posée. Ce ne seront pas forcément les variables dont on aura besoin pour travailler sur les conflits d'intérêt par exemple. Mais peut-on mettre la base, qu'on est en train de constituer, à disposition d'autres ? Je ne sais pas. Dans l'idéal, oui. Ce serait super. On pourrait faire ça, oui, une fois qu’on aura terminé notre enquête et qu'on aura pu en tirer ce qu'on voulait pour ces questions-là.

[>R1]: D'autant plus que ça ne demanderait pas forcément un énorme travail. Cette base pourrait être facilement compréhensible par un chercheur du domaine ?

[>R2]: Bien sûr, ça n'est pas une question de non compréhension. Simplement, la base peut paraître incomplète sur des tas d'aspects. Non pas parce que nous n'avons pas les données, mais parce que nous n'avons pas jugé bon de les faire figurer dans la base.

[>R1]: J'imagine qu'un certain nombre de chercheurs dans le laboratoire conçoivent, comme vous, des bases de données. Que deviennent ces bases de données actuellement ?

[>R2]: Ben, ffff... La plupart du temps, elles sont un peu la propriété du chercheur. Alors, pas la base sur les députés, parce qu'il s'agit d'un projet collectif, donc c'est un peu lié au laboratoire. Mais, pour autant, elle n'est pas ouverte à n'importe qui. On a pu la partager avec des collègues d'autres universités, qui souhaitaient s'en inspirer ou reprendre certaines données. Je trouve ça plutôt bien, parce qu'on a travaillé sur cette base : on a pris du temps, de l'argent, pour le faire et c'est bien que ça serve à d'autres. Mais c'est vrai que, à part ce cas et quelques autres, chacun est un peu jaloux de ses propres données. Ça n'est pas complètement entré dans les mœurs, je crois. Et encore, je pense que c'est peut-être plus facile pour des données de type base biographique que pour des entretiens par exemple. Avec vous, là, j'ai signé un consentement éclairé. Je suis d'accord, mais est-ce que j'ai envie que tout le monde soit au courant ? C'est vrai que je ne vous dis pas des secrets, mais qu'est-ce qu'un autre peut faire de ces entretiens ? Voilà, ça devient assez compliqué et jusqu'à présent on n'y a pas vraiment été confronté...

[>R1]: C'est vrai qu'il y a cette question de la personne interrogée. Est-ce que la personne est d'accord pour que ce qu'elle dit soit réutilisé ?

[>R2]: Voilà. Réutilisé dans quel cadre ? Elle n'en sait rien. C'est là où ça poserait problème. Parce que ça a des effets. Dans le projet d'archives orales des institutions européennes – j'en parlais ce matin avec les étudiants –, vous avez des tas d'anciens hauts fonctionnaires qui ont livré un témoignage oral à des historiens. On peut avoir accès à tous ces récits sous forme d'enregistrement ou de transcription aux archives historiques. Mais ça veut dire que ces hauts fonctionnaires ont donné leur discours sur un mode « je vous raconte toute mon histoire et l'histoire de l'institution, que j'ai servie, pour la postérité ». Donc il y a une part de reconstruction historique. Les fonctionnaires se mettent aussi en avant. Ça ne veut pas dire

qu'ils ne sont pas honnêtes avec eux-mêmes. Leur intention est plus large : ils livrent un témoignage pour la postérité. Évidemment, quand on est historien ou sociologue et qu'on travaille avec ces sources-là, on le sait et on les prend pour ce qu'elles sont. Mais ça n'est peut-être pas la même chose dans un entretien sociologique ou dans les entretiens que vous faites là. C'est un vrai enjeu, surtout s'il y a cette question de protection des données personnelles.

[>R1]: Dans la base de données des déclarations d'intérêt, est-ce qu'il y a les noms des experts ?

[>R2]: On n'est pas censé mettre les noms. Mais si on ne les met pas, on n'y comprend rien. Donc on met les noms, mais évidemment, dans le traitement qu'on va en faire et dans ce qu'on va en sortir, il n'y aura rien de nominatif. C'est le principe d'une base quanti. En revanche, il faudra sans doute qu'on donne des exemples, afin d'illustrer les tendances d'une population à partir de cas particuliers. Donc ces cas particuliers seront nominatifs. En général, ce ne sont pas des informations secrètes. Ce sont des informations qu'on obtient parce que les gens mettent leur CV en libre accès.

[>R1]: Finalement la base n'a pas à être anonymisée, si les sources sont des informations rendues publiques ?

[>R2]: Oui, mais on n'a pas le droit de faire des fichiers nominatifs.

[>R1]: Quelle règle vous l'impose ?

[>R2]: C'est la CNIL. Je crois qu'il y a des choses qui ont changé avec la loi du printemps 2018. Mais disons qu'on doit faire une déclaration à la CNIL, en décrivant ce que l'on fait et en précisant qu'on le fait à des fins de recherche, et non dans un but commercial par exemple. En toute rigueur, il faudrait que chacune des personnes, sur lesquelles on travaille, soit informée qu'on travaille sur elle. Mais comme ce sont des gens qui sont plus ou moins des personnages publics et que les informations qu'on obtient proviennent soit de leur site internet, soit des déclarations qu'ils ont faites (déclarations qui sont elles-mêmes publiques), les données ne sont pas top secrètes. Tous les juristes ne sont pas d'accord sur cette question : jusqu'où va la protection personnelle ?

[>R1]: Jusqu'à maintenant, la CNIL vous disait que vous ne pouviez pas générer de fichiers nominatifs ?

[>R2]: La CNIL nous dit qu'on peut le faire, mais qu'il faut que les personnes sur lesquelles on enquête soient prévenues. Par exemple, pour la base de données sur les députés européens, les discussions ont été assez compliquées, parce qu'on disait : « d'accord, mais certains d'entre eux sont morts, donc pour avoir leur consentement, c'est drôle votre truc ». Il me semble qu'en 2018 ça a changé. Ils ont levé le truc, en disant : « voilà, ce sont des personnages publics... ». Là où on est susceptible de nous surveiller, c'est qu'il ne faut pas qu'on ait de fichiers nominatifs. Par exemple, si on a des entretiens, il ne faudrait pas que dans nos ordinateurs on ait le nom des gens qui ont dit ceci. Il faudrait qu'on anonymise tout de suite. Mais, quand on publie, on anonymise de toute façon toujours. Quelqu'un qui connaît bien le sujet pourrait recouper les informations et retrouver la personne dont il s'agit. Mais voilà, on se débrouille. Jusqu'à présent, on n'a jamais eu de problème de ce type-là. Et puis, on n'est pas non plus dans une démarche où on va accuser quelqu'un, où on va lui faire dire des choses qu'il aurait voulu cacher.

[>R1]: Oui, il n'y a pas de polémique possible. En tout cas, ça n'est pas le but.

[>R2]: Voilà.

5 - 31:21 > 39:10 Conservation des données

[>R1]: Comment procédez-vous pour les entretiens ? Vous faites comme moi : vous enregistrez et vous retranscrivez ?

[>R2]: Oui, on enregistre, on retranscrit, et ensuite on fait une petite fiche analytique (qui est la personne, les différents points abordés, etc.). En fait, il y a deux types d'entretiens. Il y a les entretiens hyper informatifs, comme ceux qu'on a faits avec des hauts fonctionnaires ou des gens en charge d'un service. Ce ne sont pas des informations qui dépendent de la personne interrogée. Et puis il y a des entretiens où on ne cherche pas juste de l'information, mais où on essaie de saisir un peu la conception que les personnes ont des conflits d'intérêt par exemple, où on essaie d'avoir une représentation du monde. Là ce sont des entretiens dont on va citer des extraits. On analysera ces extraits comme une manière de se représenter ce qu'est le conflit d'intérêt et on essaiera de mettre en évidence plusieurs définitions du conflit d'intérêt, en rapportant toujours tel ou tel propos à la personne qui l'a dit, en fonction de son histoire, de

sa position dans l'administration, le champ sanitaire ou le champ d'expertise. Ce sont donc deux formats d'entretiens assez différents.

[>R1]: Quelle utilisation faites-vous de l'entretien informatif ?

[>R2]: Disons qu'on en cite moins d'extraits, car ce sont souvent des informations factuelles. On n'a pas besoin de rendre compte de la manière dont on nous a dit telle ou telle information. Ce sont des informations qu'on aurait pu trouver dans des documents, mais comme on ne dispose pas de document qui explique ça, on va directement demander aux personnes qui, par exemple, ont mis en place un service.

[>R1]: Conservez-vous les entretiens tout au long du projet de recherche ?

[>R2]: Oui.

[>R1]: Les conservez-vous également au terme du projet ?

[>R2]: Oui. Moi j'ai tous les entretiens que j'ai menés depuis ma thèse. Je n'ai peut-être pas tous les fichiers audio, mais j'ai en tout cas toutes les retranscriptions.

[>R1]: Avez-vous tendance à être plutôt prudente à ce niveau-là ? Par exemple, faites-vous souvent des sauvegardes sur double disque ?

[>R2]: Oui.

[>R1]: Est-ce une politique du laboratoire ou bien une initiative personnelle ? Vos données sont-elles sauvegardées sur un disque dur personnel ?

[>R2]: Oui. Je sais qu'on ne devrait pas. Si on se balade avec notre ordinateur et qu'on se le fait voler par une personne malveillante, qui trouverait ces entretiens pouvant être retenus contre les personnes... Il faudrait que les entretiens soient stockés sur un ordinateur sécurisé.

[>R1]: Y êtes-vous sensibilisée ?

[>R2]: Oui, oui. On a des réunions d'information là-dessus. On nous dit : « il faut chiffrer votre ordinateur » ou « il faut être prudent, surtout avec les portables qu'on promène partout ». Je pense que les gens qui travaillent sur des sujets un peu plus sensibles (comme le phénomène de radicalisation ou les victimes de violences), eux, font super attention. Mais bon, moi j'ai travaillé sur des hauts fonctionnaires, sur des hommes politiques... Fffff, voilà. C'est vrai que ça n'est peut-être pas prudent. Chaque fois qu'on sort de ces réunions, on se dit

« oh la la, je fais n'importe quoi, ça ne va pas ». Et puis on relativise. Quand l'enquête est finie, en général on la conserve sur un autre support (on imprime les documents et on les range dans des boîtes d'archive).

[>R1]: Qui organise les réunions d'information, dont vous parlez ?

[>R2]: Ici à l'Université de Strasbourg, c'est soit l'Université, soit le CNRS. Les réunions s'adressent aux directeurs d'unité, qui ensuite doivent rapporter les informations aux membres du labo. Les ingénieurs d'études, ici, sont très sensibilisés. Ce sont eux qui sont au contact de la CNIL, qui gèrent les bases de données, etc. Ils sont très au fait de ces questions et ils nous disent ce qu'il faut faire (déclarer à la CIL, etc.). Pour les entretiens, en revanche, c'est vrai qu'on n'est pas...

[>R1]: Pour ce qui est de la base de données, qui va être créée dans le cadre du projet, qui la conservera : les ingénieurs d'études sur les serveurs du laboratoire ou bien vous-mêmes ?

[>R2]: Ce sera un peu tout le monde, puisqu'on est plusieurs dans l'équipe à alimenter la base. On travaille avec l'espace partagé MyCore du CNRS, dont on nous a dit qu'il était sûr. Donc la base est sur MyCore. C'est un fichier Excel.

[>R1]: Vous n'êtes pas la seule à travailler sur cette base ? Tous les partenaires y contribuent ?

[>R2]: Oui, il y a tous ceux du projet M.

[>R1]: D'accord, c'est une base commune.

6 - 39:10 > 42:49 Ouverture des données

[>R1]: Qu'est-ce que vous pensez de toutes ces questions d'ouverture des données, de réutilisation... ?

[>R2]: Je pense qu'on travaille dans une institution et un laboratoire et qu'à ce titre-là ce qu'on fait doit être propriété de l'institution et du laboratoire. Donc je suis sur une ligne très collectiviste de ce point de vue-là, en disant : « Voilà, on n'est pas des entrepreneurs indépendants. Ce qu'on fait, on le fait aussi grâce aux moyens qu'on met à notre disposition dans le cadre du laboratoire, du CNRS, de l'Université, etc. Donc ce qu'on produit (nos

recherches) est quand même estampillé par nos institutions ». Donc ces données devraient être propriété du laboratoire. Mais ça voudrait dire – et j'en suis bien consciente – que chacun renonce à en tirer les produits symboliques, scientifiques et personnels, qui font avancer une carrière ou qui font qu'on peut publier un livre ou un article. Ça c'est une première chose. Mais je ne suis pas pour imposer à tout le monde de déposer ses archives, parce qu'on n'y arrivera pas et que c'est bien beau de déposer des archives, mais encore faut-il qu'il y ait un usage derrière. Et, de fait, ce qui se passe, c'est qu'une fois qu'un terrain ou un objet a été étudié par quelqu'un, il n'y a pas d'intérêt à y retourner. Un nouveau chercheur a tout intérêt à avoir son domaine à lui, plutôt que de se mettre dans les pattes de quelqu'un et dire « ah ben oui, tu fais juste » ou faire de la réactualisation. Alors, c'est intéressant 50 ans après, parce qu'on se dit : « Ah tiens, machin avait montré que… etc. Qu'en est-il 50 ans après ? ». Là, ça pourrait être intéressant. Voilà, il y a cette dimension-là. Que vous dire d'autre sur ces données ? Oui, ça pose aussi un problème, quand on est plusieurs sur un même projet. Qui a la propriété des données ? Qui a le droit d'utiliser quoi ? Là, c'est sûr qu'il faut bien s'entendre au départ, afin qu'il n'y ait pas de querelles telles que « il m'a piqué mes données » ou « moi j'ai fait tout le boulot et c'est lui qui publie l'article ». Voilà, il faut être bien sûr que les articles sont coécrits par tous ceux qui ont contribué. Ça n'est pas évident. Mais, en ce qui me concerne, j'ai toujours bien réussi à organiser ça. Dans les enquêtes collectives que j'ai encadrées, on a publié collectivement – alors, chacun son chapitre, mais dans un ouvrage collectif qui faisait sens, surtout pour les jeunes chercheurs.

7 - 42:49 > 49:53 Réutilisation des données

[>R1]: Au sein du laboratoire, vos collègues vous ont-ils déjà demandé d'avoir accès à telle retranscription ou telle base de données ?

[>R2]: Non. Je sais ce sur quoi travaille tout le monde, parce que j'ai été directrice du laboratoire jusqu'à l'année dernière. Donc, effectivement, j'avais une vue peut-être plus large que certains. Mais personne n'a demandé à avoir accès à mes données. Parce que chacun est un peu dans un domaine spécifique. On parle de nos sujets respectifs, mais de là à réutiliser des données... A part ce projet sur les élites européennes, qui est plus collectif... Ça m'est arrivé de donner des retranscriptions d'entretiens à un collègue, parce que je savais qu'il

travaillait sur tel sujet et que ça m'était arrivé de rencontrer telle personne mais sur un sujet autre que le sien. En gros, sur 15 pages de retranscription, il y avait une page qui pouvait l'intéresser. J'ai fait ça une ou deux fois. Mais, chaque fois, le collègue avait déjà les informations, parce qu'il avait lui aussi fait un entretien avec la personne. Donc ça ne lui a rien apporté de plus. Ça lui a peut-être donné un point de vue sur un autre sujet que ce que son même personnage avait pu lui dire.

[>R1]: C'était de votre propre fait ?

[>R2]: Oui. Mais c'est assez rare. Après, je pense qu'on n'utilise pas assez les données qui sont produites. C'est une réflexion qu'on a eue au niveau de la MSH. C'est aussi la raison pour laquelle on a mis en place de cette plateforme universitaire des données. Parce qu'effectivement ça coûte très cher de produire des données et que ces données sont insuffisamment exploitées par rapport à tout ce qui existe. Et on peut vraiment faire des choses, ça c'est super.

[>R1]: Vous faites référence aux données de la statistique publique ?

[>R2]: Oui, mais je crois que c'est amené à être élargi à d'autres données. Par exemple, il y a l'entreprise BeQuali. On n'est pas forcé d'avoir son petit stock à soi, artisanalement constitué. Il y a une forme de mise en commun. Le problème c'est que, nous, on insiste beaucoup sur les conditions de production de ces données. Elles font partie des données. Quand on récupère les données d'autres, on n'a pas forcément les conditions de production qui vont avec. Et ça c'est un vrai problème. Alors, pour la statistique publique, on le sait, parce qu'on nous donne tout un ensemble de choses : ça n'est pas juste le tableau, on a également les non réponses, la méthode de codage des réponses aux questionnaires, etc. Il y a plus de métadonnées que de données. Ce sont ces métadonnées qui sont importantes. Je ne suis pas certaine qu'on puisse avoir tout ça, si on mettait en commun les entretiens que tout le monde a faits. C'est ça la vraie difficulté.

[>R1]: Finalement il faudrait un service à part entière pour s'occuper de tout ce travail de contextualisation des données. L'INSEE, par exemple, c'est leur job. Mais est-ce que c'est le job d'un chercheur ?

[>R2]: Un chercheur c'est son job de documenter ses données, c'est ce qu'il fait tout le temps. Dans le projet M., on passe notre temps à expliquer aux autres comment on a eu telle info, ce

qu'on pense qu'elle veut dire et comment il faudrait continuer ou réorganiser. Le but c'est que le partenaire puisse utiliser ces données. Il faut qu'il sache tout ça, parce que c'est un travail collectif. Donc on est tout le temps en train de le faire. Et quand on écrit un article ou un livre, on est tout le temps en train d'expliquer comment on a produit ces données ou comment on les a récupérées et ce qu'on en a fait. Ça c'est notre boulot. Mais est-ce que c'est notre boulot de le faire comme le ferait un archiviste ? Je ne sais pas. C'est ça l'enjeu principal. Parce qu'après on va crouler sous les données. Ce n'est pas le problème de la masse des données – il y en a déjà trop. Le problème c'est : qu'est-ce qu'on veut en faire ? Parce qu'avoir des archives pour avoir des archives...

[>R1]: Savez-vous pourquoi les données disponibles sont peu utilisées ? Je pense à celles de l'INSEE ou du CASD.

[>R2]: Je ne sais pas. C'est une culture scientifique qui n'est peut-être pas encore suffisamment développée. On n'a pas forcément le réflexe de consulter les recensements et enquêtes qui existent déjà et qui nous fournissent par exemple des données sur l'état de la santé ou de l'éducation en France. Je pense que c'est plutôt de l'ordre du réflexe ou de l'habitude de chercheur. On essaie de formaliser ça dans nos formations de master, de doctorat, pour que quelqu'un qui commence une thèse ne réinvente pas des données, mais parte de ce qui a déjà pu être fait et vienne compléter ces choses-là.

8 - 49:53 > 53:54 Valeur des données

[>R1]: Les données générées dans le cadre du projet M. peuvent-elles être utiles à la société ou aux entreprises ? Quelle utilisation celles-ci pourraient en faire ?

[>R2]: Pour les entreprises, je ne sais pas. Pour la société, sans doute, mais peut-être pas directement comme ça. En tout cas, on n'y a pas encore pensé. Parce qu'on n'est pas encore nous-mêmes au bout du processus. Mais, par exemple, O. et M. ont un partenariat avec *Alternatives économiques* pour publiciser leurs données sur l'Assemblée nationale française. Ils ont fait un énorme travail de collecte de données sur les parlementaires depuis 60 ans et ils mettent à disposition ces données sur le site *Alternatives*, sous forme d'infographies, pas seulement descriptives mais qui viennent aussi alimenter une réflexion sur la professionnalisation politique. Ça permet notamment d'aller à l'encontre de certains discours

un peu faciles, de rétablir des vérités objectives et, à la fois, de réfléchir sur la place et le rôle du député dans la vie politique, dans la vie sociale, etc. Donc c'est possible et je pense qu'il faudrait le faire. Mais c'est encore une énième profession qu'on nous rajoute.

[>R1]: Vous dites que O. et M. ont mis en ligne leurs résultats sous forme d'infographies. Ont-ils également mis à disposition les données brutes sous-jacentes ?

[>R2]: Pour l'instant, ils n'ont pas mis à disposition les données brutes. Mais vous avez des ONG qui font ce travail, c'est-à-dire qui mettent à disposition des données pouvant être retravaillées ensuite. C'est ce qu'on a fait avec la base Transparence Santé : on en a extrait les données brutes. Il y avait plein d'erreurs d'ailleurs, plein de doublons. On a passé un temps énorme à nettoyer la base.

[>R1]: Vous avez pu l'exporter ?

[>R2]: Oui. Mais il suffit qu'on ait seulement l'initiale du prénom (et pas le prénom en entier), ça fait deux individus différents, alors qu'en réalité c'est la même personne. Parfois il y a une petite coquille, donc l'ordinateur met deux lignes différentes... Il y a toutes ces erreurs de saisie ou de non homogénéité des données, qui font que, oui, il faut nettoyer les données. Mais comme n'importe quelles données.

[>R1]: Qui rentre ces données dans la base ?

[>R2]: Ici c'est chaque industriel qui rentre ses données. Pour le Parlement européen, vous avez les ONG comme *Regard citoyen* ou *Vote Watch*. Il y a des tas d'ONG qui militent pour la transparence et la mise à disposition des données, qui publient des tas de trucs (des wikis qui peuvent alimentés, etc.)… Il faut savoir ce que l'on fait de ça.

Annexe 13 - Exemples de tableaux thématiques comparés

Exemple 1 : Tableau thématique sur la valeur des données

	Verbatims sur la thématique « valeur des données »	
Chercheur 3	"En fait, la vraie valeur de la technologie, ça n'est pas la machine. La vraie valeur c'est le savoir-faire qui a été développé à côté. Il y a des petits trucs, qu'on n'écrit pas dans les publis, parce que ce sont des détails. Mais ce sont des détails qui font que ça fonctionne ou que ça fonctionne mal. [...] Or ce savoir-faire, et bien justement, les gens aujourd'hui le recherchent."	Certaines données sont plus précieuses que d'autres, du fait de leur coût d'acquisition. "Pour les données très sensibles comme celles-là (sensibles parce que c'est 1500€ le séquençage) on a un triple stockage".
	La préciosité des données à ses yeux le conduit à mettre en place des solutions personnelles pour les sauvegarder : "Les paranos comme moi ont un disque au labo, un disque à la maison et un disque sur eux. Comme ça, on ne risque jamais rien !".	"On va poser le brevet et on va faire gagner de l'argent à nous peut-être, mais surtout à nos tutelles derrière. Parce que c'est aussi notre responsabilité, je crois, quelque part. Je veux dire : nos tutelles nous versent tant et tant d'argent ou l'ANR nous verse tant et tant ; derrière, on rend le truc d'une autre façon."
Chercheur 35	La recherche appliquée s'inscrit dans la logique de la propriété intellectuelle, qui sous-tend exclusivité de diffusion et intérêts marchands. "Si on ne protège pas les données, on ne sera pas en mesure de les transférer à un industriel. Parce que l'industriel, ça ne l'intéresse pas si c'est public. Je pense à des entreprises qui veulent exploiter, faire de la licence sur ce qu'on fait ; eux, ils veulent des droits de licence pour être sûrs d'être en mesure d'exclure toute personne qui voudrait faire la même chose."	
Chercheur 37	"Typiquement, quand on fait une mesure isotopique, on ne va pas la mettre sur internet le lendemain. Parce qu'obtenir un rapport isotopique, ça représente des mois de travail." Données qui demandent un travail d'acquisition tellement important qu'elle ne peut pas se permettre de les mettre à la disposition de tous.	Pas de réflexion sur la valeur des données : "je n'y ai jamais vraiment pensé". Il lui est seulement possible d'estimer combien a coûté la donnée : "Je ne sais pas quelle est la valeur économique ou financière de la donnée. Par contre, je sais ce qu'elle a coûté. [...] Le projet, il a un coût. Les gens qui vont sur le terrain pour récupérer des échantillons, ça coûte. L'échantillon, quand on le ramène au laboratoire et qu'on le traite, ça coûte. Le doctorant qui va travailler dessus, ça coûte. Il y a aussi mon salaire."
Chercheur 18	"Si vous tombez sur un résultat majeur, vous allez chercher à le publier dans un journal de forte réputation et, par ce biais-là, vous allez crédibiliser de futures demandes de financement. Donc, valeur économique indirecte."	

Chercheur 2	Il utilise fréquemment le terme de « valeur ». Probablement parce qu'il s'agit d'un projet de « valorisation ».	La valeur d'un résultat ne repose pas uniquement sur sa véracité. Elle est aussi liée aux pratiques de gestion utilisées. "A partir du moment où on va vouloir valoriser ces résultats, qu'on va trouver un partenaire industriel qui s'intéresse à ces résultats, il va falloir que l'industriel soit convaincu des données que nous avons générées. Ça suppose de mettre en place un certain nombre de pratiques (notamment l'archivage des données)." "Ca nous oblige à avoir des pratiques légèrement différentes des pratiques académiques. C'est-à-dire qu'on va travailler quasiment comme une start-up. […] Ça se rationnalise. C'est de la gestion de projet, qui se professionnalise, qui n'est plus académique. On utilise les mêmes méthodes de management de projet que dans une société."
Chercheur 51	"Je pense qu'on n'utilise pas assez les données qui sont produites. C'est une réflexion qu'on a eue au niveau de la MSH. C'est aussi la raison pour laquelle on a mis en place de cette plateforme universitaire des données. Parce qu'effectivement ça coûte très cher de produire des données et que ces données sont insuffisamment exploitées par rapport à tout ce qui existe."	
Chercheur 52	"Je pense que les données qu'on produit en sciences sociales se périment moins vite peut-être que celles qu'on produit en sciences dures."	
Chercheur 22	"La valeur vient du fait que ces données sont mises dans un fichier Excel et comparées semaine après semaine, qu'on voit une évolution et qu'on peut soumettre cette évolution à un test statistique et voir si on a un résultat ou non." Une donnée n'a de valeur (scientifique) que si elle est associée et comparée à d'autres données.	"La valeur qu'on peut donner aux données, c'est toute la masse d'argent qui a été investie pour les obtenir. Par masse d'argent, je pense aux animaux, aux réactifs, aux salaires…"
Chercheur 16	"Pour moi, partager les données, parce qu'on travaillait sur le même sujet, nous permettait d'écrire plus d'articles ensemble et d'être plus productifs." L'échange de données est un atout pour augmenter le nombre de ses publications.	
Chercheur 29	"Là il n'y a pas trop de valeur économique. C'est plutôt une valeur « connaissance en recherche fondamentale ». [...] On prouve qu'on est productif et que l'argent est « bien » investi, qu'il a permis de produire des connaissances nouvelles, qui seront ensuite citées par d'autres collègues."	

Chercheur 33	"Ca, ça vaut de l'or. Ca se vend, même. Parce que ça prend des heures à faire. Avoir une image annotée comme ça entièrement, c'est ce que vendent Google, Facebook, tout ça. Ce qu'ils vendent, ce sont leurs données." A l'heure actuelle, les images annotées (tous types d'images annotées) ont une très grande valeur. Parce qu'elles sont nécessaires à l'entraînement des algorithmes, qui aujourd'hui sont utilisés dans de nombreux domaines. Mais aussi parce que les annotations ne peuvent être faites qu'à la main, par des individus. Annoter mille « objets » (pour entraîner un algorithme)	
Chercheur 41	Les bases de données naturalistes ont une valeur pour le Ministère de l'Environnement, qui aurait la possibilité de les vendre à des bureaux d'études (qui en font une exploitation commerciale). "En ce moment, il y a quand même une sensibilisation des acteurs sur ces bases de données. Ils commencent à comprendre que ça a de la valeur. Il faut savoir que ça se monnaye. C'est le réseau des associations de naturalistes Odonat-Grand Est. Ils centralisent les demandes notamment pour les données naturalistes. Les données ont une valeur assez importante, notamment dans le cadre des bureaux d'études, qui doivent faire des expertises faune-flore pour des études d'impact ou des choses comme ça. Du coup, clairement, ils vendent les données. Ils font des extractions de bases et ils les vendent."	
Chercheur 40	« Il y a un aspect un peu "précieux" de la donnée, qui est qu'on ne montre pas la donnée avant de l'avoir soi-même exploitée. C'est tout l'enjeu de la compétition scientifique. »	

Exemple 2 : Tableau thématique sur l'influence du cadre institutionnel

	Verbatims sur la thématique « influence du cadre institutionnel »
Influence du cadre institutionnel sur la diffusion des données	
Chercheur 41	Obligation de dépôt des données génétiques : "Pour les données génétiques, j'ai l'obligation de les déposer dans une base de données internationale, pour attester de la véracité et de la reproductibilité de mes analyses".
Chercheur 1	En génétique, "il y a une règle : on ne pourra pas publier, si on ne met pas les données à disposition".
Chercheur 2	En biologie structurale, le dépôt des données dans la Protein Data Bank conditionne la publication de l'article.
Chercheur 40	Les demandes de mise à disposition des données brutes sous-jacentes aux publications ont augmenté avec l'avènement du numérique comme nouveau media de la communication scientifique. Car facile, rapide et peu cher. Les commentaires des relecteurs jouent un rôle important dans la mise à disposition de telles ou telles données. "[>Chercheur]: Maintenant, avec le contenu dématérialisé et la possibilité de mettre des annexes électroniques, de plus en plus on est incité à donner les données brutes dans des tableaux, de façon complémentaire. C'est-à-dire que vous ne trouverez pas les données brutes dans le corps de l'article mais dans des liens sur lesquels vous cliquez, si vous avez la version en ligne de l'article. [>Enquêteur]: Est-ce l'éditeur qui préconise de mettre les données à disposition ? [>Chercheur]: Ca dépend. Ca peut être soit les reviewers -ceux qui sont chargés de valider ou non l'étude sur le plan scientifique-, soit l'éditeur effectivement."
Chercheur 41	"Ca nous est de plus en plus demandé par la DREAL ou le Ministère de l'Environnement. Ils demandent à ce qu'on fasse remonter les données sur la biodiversité, parce que ce sont des données acquises sur des fonds publics et qu'elles sont importantes justement pour conserver la biodiversité. J'ai oublié le nom de toutes ces demandes qui émanent du Ministère et qui sont relayées par les DREAL... On nous demande de faire remonter, au moins au niveau régional, nos données sur la biodiversité, pour établir de vraies bases de données géospatialisées de la distribution du vivant, afin de mieux le conserver et de mettre en place des mesures de conservation plus pertinentes. Donc, ce que je voulais vous dire, c'est que normalement on serait censés les faire remonter à un niveau "étatique", notamment à travers la mise en place de l'AFB -l'Agence Française pour la Biodiversité."
Chercheur 13	Le réseau des Zones Ateliers (INEE-CNRS) demande à ce que les données soient versées dans des bases de données spécifiques : "Par exemple, une partie des programmes de biologie, soutenus par l'Institut Polaire, sont dans ce qu'on appelle une Zone Atelier. C'est une structure de l'INEE du CNRS. Il y a une quinzaine de Zones Ateliers. Et, dans ce contexte-là de réseau -parce qu'on est un réseau de Zones Ateliers -, on nous demande de plus en plus de mettre des données dans des bases de données, qui sont consultables à l'échelle internationale."
Chercheur 48	L'ouverture des données est enclenchée par les mandats des financeurs du projet de recherche : "Moi j'obtiens un financement de la CAF, à condition de livrer aux chercheurs dans Quetelet et dans la PUD mon enquête MFV. On a des financeurs mais on est obligé de la donner. Parce que c'est de la statistique publique. Et elle a été labellisée par le comité du label, par la CNIL, etc."
Influence du cadre institutionnel sur la manière de gérer les données	
Chercheur 45	"Tous les laboratoires de recherche, notamment qui sont sous la tutelle du CNRS, doivent respecter un certain nombre de protocoles de sécurité en lien avec ces informations. La RGPD est passée par là pour renforcer les choses. Mais, nous, ça fait quand même une bonne quinzaine d'années qu'on a une culture de la protection des données. Via le CNRS mais aussi l'ensemble des réglementations qui existent, comme ce qui est en lien avec la CNIL."

Chercheur 51	La manipulation de données personnelles est soumise à des règles dictées par la CNIL. Ces règles sont souvent peu pratiques pour les chercheurs. Exemple : anonymiser les entretiens dès la retranscription rend l'analyse plus compliquée (difficultés pour s'y retrouver). « On doit faire une déclaration à la CNIL, en décrivant ce que l'on fait et en précisant qu'on le fait à des fins de recherche, et non dans un but commercial par exemple. En toute rigueur, il faudrait que chacune des personnes, sur lesquelles on travaille, soit informée qu'on travaille sur elle. Mais comme ce sont des gens qui sont plus ou moins des personnages publics et que les informations qu'on obtient proviennent soit de leur site internet, soit des déclarations qu'ils ont faites (déclarations qui sont elles-mêmes publiques), les données ne sont pas top secrètes. » « Là où on est susceptible de nous surveiller, c'est qu'il ne faut pas qu'on ait de fichiers nominatifs. Par exemple, si on a des entretiens, il ne faudrait pas que dans nos ordinateurs on ait le nom des gens qui ont dit ceci. Il faudrait qu'on anonymise tout de suite. Mais, quand on publie, on anonymise de toute façon toujours. »
Chercheur 16	"On peut garder les données pendant cinq à dix ans, si d'autres travaux sont ré-effectués dessus. C'est une période qui est renouvelable, c'est-à-dire que, si un nouveau travail est effectué sur ces données, on peut prolonger à nouveau la conservation." Les données de la cohorte peuvent être conservées tant qu'il y a des recherches menées dessus. D'où l'intérêt de toujours renouveler les projets de recherche, pour pouvoir conserver ces données.
Chercheur 52	"Il n'y a que pour l'EDP, où on a eu effectivement des contraintes." Les données de l'Echantillon Démographique Permanent (données sensibles) sont soumises à une obligation de « regroupement » au moment de la publication des résultats.

Sommaire détaillé

Remerciements.....3

Résumé.....4

Abstract.....5

Sommaire.....6

Liste des tableaux.....8

Liste des figures.....9

Introduction.....11

Première partie - Qu'est-ce qu'une donnée de la recherche ?.....23

1. Des tentatives de définition.....25

1.1. Quand a-t-on commencé à définir les données de recherche ?.....25

1.2. Définitions par l'énumération.....26

1.3. Définitions sous forme de typologies.....28

1.3.1. Typologie selon l'approche méthodologique.....28

1.3.2. Typologie selon le niveau de traitement des données.....31

1.3.3. Typologie selon l'origine des données.....31

2. Vers une non définition.....34

2.1. Le concept général de « donnée » : un concept fuyant.....34

2.1.1. Étymologie.....34

2.1.2. Donnée, Information, Connaissance.....35

2.1.3. « Information as thing ».....37

2.2. Les données de recherche se définissent relativement à un contexte épistémologique 39

2.2.1. Quand une entité devient-elle une donnée ?.....39

2.2.2. Les données comme preuves.....41

2.2.3. Une définition multidimensionnelle.....43

3. Conclusion.....45

Deuxième partie - Les politiques publiques de gestion et d'ouverture des données de la recherche.....47

1. Mouvements à l'origine des politiques de gestion et d'ouverture des données de la recherche.....50

1.1. Déclaration de Berlin (2003).....50

1.2. Recommandations de l'OCDE (2004).....52

1.2.1. Déclaration sur l'accès aux données de la recherche publique.....52

1.2.2. Une logique d'innovation.....54

1.2.2.1. Le concept d'économie de la connaissance.....54

1.2.2.2. Rôle de la science dans l'économie de la connaissance.....55

2. Influence de l'Open Data.....56

2.1. Philosophies du mouvement Open Data.....56

2.1.1. Transparence et plus grande participation des citoyens.....57

2.1.2. Innovation et création de valeur.....57

2.2. L'Open Data devenu objet politique.....60

2.2.1. Les États-Unis : précurseurs des politiques d'Open Data.....60

2.2.2. L'Open Data dans la politique : une démarche consensuelle mais des objectifs originels affaiblis.....61

2.3. Problématiques de réutilisation des données ouvertes.....63

3. Politiques et initiatives de l'Union européenne.....65

3.1. La vision d'une science au service de l'économie.....65

3.1.1. Création d'un Espace Européen de la Recherche (2000).....65

3.1.2. Renforcement de l'EER par le libre accès aux résultats scientifiques (2012).…66

3.1.2.1. Première prise de position en faveur du libre accès aux résultats scientifiques (2007).....67

3.1.2.2. Les raisons d'une politique de libre accès aux résultats de la recherche.….68

3.1.2.3. La mise en place de moyens d'action (2012).....70

3.1.2.4. Des recommandations aux États membres (2012).....72

3.2. Des programmes de financement de la recherche de plus en plus prescriptifs.....73

3.2.1. La Commission européenne, financeur de la recherche.....73

3.2.2. L'initiative première du Conseil européen de la recherche.....73

3.2.3. La politique d'ouverture des données de la recherche dans le programme Horizon 2020.....74
3.2.3.1. Le dépôt des données en ligne.....74
3.2.3.2. La rédaction de plans de gestion de données.....76
3.2.3.3. Des données compatibles avec les principes FAIR.....76
3.2.4. Vers une politique renforcée dans le prochain programme Horizon Europe ?…77
3.3. Des infrastructures de recherche mises au service de l'Open Science.....78
3.3.1. Les infrastructures de recherche financées par la Commission européenne.....78
3.3.1.1. La question des données dans les infrastructures de recherche.....78
3.3.1.2. La politique européenne de financement des infrastructures de recherche..80
3.3.2. Le cloud européen pour la science ouverte (EOSC).....81
3.3.2.1. Cadre d'émergence de l'EOSC: le passage de l'économie mondiale au numérique.....81
3.3.2.2. Structure du cloud européen pour la science ouverte.....82
3.3.2.3. Engouement autour du projet.....84
3.3.2.4. Quels utilisateurs pour l'EOSC ?.....85
4. Politiques et initiatives de l'État français.....88
4.1. Législation relative à la diffusion des données scientifiques.....88
4.1.1. A l'origine : la loi CADA (1978 / 2005).....88
4.1.2. La loi Valter et la loi pour une République numérique (2015 / 2016).....88
4.2. Initiatives du Ministère de la Recherche en matière de données de recherche.....91
4.2.1. La Bibliothèque Scientifique Numérique (2009-2017) et son successeur, le Comité pour la Science Ouverte (2018-).....91
4.2.2. Un plan national pour la science ouverte (2018).....93
5. Conclusion.....96

Troisième partie - Les services d'appui à la gestion et au partage des données de recherche.....101
1. Terrain et méthodologie.....105
1.1. Un recensement des services de données par la Bibliothèque Scientifique Numérique 105

1.2. Phase 1 : une première cartographie....108
1.2.1. Identification de l'offre de services....108
1.2.2. Élaboration d'une typologie des services....108
1.2.3. Conception d'une grille d'analyse pour chaque type de service....110
1.2.4. Recensement et analyse des services....110
1.3. Phase 2 : la conception d'un répertoire en ligne, Cat OPIDoR....110
1.3.1. Objectifs du répertoire....111
1.3.2. Caractéristiques du répertoire....111
2. Le paysage national des services de gestion et d'ouverture des données : constats 114
2.1. Des services relativement récents....115
2.2. Des services extrêmement divers....116
2.2.1. Divers par leur nature....116
2.2.2. Divers par le périmètre de leur public cible....118
2.2.3. Divers dans leur utilisation....126
2.3. Organisation des services....129
2.3.1. Structures d'appartenance....130
2.3.2. Financement....132
2.3.3. Ressources humaines....133
2.4. Spécificités des services fournis par des professionnels de l'information scientifique et technique....134
2.4.1. A l'origine de l'offre de services des professionnels de l'IST : la défense du libre accès aux résultats de la recherche et le repositionnement de la profession....135
2.4.2. Nature de l'offre de services conçue par les professionnels de l'IST....136
2.4.3. Quelles tendances : des services techniques ou informatifs ?....142
2.5. Evolution des services de gestion et d'ouverture entre 2016 et 2019....144
2.5.1. Evolution des entrepôts et annuaires de données....144
2.5.2. Evolution des autres types de services....149
3. Conclusion....151

Quatrième partie - Les données dans les pratiques de recherche....153
1. Terrain et méthodologie....156
1.1. Une enquête sous forme d'entretiens semi-directifs....156

1.2. Méthode d'échantillonnage.....156
1.2.1. L'Université de Strasbourg pour terrain d'enquête.....156
1.2.2. Populations sondées.....157
1.3. Conduite des entretiens.....160
1.4. Méthode d'analyse des entretiens.....161
2. Résultats et discussion.....162
2.1. Résultats de participation à l'enquête.....162
2.2. Terminologie : le sens du terme « donnée » dans le discours des chercheurs.....164
2.3. Valeur des données dans un écosystème centré sur la publication scientifique.....170
2.3.1. Publier pour gagner en « crédibilité ».....170
2.3.1.1. Le cycle de crédibilité.....170
2.3.1.2. Un mécanisme plus ou moins perceptible selon les domaines de recherche 171
2.3.1.3. Faible valeur d'échange des données.....173
2.3.2. La valeur indirecte des données : une valeur d'usage.....175
2.3.2.1. Différentes perceptions de leur valeur.....175
2.3.2.2. Une valeur pour soi.....176
2.3.2.3. Variation de la valeur de la donnée.....176
2.4. Traitement des données (collecte, analyse, interprétation) : des méthodes et outils variés et évolutifs.....178
2.5. Gestion des données (stockage, curation, conservation...).....180
2.5.1. Premier facteur de variation : les ressources consacrées à la gestion des données dans le laboratoire ou le projet de recherche.....181
2.5.1.1. Inégalité des laboratoires.....181
2.5.1.2. Nouvelles logiques de financement de la recherche.....186
2.5.2. Deuxième facteur de variation : le degré de sensibilité des données.....187
2.5.2.1. Gestion des données personnelles.....188
2.5.2.2. Gestion des données issues de la recherche appliquée.....197
2.6. Ouverture des données.....201
2.6.1. Le discours plutôt favorable des chercheurs.....201

2.6.2. Nature du partage dans la culture scientifique : un échange censé contribuer à la renommée du chercheur....206
3. Conclusion....210

Cinquième partie - Adéquation entre les services de données et les pratiques des chercheurs....213

1. Enquête sur l'utilisation de services de données par les chercheurs....216
1.1. Méthodologie....216
1.2. Avis des chercheurs sur Cat OPIDoR....218
1.3. Un recours peu fréquent aux services de données....221
1.4. Deux ordres de services....223
2. Utilisation par les chercheurs des services nés sous l'influence du mouvement d'ouverture des données....224
2.1. Une méconnaissance de ces services....224
2.2. Un rapport privilégié à la publication....225
3. Utilisation des services disciplinaires par les chercheurs....228
3.1. Exemples de services disciplinaires....229
3.1.1. Des services d'acquisition de données....229
3.1.1.1. Plateformes technologiques....229
3.1.1.2. Observatoires astronomiques....230
3.1.1.3. Stations de recherche....232
3.1.1.4. Cohortes de populations....232
3.1.2. Des services de réutilisation de données....234
3.1.2.1. Banques de données d'enquêtes quantitatives en sciences sociales....234
3.1.2.2. Banques de données omiques....235
3.1.2.3. Archives des télescopes....236
3.1.2.4. Points communs des services de réutilisation....236
3.2. Caractéristiques des services disciplinaires....238
3.2.1. Une majorité de services internationaux....238
3.2.2. Des services publics et privés....238
3.2.3. Le recours à des services, mais pas seulement....239

3.2.4. Vers une partition entre producteurs et utilisateurs de données ?.....241
3.3. Quels services complémentaires pour les données de la recherche ?.....242

Conclusion.....249

Bibliographie.....263

Sources.....281

Sitographie & Acronymes.....293

Annexes.....307

Annexe 1 - Offre de poste pour la réalisation d'une cartographie nationale des services de données scientifiques.....309
Annexe 2 - Équipe projet de Cat OPIDoR (octobre 2016 – septembre 2017).....312
Annexe 3 - Exemples de grilles d'analyse de deux types de services de données.....313
Exemple 1 : Grille d'analyse des services de type accompagnement.....313
Exemple 2 : Grille d'analyse des services de type entrepôt de données.....314
Annexe 4 - Cat OPIDoR : Page d'accueil.....316
Annexe 5 - Cat OPIDoR : Page de résultats du domaine Sciences & Technologies (affichage sous forme de tableau).....317
Annexe 6 - Cat OPIDoR : Page de résultats du sous-domaine Sciences du système Terre (affichage sous forme d'index).....318
Annexe 7 - Cat OPIDoR : Champs descriptifs d'un service.....319
Annexe 8 – Tableau analytique des 44 services de gestion et d'ouverture des données étudiés dans la 3ème partie.....320
Annexe 9 - Formulaire de consentement éclairé.....342
Annexe 10 - Guide d'entretien (1er panel, novembre 2017 – octobre 2018).....344
Annexe 11 - Guide d'entretien (2nd panel, mars – mai 2019).....345
Annexe 12 – Exemples d'entretiens retranscrits.....346
Entretien avec le chercheur 2 (chimie).....346
Entretien avec le chercheur 16 (éthologie).....371
Entretien avec le chercheur 18 (neurosciences).....388

Entretien avec le chercheur 26 (astronomie).....408
Entretien avec le chercheur 37 (géologie).....425
Entretien avec le chercheur 43 (géographie).....446
Entretien avec le chercheur 51 (sciences politiques).....462
Annexe 13 - Exemples de tableaux thématiques comparés.....476
Exemple 1 : Tableau thématique sur la valeur des données.....476
Exemple 2 : Tableau thématique sur l'influence du cadre institutionnel.....479

Sommaire détaillé.....481

le cnam

Violaine REBOUILLAT

Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs

Résumé : Cette thèse s'intéresse aux données de la recherche, dans un contexte d'incitation croissante à leur ouverture. Les données de la recherche sont des informations collectées par les scientifiques dans la perspective d'être utilisées comme preuves d'une théorie scientifique. Il s'agit d'une notion complexe à définir, car contextuelle. Depuis les années 2000, le libre accès aux données occupe une place de plus en plus stratégique dans les politiques de recherche. Ces enjeux ont été relayés par des professions intermédiaires, qui ont développé des services dédiés, destinés à accompagner les chercheurs dans l'application des recommandations de gestion et d'ouverture. La thèse interroge le lien entre philosophie de l'ouverture et pratiques de recherche. Quelles formes de gestion et de partage des données existent dans les communautés de recherche et par quoi sont-elles motivées ? Quelle place les chercheurs accordent-ils à l'offre de services issue des politiques de gestion et d'ouverture des données ?

Pour tenter d'y répondre, 57 entretiens ont été réalisés avec des chercheurs de l'Université de Strasbourg dans différentes disciplines. L'enquête révèle une très grande variété de pratiques de gestion et de partage de données. Un des points mis en évidence est que, dans la logique scientifique, le partage des données répond un besoin. Il fait partie intégrante de la stratégie du chercheur, dont l'objectif est avant tout de préserver ses intérêts professionnels. Les données s'inscrivent donc dans un cycle de crédibilité, qui leur confère à la fois une valeur d'usage (pour la production de nouvelles publications) et une valeur d'échange (en tant que monnaie d'échange dans le cadre de collaborations avec des partenaires). L'enquête montre également que les services développés dans un contexte d'ouverture des données correspondent pour une faible partie à ceux qu'utilisent les chercheurs. L'une des hypothèses émises est que l'offre de services arrive trop tôt pour rencontrer les besoins des chercheurs. L'évaluation et la reconnaissance des activités scientifiques étant principalement fondées sur la publication d'articles et d'ouvrages, la gestion et l'ouverture des données ne sont pas considérées comme prioritaires par les chercheurs. La seconde hypothèse avancée est que les services d'ouverture des données sont proposés par des acteurs relativement éloignés des communautés de recherche. Les chercheurs sont davantage influencés par des réseaux spécifiques à leurs champs de recherche (revues, infrastructures…). Ces résultats invitent donc à reconsidérer la question de la médiation dans l'ouverture des données scientifiques.

Mots-clefs : Pratiques de recherche, Communication scientifique, Partage des données, Politiques de science ouverte, Services de données, Professionnels de l'information scientifique et technique

Abstract: The thesis investigates research data, as there is a growing demand for opening them. Research data are information that is collected by scientists in order to be used as evidence for theories. It is a complex, contextual notion. Since the 2000s, open access to scientific data has become a strategic axis of research policies. These policies has been relayed by third actors, who developed services dedicated to support researchers with data management sharing. The thesis questions the relationship between the ideology of openness and the research practices. Which kinds of data management and sharing practices already exist in research communities? What drives them? Do scientists rely on research data services?

Fifty-seven interviews were conducted with researchers from the University of Strasbourg in many disciplines. The survey identifies a myriad of different data management and sharing practices. It appears that data sharing is embedded in the researcher's strategy: his main goal is to protect his professional interests. Thus, research data are part of a credibility cycle, in which they get both use value (for new publications) and exchange value (as they are traded for other valuable resources). The survey also shows that researchers rarely use the services developed in a context of openness. Two explanations can be put forward. (1) The service offer comes too early to reach researchers' needs. Currently, data management and sharing are not within researchers' priorities. The priority is publishing, which is defined as source of reward and recognition of the scientific activities. (2) Data management services are offered by actors outside the research communities. But scientists seem to be more influenced by internal networks, close to their research topics (like journals, infrastructures…). These results prompt us to reconsider the mediation between scientific communities and open research data policies.

Keywords: Research practices, Scientific communication, Data Sharing, Open science policies, Data services, Scientific and technical information staff