Construction(s) et contradictions des données de recherche en SHS

Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Alexandre Serres

Construction(s) et contradictions des données de recherche en SHS

Construction(s) and contradictions of research data in the Humanities and Social Sciences

Marie-Laure Malingre[1], Morgane Mignon[2], Cécile Pierre[3], Alexandre Serres[1]

[1] URFIST (Unité Régionale de Formation à l'Information Scientifique et Technique) de Rennes, France
[2] MSHB (Maison des Sciences de l'Homme en Bretagne), Rennes, France
[3] SCD (Service Commun de Documentation), Université Rennes 2, France

RÉSUMÉ. La structuration et le partage des données s'imposent depuis cinq ans au monde de la recherche, à travers des injonctions politiques (de Horizon 2020 au Plan national pour la science ouverte). L'analyse de l'enquête menée en 2017 auprès des chercheurs de l'université Rennes 2 sur leurs pratiques, représentations et attentes en matière de données conduit à interroger le terme lui-même. Variable et complexe, contrairement à ce que suggère le mot « donnée », la notion ne va pas de soi. L'article s'efforcera de montrer qu'elle fait l'objet d'une triple construction, épistémologique, intellectuelle et politique, dans les discours des chercheurs et des acteurs institutionnels, en tension avec les pratiques constatées sur le terrain.

ABSTRACT. In the last decade, political injunctions to curate and share research data have increased significantly. A survey conducted in 2017 in Rennes 2, a french Humanities and Social Sciences university, enabled us to question the habits and representations of the researchers in this matter, but also the term of "data" itself. Contrary to the idea that data are given, which is implicit in the french word "données", the notion of "data" is far from being self-evident and actually proves to be complex and multifaceted. This article aims at showing that a triple redefinition and construction of research data is at stake in the discourses of researchers and institutional stakeholders: it operates at epistemological, intellectual and political levels. These concepts of data conflict with existing practices in the field.

MOTS-CLÉS. données de recherche, SHS, pratiques des chercheurs, enquête, France, politique des données, partage des données.

KEYWORDS. research data, humanities, social sciences, researcher practices, survey, France, data policy, datasharing.

Introduction

Avec l'ouverture des publications, prônée depuis plusieurs années mais qui peine encore à se généraliser, la structuration et le partage des données s'imposent depuis cinq ans au monde de la recherche. Rassemblées sous le deuxième axe du Plan national pour la science ouverte exposé en juillet 2018, ces thématiques se traduisent dans le plan d'action 2018 de l'Agence Nationale de la Recherche par l'obligation de soumettre des plans de gestion de données dans toute réponse aux appels à projets de l'ANR.

Comment cet objet complexe qu'est la donnée est-il perçu par les chercheurs en SHS ? Les discours autour des données de recherche sont-ils connus et adoptés au niveau local par les individus et dans les laboratoires ? Quelles sont les caractéristiques des pratiques quotidiennes relatives aux données en SHS ? Quels freins lever pour les faire évoluer et se conformer aux injonctions de gestion et de partage ? Autant d'interrogations pouvant orienter et conditionner le succès d'une politique d'établissement.

L'URFIST de Rennes, la Maison des Sciences de l'Homme en Bretagne et le SCD de l'université Rennes 2 ont mené en 2017 une enquête auprès des chercheurs du campus portant sur leurs pratiques en matière de données. Disponibles sur HAL [SER et al. 17], les résultats de cette enquête ont fourni un point de départ à la réflexion développée tout au long de cet article. Écrit à plusieurs mains, il croise les regards de ses auteurs sur un phénomène multiforme. Favoriser l'ouverture et le partage, c'est prendre en compte la réalité du terrain, les différentes données manipulées et les facteurs positifs ou les freins qui opèrent des tensions contradictoires dans les pratiques et dans l'imaginaire des chercheurs.

L'analyse de l'enquête, à la fois quantitative et qualitative, conduit à interroger la notion même de donnée. En nous appuyant sur les résultats statistiques et les *verbatim* de chercheurs lors des entretiens, nous nous efforcerons d'appréhender la variété et la complexité du terme et argumenterons que, contrairement à ce que suggère le mot « donnée », la notion ne va pas de soi, et qu'elle fait l'objet d'une triple construction, épistémologique, intellectuelle et politique, dans les discours des chercheurs et des acteurs institutionnels.

1. Méthodologie

1.1. *Méthodologie et conclusions de l'enquête*

Pour cerner les usages et les représentations de chercheurs sur la gestion de leurs données, l'enquête qui a servi de socle à cet article s'est engagée dans une double démarche, à la fois quantitative et qualitative. Le questionnaire statistique, qui reprenait en partie dans sa trame l'enquête menée à l'Université de Lille 3 en 2015 [PRO 15], comportait 32 questions, dont 13 obligatoires, avec peu de questions ouvertes. Il était structuré autour de plusieurs thématiques : outre le profil des répondants, étaient abordés la typologie des données, les pratiques de stockage et d'archivage, les pratiques de partage et de diffusion, ainsi que les besoins et attentes liés à l'usage des données en recherche. Le questionnaire a été testé auprès d'un panel de chercheurs, il a ensuite été diffusé en ligne pendant deux mois, en direction des chercheurs, enseignants-chercheurs et ingénieurs de recherche, soit 496 personnes issues de 19 laboratoires. En prolongement, 21 entretiens individuels semi-directifs ont été menés auprès de volontaires afin de préciser les réponses obtenues, en abordant le cas échéant de nouvelles questions. Leur transcription et leur analyse ont été réalisées avec le logiciel Sonal[1].

Au final, le matériau recueilli a permis de tirer de cette enquête trois principales leçons : premièrement, la complexité et la spécificité des données de recherche, variant avec la communauté de recherche et la méthode de recueil et de production ; ensuite, la transversalité et la place cruciale que prennent les questions de la sécurisation, du stockage et de l'archivage des données, face à des pratiques qui restent individuelles et peu systématisées ; enfin, l'écart constaté entre les déclarations, les représentations et les pratiques des chercheurs sur la question du partage et de l'ouverture des données. En témoignent des interrogations ou des inquiétudes, une hétérogénéité de pratiques selon le champ disciplinaire et l'écosystème de recherche.

Confronter les résultats de notre enquête avec ceux de Lille 3, également spécialisée en SHS, permet de dégager à la fois des convergences fortes et des spécificités[2]. La première spécificité est méthodologique, car l'enquête rennaise porte sur un échantillon plus restreint (enseignants-chercheurs uniquement), et comprend aussi un volet qualitatif (réalisé par Joachim Schöpfel *a posteriori* en 2018 [SCH 18]). Le contexte des répondants est comparable, même si le taux de réponse de cette enquête est plus élevé (28,8 % de répondants à Rennes contre 15 % à Lille). On retrouve dans les deux cas une plus forte mobilisation des professeurs et les mêmes communautés disciplinaires majoritaires (linguistes et psychologues).

Une large convergence de réponses peut être observée sur plusieurs points, consolidant une certaine homogénéité des pratiques en SHS à l'échelle macro :

– hiérarchie des catégories de données sources et de données produites[3] les plus utilisées (prédominance du texte, suivi par les enquêtes et entretiens, puis observations) ;

[1] Disponible sur http://www.sonal-info.com/

[2] Nous n'avons pas retenu de comparaison avec des enquêtes étrangères, faute de temps et en raison de la divergence des structures universitaires et des domaines disciplinaires. Toutefois, un éclairage indirect peut être trouvé à travers l'analyse croisée conduite par les auteurs de l'enquête de Lille 3 avec les enquêtes de Berlin, Strasbourg et Liber [PRO 15, p. 26-27].

[3] Cette typologie des catégories de données oppose les *données sources*, base du travail du chercheur, et les *données produites* ou *données résultats*, issues de son activité de recherche. Elle est reprise de l'enquête de Lille 3 [PRO 15, p 14], qui s'appuie elle-même sur le travail de l'université Humboldt de Berlin.

– prédominance massive du stockage personnel et utilisation de supports multiples ;

– des pratiques de partage encore largement minoritaires : selon 54 % des répondants, personne d'autre ne peut accéder à leurs données ;

– des analogies dans les besoins d'archivage sécurisé des données, ainsi que sur les services de conseil : les conseils techniques et juridiques sont plébiscités dans les deux cas. En revanche, dans l'enquête lilloise, les conseils pour la gestion des données en général arrivent en tête et les conseils relatifs à la publication et la citation des données de recherche recueillent un assentiment nettement plus large.

Toutefois on observe des évolutions entre les deux enquêtes, réalisées à deux ans d'intervalle, tant sur les pratiques de stockage que sur le partage :

– le pourcentage de répondants estimant le volume de leurs données supérieur à un To est plus élevé à Rennes 2 (11 %) qu'à Lille 3 (6 %). Le recours au stockage sur le *cloud* semble plus répandu (27 % contre 19 %) ;

– l'idée du partage se diffuse aussi (+ 15 % d'opinions favorables) et se vérifie en pratique : le pourcentage de données non accessibles est en baisse (- 10 %) et les pratiques de partage en progression (+ 7 %). Mais un noyau comparable de chercheurs réticents demeure (10 % contre 12 % à Lille). Le taux de partage reste inférieur aux autres enquêtes européennes.

1.2. *Méthodologie de l'article*

Pour présenter la méthodologie de cet article, il convient d'abord de rappeler l'origine de celui-ci : il est le résultat d'une demande, à la suite d'une communication réalisée dans le cadre de la journée d'étude ADOC « Variété des données en SHS », qui s'est déroulée en mai 2018 à Nantes. Cette communication présentait les résultats de l'enquête menée dans notre université de SHS sur les pratiques des chercheurs en matière de données. Passer d'un rapport d'enquête et d'une communication de journée d'étude à un article de recherche impliquait pour les auteurs un double effort : d'une part, un effort de synthèse et d'originalité dans la présentation des résultats pour ne pas répéter le rapport d'enquête ou la communication ; d'autre part, une volonté d'approfondissement, voire de dépassement, de nos problématiques initiales. En cherchant à répondre au mieux au thème de la variété des données en SHS, nous avons alors choisi de privilégier « l'approche constructiviste » des données, en mettant en exergue les trois axes de la construction des données de recherche en SHS : la construction épistémologique, intellectuelle et politique. Dès lors, la présentation des résultats de l'enquête pouvait se réorganiser autour de ces axes, et surtout s'enrichir d'un nouveau travail de réflexion et d'analyse. Celui-ci s'est traduit de trois manières :

– d'abord, un effort de réflexion théorique sur les aspects sémantiques et épistémologiques des notions mêmes de donnée et de donnée de recherche ; il s'agissait de rappeler, aussi synthétiquement que possible, à la fois la complexité théorique de cette notion, ses articulations avec la notion voisine d'information, et surtout le caractère construit des données de recherche, au plan épistémologique ;

– ensuite, il convenait, en les respectant fidèlement, de réorganiser les résultats de notre enquête autour de l'axe de la construction intellectuelle, scientifique, des données de recherche ; nous avons voulu mettre l'accent sur l'une des principales leçons de cette observation des pratiques : le rapport affectif, personnel, complexe, des chercheurs de SHS avec leurs données, qu'ils construisent d'un bout à l'autre du cycle de recherche ;

– enfin, il fallait souligner aussi le caractère construit des « discours sur » les données, tenter de retracer leur généalogie, identifier leurs lignes de force : pour ce faire, et dans les contraintes d'un article, il s'agissait de réaliser une rapide analyse de corpus des principaux textes politiques. Pour cette partie entièrement nouvelle (par rapport à l'enquête), le principal objectif était d'identifier la diversité des acteurs et des discours politiques sur l'ouverture des données, de résumer leur progressive convergence depuis quelques années et d'en repérer les principaux arguments.

Quatre parties, d'inégale importance, structurent donc l'article : une évocation de la variété et de la complexité des données de recherche, à travers les témoignages des chercheurs interrogés ; un rappel de la dimension épistémologique de la construction des données ; les témoignages et les observations de la construction intellectuelle, scientifique, des données de recherche par les chercheurs, à partir des résultats de l'enquête ; la construction des discours politiques sur le partage des données, à travers une analyse sélective de corpus de textes.

En fil rouge sont évoquées les pratiques réelles des chercheurs en SHS, confrontés à ces multiples complexités.

2. La donnée variable et insaisissable

L'enquête menée sur le terrain a permis de mettre en évidence la grande variété des données manipulées dans le contexte des SHS, tout en relevant la diversité des pratiques qui y sont adossées et les termes utilisés par les chercheurs pour désigner ces données.

2.1. *Variété et complexité des données et des pratiques*

Cette variété concerne à la fois les données collectées et les données produites [SER et al. 17, p. 26]. Dans l'enquête, les chercheurs ont été interrogés sur les catégories de données sources qu'ils manipulaient. Sur les neuf grands types de données proposés dans la typologie, huit catégories ont été choisies par plus de 20 % des répondants. 83 % des participants utilisent plus d'un type de données sources, et 42 % quatre catégories ou plus.

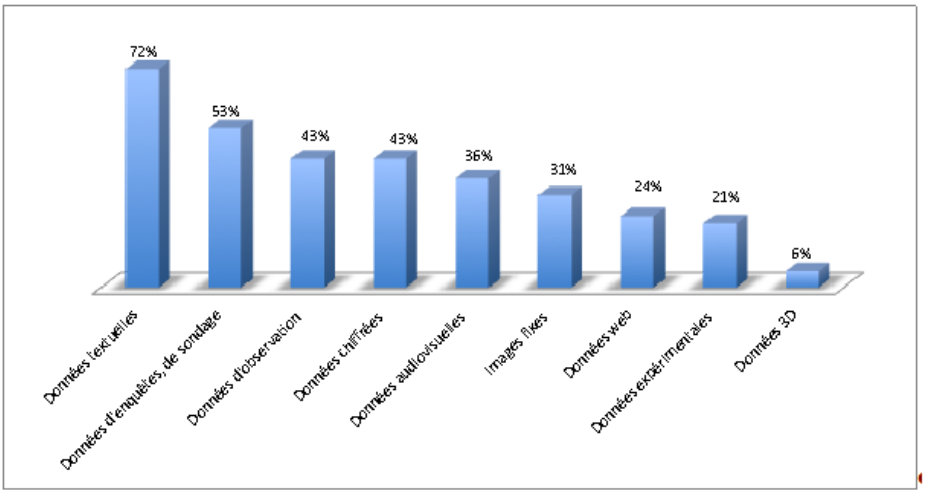

Figure 1. *Catégories des données sources (question 6)*

Par ailleurs, la même hiérarchie joue pour les données résultats : on y trouve aussi des types de données variés et complexes, dans des proportions significatives (production de bases de données et/ou programmes et applications, données spécifiques, modèles multidimensionnels) [SER et al. 17, p. 32].

Comment analyser cette variété ? Il y a tout d'abord un facteur individuel lié à chaque recherche. Pour un même chercheur, les types de données peuvent parfois varier d'un projet de recherche à l'autre, comme le pointent certains entretiens. *A contrario*, d'autres témoignent, en histoire par exemple, de la tentation de se constituer un terrain pérenne [SER et al. 17, p. 37].

L'aspect disciplinaire joue également un rôle notable. Le croisement des réponses par laboratoire ou par discipline CNU conduit à nuancer la hiérarchie globale et à repérer des tendances partagées (les répondants des laboratoires de sciences sociales et humaines, à l'exception de l'histoire, privilégient par exemple les données d'observation, d'enquêtes ou chiffrées et non pas le texte). Il confirme aussi la diversité des données manipulées ou créées à l'intérieur d'une même entité : les membres de 13

laboratoires sur 19 ont cité au moins une fois sept types de données parmi les neuf proposés [SER et al. 17, p. 27 et 31].

D'autres lignes de partage sont suggérées par les remarques de chercheurs interrogés, comme les convergences méthodologiques : « il y a des grandes catégories de méthodes, c'est quelque chose de structurant » [SER et al. 17, p. 35], y compris entre les disciplines.

Si l'on élargit le prisme aux pratiques, au-delà de la typologie des données, le support matériel a un impact sur cette variété. 45 % des données initialement collectées ou créées le sont en format non numérique, mais presque la moitié des données récoltées sont numérisées *a posteriori*. Des campagnes d'archives qui se faisaient initialement sous forme de transcriptions papier peuvent donner lieu aussi à des photos numériques. Par ailleurs, certains corpus, qui demandaient précédemment une acquisition et une construction coûteuses en temps et/ou en argent, sont désormais disponibles librement sur internet (bibliothèques numériques, données géographiques, photographies satellite, etc.).

Cela a plusieurs conséquences, avec tout d'abord le foisonnement des données accessibles [SER et al. 17, p. 40] comme des données stockées par les chercheurs : 1/3 estiment le volume de leurs données à plus de 100 Go, et 11 % à plus de 1 To, proportion en hausse par rapport à celle mesurée à Lille 3 deux ans plus tôt [SER et al. 17, p. 43]. Le deuxième élément à noter est une variété de pratiques, avec un effet générationnel entre « des collègues plus âgés qui travaillent encore à l'ancienne » et « ceux qui arrivent [qui] n'auront même plus besoin de ces services » [SER et al. 17, p. 41]. Enfin, la problématique même du partage, sous-jacente au questionnaire et aux discours politiques et scientifiques autour des données de recherche, est rendue possible par la prédominance aujourd'hui du support numérique et d'internet.

2.2. *Multiplicité des termes utilisés pour qualifier les données*

L'enquête a montré la complexité du paysage des données dans une université SHS. Mais au fond, de quoi parle-t-on ? Le *Petit Larousse* et le *Petit Robert* donnent plusieurs acceptions différentes du terme, parfois opposées : « élément admis ou connu servant de base à un raisonnement », « résultat d'observations », « l'hypothèse figurant dans l'énoncé d'un problème en mathématiques », « la représentation conventionnelle d'une information en vue de son traitement en informatique ».

Force est de constater lors des entretiens que, lorsqu'ils parlent de la nature des données collectées ou produites, très peu de chercheurs utilisent spontanément le terme « données », sauf en écho à la question posée par les enquêteurs. Il s'agit d'un terme exogène gênant, au point qu'il génère dans un cas un besoin de précision et l'utilisation d'un synonyme plus approprié (« on a toute une série de données, d'enregistrements »). Le terme apparaît flou et quelques interrogations s'expriment sur la frontière avec les publications (« je ne sais pas si cela entre dans les données de recherche, toute la politique de diffusion scientifique et des articles mis en ligne », « est-ce que les données ce sont les données brutes, ou est-ce que les résultats on les appelle aussi données ? »).

La donnée s'invite toutefois dans certains discours de chercheurs pour évoquer les données numériques récupérées sur internet (réseaux sociaux), lorsqu'il est question de gestion (stockage et perte, partage, données sensibles, données payantes ou gratuites) ou de structuration (bases de données), donc d'informatique et d'une vision plus globale ou analytique.

Quels termes utilisent alors les chercheurs pour parler de leurs données et que traduisent-ils ? On y trouve tout d'abord le reflet de la diversité observée dans les types de données, avec une multitude de mots, variables selon les disciplines. Pour les données sources, les registres les plus utilisés sont des termes disciplinaires spécifiques (sources, pour les historiens) ; des termes génériques relatifs à un type de données (images, photos, textes, documents), évoquant un ensemble (corpus, listes) et, dans une moindre mesure, liés à des méthodes de collecte (enregistrements, captures d'écran, relevés GPS, enquêtes, entretiens, etc.). La dimension humaine est sensible, quoique moins fréquente, avec la notion d'interaction (témoignages, biographies langagières, correspondances, traces), y compris dans une

dimension impalpable et très subjective (des échanges non méthodiques « qui peuvent avoir l'air de rien du tout », le « bouche à oreille »).

Pour les données produites, on peut noter l'importance de la référence au travail et au traitement (transcriptions, analyses, descriptions, photos retraduites, encodages, versions). Le format informatique est aussi cité (PDF, CSV, Excel, Word, Matlab, bases de données, etc.). On retrouve enfin, comme pour les données sources, les termes génériques liés à un type de données (textes, documents, photos).

Le choix des mots privilégiés par les chercheurs véhicule donc clairement l'interaction du locuteur avec la donnée et l'action exercée (collecte, gestion ou traitement). On peut s'interroger sur ce que véhicule le mot « donnée » lui-même et ses attendus implicites.

3. La donnée n'est jamais donnée : une notion qui ne va pas de soi

La grande variété des termes utilisés par les chercheurs pour parler de leurs données ne témoigne pas seulement de la diversité de celles-ci. La difficulté à nommer et à définir ce qui constitue le matériau quotidien du chercheur est aussi l'illustration d'une triple complexité : la complexité théorique de la notion même de donnée et de son articulation avec celle d'information ; la complexité épistémologique, propre à la place des données dans le cycle de la recherche ; et la complexité sémantique, définitionnelle des données de recherche.

3.1. *Retour sur la notion de donnée*

Revenons sur la première de ces complexités : qu'entend-on par donnée et qu'est-ce qui la distingue de l'information ? Comme sa voisine « information », le terme de donnée est porteur d'une double difficulté, sémantique et théorique :

– d'une part, les acceptions du terme varient selon les disciplines, les approches, les contextes : la donnée ne se définit pas exactement de la même manière en mathématiques, en médecine, en informatique, en SHS, etc. Il suffit d'évoquer la multiplicité des occurrences du terme (bases de données, données personnelles, données médicales, données d'un problème, etc.), pour en saisir immédiatement la diversité sémantique ;

– d'autre part, il est difficile de définir la donnée indépendamment de l'information, et même de la connaissance, car ces trois notions ont partie liée et constituent une sorte de « triade conceptuelle », qui a alimenté de nombreuses réflexions, notamment en sciences de l'information. Pour en mesurer toute la richesse, on peut rappeler ici l'étude internationale, menée par le chercheur israélien Chaims Zins en 2003-2005, qui avait interrogé 57 chercheurs en Sciences de l'Information de 16 pays, pour leur demander leurs définitions des trois notions Information, Données et Connaissance [ZIN 07] : l'étude avait ainsi recensé pas moins de 130 définitions !

Nous nous contenterons ici de quelques considérations générales sur l'articulation données / information / connaissance. Une première réponse est fournie par Serge Abiteboul, pour qui une donnée (*data*, en anglais) peut être définie comme une « description élémentaire d'une réalité » ; le célèbre informaticien donne un exemple simple de la triade donnée-information-connaissance :

> « Des mesures de température relevées chaque jour dans une station météo, ce sont des données. Une courbe donnant l'évolution dans le temps de la température moyenne dans un lieu, c'est une information. Le fait que la température sur Terre augmente en fonction de l'activité humaine, c'est une connaissance. » (S. Abiteboul, *Sciences des données : de la logique du premier ordre à la Toile*, 2012)

Dans la même perspective, Bruno Teboul précise que « la donnée serait la traduction la plus immédiate, la plus brute, d'un fait. Elle n'est pas le fait, mais l'unité minimale d'observation qui permet de le caractériser » [TEB 17]. Ainsi peut-on dire dans un premier temps qu'une donnée est

n'importe quel élément, n'importe quel enregistrement, prélevé ou construit, permettant de décrire ou d'exprimer n'importe quelle partie de réalité : une mesure, un indice, un prélèvement, un chiffre, une lettre, une photo, etc. 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.

C'est pourquoi la donnée est généralement appréhendée comme la matrice, le substrat de l'information, elle-même à la base de la connaissance. Comme l'énonce Sylvie Leleu-Merviel, « les données, à partir desquelles s'élabore l'information intermédiaire, s'avèrent donc la matière première du processus de signifiance (le faire-sens) » [LEL 10, p. 6]. Ne pouvant malheureusement pas aller plus loin ici sur cette articulation complexe entre donnée et information, nous nous permettons de renvoyer à l'article remarquable de Sylvie Leleu-Merviel, qui présente notamment les travaux du philosophe italien Luciano Floridi [FLO 05], sans doute l'un de ceux qui ont proposé l'une des approches les plus approfondies sur la distinction donnée / information.

Une dernière remarque sur cette complexité théorique : on sait le terme particulièrement trompeur du fait de son étymologie, de sa suggestion de l'idée de « don ». On reviendra plus loin sur ce point en ce qui concerne les données de recherche, mais on peut souligner d'ores et déjà que, même dans leur dimension la plus globale, les données ne sont précisément jamais données : « Il est clair que les données ne s'imposent pas d'elles-mêmes : elles sont saisies à partir d'un acte de lecture de la part d'un sujet agissant. On dira ainsi que l'"on n'a jamais affaire à des données brutes (ou crues)" comme l'écrit Gregory Bateson » [LEL 10, p. 4] pour citer encore Sylvie Leleu-Merviel, qui ajoute :

> « Les données sont toujours sélectionnées, transformées, car on ne peut accéder à la totalité des données passées, présentes et à venir. Les données procèdent d'une différence ou d'un manque d'uniformité dans un contexte et surgissent du fait d'un certain cadrage (lequel ne se fait pas au hasard et appelle toujours un sujet agissant). La compréhension des données comme résultant d'une forme de saillance perceptive, émergence d'une figure sur un fond, confirme d'ailleurs ce trait. » (S. Leleu-Merviel, « Le sens aux interstices, émergence de reliances complexes », 2010)

Autrement dit, de même que l'information, au sens social du terme, est toujours le produit d'une interaction entre un humain et un support, est toujours construite par un regard humain (l'information n'existe pas « en soi », indépendamment du regard qui lui donne sens), de même les données sont toujours « sélectionnées » et finalement construites par l'observateur.

3.2. *Complexité épistémologique des données en SHS*

La complexité des données de recherche en SHS se manifeste sur deux plans, étroitement liés. Tout d'abord, au plan épistémologique, le caractère construit des données de recherche conditionne tout le cycle des données. Dans *L'espoir de Pandore*, Bruno Latour résume très bien ce « constructivisme » des données de recherche, y compris en STM :

> « Pour que les données de la botanique et celles de la pédologie puissent se superposer plus tard sur un même diagramme, encore faut-il que leurs deux référentiels soient compatibles. Décidément, on ne devrait jamais parler de "données" mais toujours d'"obtenues". » (B. Latour, *L'espoir de Pandore. Pour une version réaliste de l'activité scientifique*, 2001)

Cette caractéristique essentielle est encore plus manifeste en SHS, où les données de recherche, y compris dans leur forme « brute » de données primaires, sont toujours le produit d'une démarche, d'une problématique, d'un choix du chercheur. Elles sont littéralement « produites » par le chercheur, comme par exemple un corpus de textes, d'images ou de photos, qui aura fait l'objet d'une sélection, un ensemble de sites web collectés, des transcriptions d'entretiens semi ou non directifs, etc. Il existe d'innombrables exemples où les chercheurs en SHS construisent complètement leurs « données sources », pour reprendre la terminologie en vigueur.

Cette dimension épistémologique, *i.e.* la construction des données de recherche en SHS, a d'évidentes conséquences sur les représentations et les pratiques en matière de partage des données, et explique au moins trois types de réticences face au partage, qui sont apparues au cours de l'enquête : l'objection scientifique, les réticences « personnelles », voire affectives (cf. 4.1), et l'objection juridique (cf. 4.2).

Cette construction des données de recherche a été mise en avant par plusieurs chercheurs pour expliquer leurs réticences face au partage des données [SER et al. 17]. Elle est à l'origine de ce qui peut être qualifié d'« objection scientifique » au partage, fondée sur la difficile, sinon impossible, contextualisation d'un jeu de données en SHS. Selon plusieurs enseignants-chercheurs, notamment en géographie sociale, en sociologie, en sociolinguistique, en sciences de l'information et de la communication, les métadonnées, aussi complètes soient-elles, ne peuvent pas toujours rendre compte du contexte scientifique dans lequel a été élaboré un jeu de données. Décrire le jeu de données peut en effet revenir à décrire l'hypothèse de recherche, le cadre méthodologique, les critères de sélection dans la constitution d'un corpus, etc. Et cette « méta-description » peut s'apparenter assez vite à une description tout court, faisant partie intégrante de la recherche elle-même et de l'article qui rend compte des résultats. De plus, dans certaines disciplines comme la sociolinguistique, le contexte de production des données de recherche, constituées d'éléments langagiers, serait quasiment impossible à décrire et à contextualiser : selon un chercheur interrogé, « dès lors qu'on travaille sur l'ancrage social et sur les interactions entre les pratiques langagières et les pratiques sociales dans lesquelles elles sont produites et qu'elles contribuent à produire, tu ne peux plus partager » [SER et al. 17, p. 85].

Avec 27 % des réponses (sur un panel de 108 répondants), les « raisons scientifiques » (qui incluent donc cette objection de la contextualisation) figurent en quatrième position dans les facteurs de frein au partage des données [SER et al. 17]. Et dans les entretiens, cet item de la contextualisation, particulièrement commenté par quelques chercheurs, était considéré comme le véritable frein au partage des données.

Cette deuxième dimension des données de recherche, que nous appelons scientifique ou méthodologique, s'applique de manière très différenciée selon les disciplines, et ce n'est pas un hasard si ce sont les disciplines les plus « sociales » qui mettent le plus en avant cette objection. Des données d'observation (par exemple la captation des mouvements humains), ou des données purement textuelles (de corpus, etc.) seront plus facilement objectivables, car décontextualisées, par les chercheurs.

3.3. *Complexité sémantique*

Dernier facteur de complexité concernant les données de recherche : la difficulté à en donner une définition standard, unanimement reconnue. Ainsi peut-on comparer ces deux définitions officielles, qui divergent légèrement. La première, la plus connue et la plus citée, est celle donnée par l'OCDE en 2007 :

> « 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. » (OCDE, « Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics », 2007) (souligné par nous).

La seconde est plus récente et provient du projet de refonte de la directive européenne sur les informations du secteur public, en avril 2018 :

> « Documents se présentant sous forme numérique, autres que des publications scientifiques, qui sont recueillis ou produits au cours d'activités de recherche scientifique et utilisés comme éléments probants dans le processus de recherche, ou dont la communauté

scientifique admet communément qu'ils sont nécessaires pour valider des conclusions et résultats de la recherche » (Commission européenne, « Proposition de directive du Parlement Européen et du Conseil concernant la réutilisation des informations du secteur public », 2018) (souligné par nous).

Si l'on observe une convergence entre les deux définitions sur la fonction de validation des résultats de recherche attribuée aux données de recherche, deux différences sautent aux yeux : dans la définition de la Commission européenne, les données de recherche sont assimilées à des « documents », et surtout ces documents sont exclusivement liés au support numérique. Ce qui laisse voir une vision assez restrictive de la nature des données de recherche, inapplicable notamment en SHS où les chercheurs travaillent très souvent sur des données non numériques.

Difficultés des institutions à s'entendre sur leur définition, difficultés des chercheurs en SHS à nommer leurs données et à les distinguer parfois des publications, difficultés enfin de compréhension de la terminologie utilisée par les spécialistes de l'IST (par exemple les distinctions entre « données sources » et « données résultats », ou entre « données chaudes » et « froides », sont loin d'aller de soi pour la majorité des enseignants-chercheurs) : la complexité sémantique autour des données de recherche est bien réelle et s'ajoute aux précédentes.

4. L'artisanat des données en SHS

Le constat selon lequel il n'y aurait finalement pas de données, mais seulement des construits [SER et al. 17, p. 36] a été particulièrement affirmé lors des entretiens menés avec les chercheurs. Cette perception des données, aussi bien sources que produites, comme un construit intellectuel, est un élément essentiel dans le retour réflexif qu'opère le chercheur sur la gestion de ses données et dans le positionnement qu'il peut avoir, notamment face aux possibilités d'ouverture. Sans nier la variabilité des pratiques individuelles, elle renvoie aussi à l'idée d'un écosystème de recherche, lié au champ disciplinaire ou dans certains cas, plutôt à une méthodologie transdisciplinaire ; cet écosystème, dans lequel le caractère artisanal et individuel de la recherche en sciences humaines et sociales est facteur d'une appropriation affective, induit de fait des logiques et des approches spécifiques quant à la gestion, à la réutilisation, au partage des données.

4.1. *De la construction intellectuelle à l'appropriation affective*

La recherche en SHS peut revêtir un caractère d'artisanat, au sens où la science est « façonnée » par le chercheur. Cela est particulièrement vrai de certaines disciplines comme l'histoire ou les lettres, dans lesquelles le chercheur « peut avoir encore un rapport très artisanal à la recherche » [SER et al. 17, p. 87]. Et dans ces champs disciplinaires, il en va de la collecte et de la gestion des données comme des autres étapes de la recherche ; le travail sur les données est une construction « à la main » au sens propre du terme quand ce travail mobilise encore l'écriture manuscrite, même si le numérique devient la règle : « je pense que nos collègues ont tous dans leurs cartons chez eux, dans leur grenier en fait, enfin maintenant dans leurs disquettes, dans leurs disques durs, les résultats de leurs dépouillements » [SER et al. 17, p. 87]. Nous avons déjà montré que les données sources à la base du travail du chercheur sont toujours, pour une large part, non numériques : 45 % de ces données sont non numériques, à égalité avec les données numérisées, et seulement 10 % sont décrites comme nativement numériques [SER et al. 17, p. 28]. Il peut donc y avoir, dans un certain nombre de cas, tout un long travail « manuel » d'entrée de données dans des bases personnelles, comme cela a été rapporté pour la recherche en prosopographie, de numérisation, de transcription, de codage ou de ressaisie de texte.

Il s'avère cependant que le poids de la méthode scientifique est au moins aussi important que le champ disciplinaire dans la perception d'un processus d'élaboration long et minutieux de la donnée : c'est ainsi la construction de « tout un appareil méthodologique » [SER et al. 17, p. 36], soit propre au champ disciplinaire, soit partagé par plusieurs disciplines, qui est à l'origine du prélèvement de données. Et pour les chercheurs, le travail sur les données, « ce n'est pas tant une question de

discipline, mais de paradigme de recherche avec des méthodes qui ne sont pas les mêmes ». On retrouve ce constat d'une influence déterminante des méthodes et des équipements dans la dernière étude de Joachim Schöpfel :

> « Les chercheurs de la même discipline ou du même laboratoire n'ont pas nécessairement les mêmes pratiques et approches des données ; et cela veut dire aussi que d'autres chercheurs appliquent les mêmes "stratégies de données" du fait d'une même approche méthodologique (par exemple, des enquêtes, des enregistrements d'entretiens, ou des analyses de réseaux) et ceci malgré une origine disciplinaire différente. » (J. Schöpfel, *Vers une culture de la donnée en SHS : Une étude à l'Université de Lille*, 2018)

Au travers de la méthode, c'est le regard du chercheur qui donne sens et valeur à la donnée.

Les formulations employées pour décrire le travail sur les données en SHS expriment particulièrement bien l'importance, la longueur, l'érudition de la « fabrique » des données par le chercheur qui les collecte et les traite ; relevons simplement à cet égard les mentions récurrentes d'un « travail de bénédictin », ou encore d'un travail « en moniale » [SER et al. 17, p. 37]. Derrière ces expressions, il y a encore la notion d'un travail individuel, voire individualiste, qui peut être solitaire, qui demande un effort de longue haleine et s'effectue dans le cadre d'une indépendance revendiquée du chercheur : celui-ci consacre temps et énergie pour sélectionner, organiser, traiter et exploiter ses données (« des heures et des heures qui ont été passées à monter une donnée ») [SER et al. 17, p. 91], et dans certaines disciplines en particulier, comme cela a pu être noté pour l'histoire, s'exprime un attachement très fort à l'indépendance du chercheur, à l'originalité, à la personnalisation de ses pratiques en matière de gestion de données. En effet, elles concourent à la constitution d'un terrain de recherche spécifique et pérenne, crucial pour le chercheur, et qu'il n'est pas forcément prêt à partager : « quand on se constitue un terrain comme ça, c'est quelque chose de très précieux, car c'est aussi la condition du maintien dans un niveau d'échange scientifique assez élevé » [SER et al. 17, p. 37]. Notons que cet argument de l'originalité du terrain n'est pas valable pour toutes les disciplines SHS, mais qu'il reste toutefois assez partagé.

Le temps de collecte des données est aussi perçu par les chercheurs comme un temps d'élaboration de la pensée [SER et al. 17, p. 37]. Et dans cette mesure, le chercheur évoque non pas une collecte « brute », mais certaines méthodes de recueil et certains types de données indiquant davantage un processus de travail et un construit progressif (transcription manuelle, reprise de notes, etc.). Finalement, la construction des données intervient à tout moment de la recherche, dans une sorte de *continuum* allant jusqu'à la publication et il peut d'ailleurs sembler difficile pour les chercheurs de définir la donnée traitée [SER et al. 17, p. 36] : « c'est difficile de savoir où commence le traitement », « comment situer les données, par rapport à ce processus d'élaboration » [SER et al. 17, Ann. 1, p. 4]. Le travail sur les données apparaît comme un processus évolutif, mobile, où la donnée peut être successivement construite, déconstruite, reconstruite :

> « Si tout d'un coup, je retrouve un fragment et je complète le nom, les dates de mes sources, elles peuvent changer tout d'un coup 20 ans après, on décale tout de 50 ans, donc il faut inventer un système où on puisse gérer cette mobilité […] ce sont toutes ces étapes de déconstruction des données que j'essaie de rendre apparentes. » (A. Serres et al., *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2*, 2017)

Le temps passé, l'effort consenti et l'énergie mise au long travail de constitution des données ont pour effet une appropriation affective [SER et al. 17, p. 38], qui en définitive fait passer les données hors du domaine de l'impensé, qui était le leur en tant que structure neutre et utilitaire, et les relie directement à l'élaboration d'un terrain de recherche original, vital pour les projets de publication. Le sentiment d'un fort investissement personnel pour la constitution de ce terrain de recherche (« j'ai mis en place une méthode, j'ai passé du temps, je les ai élaborées, j'ai travaillé, j'ai produit mes données

avec des acteurs sociaux […] ») [SER et al. 17, p. 20] conduit les chercheurs, pour certaines disciplines du moins, à considérer les données collectées, traitées, exploitées comme autant de composantes d'un territoire qui leur est propre. Mais c'est surtout dans la perte que s'exprime l'attachement aux données [SER et al. 17, p. 39]. Et lorsqu'il arrive que des données disparaissent, soient dérobées, ou encore soient devenues techniquement inaccessibles, cette perte de contrôle est vécue comme une dépossession traumatisante, preuve que le sentiment d'intimité du chercheur avec ses données s'est mué en sentiment de propriété sur celles-ci.

4.2. *De l'appropriation affective à la revendication d'une propriété intellectuelle*

Le fort rapport d'appropriation affective développé par les chercheurs vis-à-vis de leurs données se heurte parfois au cadre juridique existant en matière de données de recherche. Si les chercheurs sont indéniablement détenteurs d'un droit d'auteur au regard de leurs publications scientifiques, la situation concernant les données se révèle plus complexe. Le cadre juridique national s'est trouvé récemment reconfiguré de manière substantielle avec la loi pour une République numérique promulguée le 7 octobre 2016. Ce texte a notamment apporté deux modifications majeures concernant les modalités de diffusion et de réutilisation des données de recherche :

– d'une part, l'article 6 introduit l'application du principe d'*open data* par défaut pour les jeux de données produits et reçus par les établissements d'enseignement supérieur et de recherche [MAU 16a] ;

– d'autre part, l'article 30 instaure la libre réutilisation des données dès lors qu'elles sont issues d'une activité de recherche financée au moins pour moitié sur des fonds publics, rendues publiques par le chercheur et non protégées par un droit spécifique ou une réglementation particulière (par exemple la protection des données à caractère personnel et sensible).

Ainsi, le choix d'ouvrir ou non ses jeux de données n'est plus entièrement laissé à la discrétion des chercheurs, mais relève dans certains cas d'une obligation. Ensuite, au-delà des modes de diffusion, c'est l'attribution de la propriété intellectuelle qui se trouve au cœur des interrogations, notamment dans le cadre d'une création de base de données. En effet, dans une majorité des cas, ce sont les institutions de recherche et/ou les établissements financeurs auxquels les chercheurs sont rattachés, et non les chercheurs eux-mêmes, qui détiennent la propriété intellectuelle sur les données produites [MAU 16b].

La complexité du dispositif juridique en vigueur impose quasiment un traitement au cas par cas pour démêler à qui appartiennent réellement les données et provoque de nombreuses incompréhensions de la part des chercheurs, allant parfois jusqu'à susciter chez ces derniers l'impression d'être dépossédés de leurs matériaux de recherche. Cela peut être vécu comme une injustice, ou tout au moins en décalage avec le lourd investissement personnel que requièrent les activités de collecte et de traitement des données. En témoignent les paroles de chercheurs : « j'ai passé deux ans de ma vie à traiter, exploiter une base d'entretiens » ou encore « la vraie donnée c'est ça, c'est des heures de boulot, d'analyse » [SER et al. 17, Ann. 3, p. 5]. De fait, la constitution et le traitement d'un corpus de données désignent un ensemble d'activités à la fois invisibles, chronophages et peu reconnues institutionnellement.

L'invisibilisation du travail autour des données est double. Elle est provoquée à la fois par une disparition de la tâche elle-même, dans la mesure où le processus d'analyse des données se trouve dilué dans la publication finale qui, elle, constitue l'élément tangible de l'activité de recherche, et par une certaine dévalorisation sociale de la gestion des données pouvant être jugée comme un « travail de petites mains », voire un « sale boulot » comme le met en évidence Florence Millerand [MIL 12]. Néanmoins, les chercheurs et chercheuses font part de points de vue contrastés sur ce point. Certains d'entre eux estiment que le travail des données ne les concerne pas, là où d'autres ont intégré ces activités dans leurs pratiques de recherche et les considèrent comme une étape essentielle d'élaboration de leur réflexion [SER et al. 17, p. 42].

Afin d'expliquer leur faible implication dans les activités de gestion de leurs données, de nombreux chercheurs invoquent le problème du manque de temps. Parmi les propositions qui ont pu être formulées pour répondre à ce problème, deux d'entre elles se dégagent particulièrement. La première consiste à déléguer entièrement le travail des données auprès d'un personnel dédié. Or, il se trouve que cette option ne peut être satisfaisante puisque certaines activités (comme la description) ne peuvent être effectuées par quelqu'un d'autre que le chercheur ou la chercheuse qui a collecté et analysé les données en question. La deuxième piste envisagée est de plaider pour une science ralentie qui laisserait aux chercheurs le temps nécessaire pour réaliser ce travail eux-mêmes, tout en demeurant accompagnés par un personnel spécialiste et compétent pour les tâches qui le requièrent (telles que les opérations d'archivage par exemple). On a pu voir à ce sujet émerger le concept de *slow science*, notamment analysé par Isabelle Stengers [STE 17], qui interpelle sur la nécessité d'un ralentissement de la science dans une société elle-même soumise à l'accélération [ROS 10].

Par ailleurs, si ces activités se trouvent dévalorisées ou ne sont pas toujours perçues comme une problématique déterminante, c'est aussi parce qu'elles ne semblent pas suffisamment reconnues au niveau institutionnel. La création de bases de données pour les chercheurs en SHS ne leur permet généralement pas, ou très peu, d'avancer dans leur carrière : « je suis en --ème section CNU, et au fond je ne peux habiliter sur l'outil » [SER et al. 17, Ann. 3, p. 6]. Or, il s'agit d'un enjeu crucial dans un monde de la recherche où l'injonction à la publication est très prégnante (évaluation au moyen d'indicateurs bibliométriques et culture du *publish or perish*). Dans ce contexte, les tâches liées aux données se trouvent plus facilement reléguées au second plan, ce qui est dommageable dans la mesure où elles peuvent pourtant constituer d'importantes contributions à la connaissance. C'est l'idée que défend dans ces termes Marcello Vitali-Rosati :

> « L'article est une forme de stabilisation de la connaissance adaptée au dispositif de l'imprimé. Or, aujourd'hui, cette forme n'est pas la seule possible. […] Le numérique permet d'autres formes de stabilisation : des données structurées, par exemple, peuvent être une contribution importante à la recherche. » (M. Vitali-Rosati, « À quoi servent les publications scientifiques ? », 2018)

Le travail de curation et d'ouverture des données se révèle pleinement porteur de sens sur le plan scientifique, mais comment traduire ce sens en pratique ? Premièrement, une refonte des méthodes d'évaluation de la recherche paraîtrait nécessaire, comme le note Hans Dillaerts en s'appuyant sur une des recommandations du Comité d'éthique du CNRS (COMETS) formulée en 2015, afin d'y intégrer une meilleure reconnaissance du « travail de mise à disposition de données utilisables à partir de données brutes » [DIL 17]. Ensuite, il conviendrait de bien distinguer propriété intellectuelle et reconnaissance institutionnelle. Une revendication appuyée de la première par les chercheurs, telle qu'elle peut l'être parfois dans le but d'attester leur rôle fondamental dans l'ouverture des données de recherche, ne permet pas de compenser les lacunes de la seconde ; de fait, cette revendication apparaît inefficace, voire contreproductive. Au lieu de contribuer à verrouiller les accès, il semble davantage pertinent de réinterroger la question de la propriété intellectuelle sous le prisme de la notion des biens communs [COR 17]. L'ouverture des données de recherche pourrait alors s'inscrire dans une approche des biens communs scientifiques visant à « repenser la place de la production, du partage et de la diffusion de connaissances dans nos sociétés » [BRO 18].

En définitive, il s'avère important de déplacer la problématique : ce qui freine l'ouverture des données de recherche ne réside pas dans la nature même du principe, auquel une majorité de chercheurs adhère, mais dans des obstacles externes qui en entravent sa mise en œuvre. Parmi ces obstacles, on peut citer la dimension concurrentielle du milieu de la recherche (exacerbée par certaines « dérives de l'évaluation » [GIN 14]) qui alimente une crainte répandue du plagiat, ou encore le risque de captation économique des données par une poignée de grands éditeurs exerçant déjà leur monopole sur le marché des revues scientifiques.

De même que la « donnée » est communément envisagée comme une notion aux contours flous, le dispositif législatif qui encadre sa mise à disposition et son ouverture demeure fréquemment mal compris par les chercheurs. Pourtant, il semble qu'au-delà des diverses injonctions, ces derniers ont un intérêt notable à s'emparer de la question des données de recherche : leur appropriation de cet objet ne doit pas être uniquement affective, elle est aussi déjà profondément épistémologique (les données sont construites par un faisceau de pratiques et de méthodes, issues d'un savoir situé) et tendrait également à devenir sociale et politique.

5. Les données de recherche : une construction politique

Les données de recherche constituent une préoccupation partagée par différents acteurs politiques aux discours tantôt analogues, tantôt contradictoires. L'objectif de cette dernière partie sera d'en analyser les registres et les argumentaires, afin de les comparer avec les pratiques réelles des chercheurs sur le terrain.

5.1. *Diversité des discours politiques*

Les discours militants liés à l'Open Access et l'Open Data intègrent dès l'origine cette dimension des données de recherche, qu'il s'agisse de la déclaration de Berlin (2003)[4] ou de l'emploi de l'expression Open Data « forgée, en 1995, dans un rapport américain du National Research Council sur "l'échange complet et ouvert des données scientifiques" » [GAI 14, p 12]. Les financeurs, comme les organisations transnationales (OCDE, Union européenne) ou les institutions qui encadrent la recherche en France, se sont progressivement emparés de cette thématique. Rémi Gaillard mentionne quelques jalons historiques dans son mémoire ENSSIB de 2014 [GAI 14, p 21-23], complétés ici pour la période récente.

Au niveau français, on peut citer l'organisation en 2012 d'une journée d'étude intitulée « Données de la recherche : enjeux, perspectives, politique(s) » par le CNRS et la mise en ligne du site *Données de la Recherche* la même année ; la création du segment « Données de la recherche » du projet Bibliothèque Scientifique Numérique (2013) ; la loi Pour une République Numérique (2016) ; le lancement de la plateforme DoRANum (INIST-CNRS, Réseau des URFIST, 2016) ; le Plan national pour la science ouverte (M.E.S.R.I., 2018) ; l'obligation de déposer un plan de gestion de données dans le plan d'action 2019 de l'ANR (juillet 2018).

Au niveau européen, on peut mentionner les recommandations du Conseil scientifique du Conseil Européen de la Recherche en décembre 2007 (mise en accès libre des résultats de la recherche) ; le rapport *Riding the wave. How Europe can gain from the rising tide of scientific data*, rédigé par le Groupe d'experts sur la gestion des données scientifiques de la Commission européenne (octobre 2010) ; la recommandation aux États membres « relative à l'accès aux informations scientifiques et à leur conservation » (2012) ; le lancement d'un projet pilote pour le libre accès aux données de la recherche dans le cadre du PCRD Horizon 2020 (décembre 2013), projet étendu en 2016 à tous les appels du programme ; les modalités en discussion pour le futur PCRD Horizon Europe, avec l'obligation de plans de gestion de données pour des données de recherche FAIR[5] et ouvertes.

On voit à travers ces jalons la progression et l'amplification d'un discours de structuration et d'ouverture des données, porté par une forte volonté politique. Elle se traduit en pratique au niveau des acteurs de l'IST par l'émergence depuis 2014 de formations à destination des professionnels et des chercheurs (INIST, URFIST, ENSSIB, SCD), formations de sensibilisation puis formations pratiques pour donner des clés de mise en œuvre (création de plans de gestion de données, aspects juridiques, etc.). Thomson-ISI a lancé dès 2012 un *Data Citation Index* pour faciliter la découverte et mesurer

[4] « 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 » [DEC 03, p. 1]

[5] Acronyme pour Findable, Accessible, Interoperable, Re-usable

l'impact des données. Les éditeurs s'emparent eux aussi de ces thématiques avec la préconisation, voire l'obligation, de mettre à disposition les données comme préalable à la publication d'un article (politique des revues Nature, création de badges Open Data par Springer), la création de *data journals* ou encore le développement de services et d'outils dédiés à la curation et à la recherche de données (DataSearch et Mendeley Data pour Elsevier, Research Data Support pour Springer-Nature).

Les arguments avancés pour justifier et convaincre de l'intérêt de l'ouverture des données jouent sur plusieurs registres, de l'intégrité scientifique aux arguments économiques ou au discours démocratique : améliorer la qualité de la science en facilitant la reproductibilité et la vérification des résultats de recherche, valoriser et rendre plus visible la recherche, favoriser et accélérer le progrès scientifique, permettre de nouvelles formes de sciences, améliorer la transparence de l'information pour la société et les citoyens, ne pas dupliquer les efforts de financement, stimuler la croissance économique et l'innovation.

Ces arguments sont-ils tous utilisés indifféremment ? L'analyse de cinq textes[6] émanant de divers acteurs institutionnels au cours de l'année 2018 permet de faire plusieurs constats.

	Date	Source	Périmètre	Cible
[FRA 17]	29 nov. 2017	Gouvernement français	Europe	Union européenne
[CNR 18]	16 mars 2018	CNRS	Europe	Union européenne
[FRA 18]	4 juillet 2018	Gouvernement français (MESRI)	France	Communauté scientifique française
[COM 18b]	25 avril 2018	Commission européenne	Europe	Gouvernements européens
[ANR 18]	26 juillet 2018	ANR	France	Communauté scientifique française

Tableau 1. *Textes analysés*

Tout d'abord ces textes font partie d'un écosystème international de textes liés à l'Open Data et l'Open Science. Le Plan national pour la science ouverte fait référence aux engagements pris dans le cadre du Partenariat pour un gouvernement ouvert et cite l'initiative européenne *Amsterdam call for action on open science* (2016) mais pas la recommandation de la commission d'avril 2018. Le plan d'action 2019 de l'ANR se réfère, sur le plan français, à la Stratégie nationale de recherche et au Plan national pour la science ouverte, et sur le plan européen, à Horizon 2020 et au futur programme cadre FP9[7], ainsi qu'à l'OCDE et l'ONU. La recommandation de la Commission européenne cite uniquement des textes produits par les différentes instances de l'Union européenne (Commission, Conseil, Parlement). Les contributions du CNRS et des autorités françaises au FP9, qui s'adressent à des politiques avertis, ne font pas référence à des textes extérieurs. La référence semble faire fonction de premier argument de conviction, en montrant l'importance internationale du sujet.

Au-delà de la citation, il y a une filiation non déclarée dans le contenu des textes, entre la formulation de principes dans un cadre général et la traduction de ces principes en injonctions concrètes à destination des chercheurs. Ainsi, la plupart des éléments cités dans la recommandation de la Commission qui demande aux États membres de « définir et mettre en œuvre des politiques claires »

[6] Plan national pour la science ouverte [FRA 18] ; Position préliminaire de la France sur le 9e PCRI [FRA 17] ; Contribution du CNRS au 9e PCRI [CNR 18] ; Plan d'action 2019 de l'ANR [ANR 18] ; Recommandation 2018|790 de la Commission européenne relative à l'accès aux informations scientifiques et à leur conservation [COM 18b]

[7] FP pour *Framework Program*. Il s'agit du futur programme cadre européen pour la recherche et l'innovation qui sera effectif à partir de de 2021. Il prendra la suite du programme actuel H2020.

et des « plans d'action nationaux » en matière de gestion des données de la recherche financée par des fonds publics sont repris dans le Plan national pour la science ouverte, qui déclare vouloir adopter une « politique qui prolonge et amplifie les efforts de l'Union européenne ». L'ANR décline ce plan pour la Science ouverte, en posant comme valeur et engagement de son plan d'action 2019 l'Open Science, la diffusion et le partage ; elle impose les plans de gestion de données (mesure prônée par la recommandation pour les agences de financement nationales) mais se montre peu prolixe sur les moyens concrets du partage, en dehors d'une référence à OPIDoR.

Dans leur structure, la recommandation de la Commission européenne et le Plan national pour la science ouverte font aux données une place propre, en exergue de l'Open Access (un des trois axes du plan, 2e point sur sept de la recommandation), alors que les autres textes traitent dans la même partie l'accès ouvert aux publications et aux données. Il ne s'agit pas d'une méconnaissance de la problématique, puisque d'autres documents internes du CNRS se penchent sur les données[8], mais sans doute plutôt d'une volonté différente de communication selon les cibles, d'un côté pour mettre l'accent sur la thématique et ses implications stratégiques auprès des décideurs, et de l'autre pour insister plutôt sur l'importance globale de la diffusion et du partage de tous les résultats de la recherche auprès des chercheurs.

Tous ces textes jouent sur le registre incitatif : les leviers sont l'orientation, la conviction et la valorisation des actions vertueuses, dans le respect de l'indépendance du chercheur, plutôt que l'obligation, hormis dans le cadre contraignant d'un financement de projet (obligation de dépôt des publications et de plans de gestion de données pour l'ANR et H2020). Les textes à portée générale (Recommandation, Plan national pour la science ouverte), à visée de communication et de conviction, sont plus argumentés, puisqu'on y retrouve cinq des sept catégories d'arguments identifiés (cf. tableau 2), alors que le plan d'action de l'ANR ou les textes remontés pour le FP9 en citent au maximum deux. Ils prônent tous le partage, avec la nuance pour la position française sur le FP9 de la protection des intérêts des entreprises et d'un système « aussi ouvert que possible, aussi fermé que nécessaire ». On pourrait s'étonner que l'intégrité scientifique ne soit pas un argument sur lequel joue l'ANR. Mais de fait, il s'agit d'une notion tellement importante qu'elle est citée comme une des quatre valeurs fondamentales du plan, au même titre que l'accès ouvert et dissociée de celui-ci.

	[COM 18b]	[CNR 18]	[FRA 17]	[FRA 18]	[ANR 18]
Améliorer la qualité de la science en facilitant la reproductibilité et la vérification des résultats de recherche	4	1		1	
Valoriser et rendre plus visible la recherche					1
Favoriser et accélérer le progrès scientifique			1		
Permettre de nouvelles formes de sciences	2			4	
Améliorer la transparence de l'information pour la société et les citoyens	5	2		2	2
Ne pas dupliquer les efforts de financement	3			3	
Stimuler la croissance économique et l'innovation	1			5	

Tableau 2. *Ordre des arguments*

[8] Voir les publications et sites web de l'INIST, de l'INSU, la recommandation « Les moyens du partage des données » adoptée fin 2017 par le Conseil scientifique

Les valeurs mises en avant sont différentes suivant les textes. À l'inverse de la Commission, qui cible la communication d'abord sur les aspects économiques et l'innovation, les textes français privilégient la mise en avant de valeurs fédératrices et positives, tels que les bienfaits de l'ouverture des données pour la science (qualité scientifique, visibilité, progrès scientifique), pour la société et les citoyens. On rejoint le discours militant de l'Open Access, ce qui est plus sensible encore dans les écrits des organismes opérationnels de la recherche (CNRS et ANR), plus ambitieux et idéalistes dans la rédaction : la science est considérée non seulement comme un moyen, mais aussi comme une valeur fondatrice du social. Une valeur qui se retrouve également dans la contribution française au FP9 ; en revanche, elle n'est pas reprise dans le Plan national pour la science ouverte.

On peut citer quelques extraits à titre d'illustration : « la science doit être aujourd'hui à la base d'une société de progrès, dans laquelle les avancées technologiques ou sociales profitent au plus grand nombre » (CNRS) ; « la diffusion, le partage et l'archivage pérenne des publications scientifiques et des données de recherche […] contribuent […] à faire de la science un bien commun » (ANR) ; « promouvoir une société de la connaissance basée sur les valeurs d'ouverture, de progrès et d'esprit critique » (Contribution française FP9).

Les autres textes restent sur la notion d'avancée : « le libre accès permet […] d'accélérer le progrès scientifique » (CEE) ; « la science ouverte constitue […] un progrès scientifique et un progrès de société » (Plan national pour la science ouverte).

La place du citoyen est aussi envisagée plus activement par les organismes opérationnels : il est question d'« informer » dans la Recommandation, de « démocratisation » et de « confiance des citoyens » dans le Plan national, alors que l'ANR et le CNRS vont jusqu'à l'appropriation (« bien commun » pour l'ANR, « éclairage aux débats citoyens » et « sciences participatives », CNRS FP9). Cette différence de registre peut s'interpréter par la cible visée.

Il est intéressant aussi de constater que les arguments prioritaires retenus par les politiques ne collent pas forcément aux facteurs d'incitation au libre accès dégagés dans l'enquête rennaise. La visibilité de la recherche, plébiscitée comme premier facteur favorable au partage par les chercheurs, n'est mentionnée que par l'ANR. *A contrario,* le souci de validation des résultats, très présent dans le discours politique, n'est cité que par 11 % des répondants [SER et al. 17, p 60]. L'adhésion aux valeurs du libre accès, mise en valeur dans la rédaction politique, reste donc une valeur sûre et un élément où se rejoignent les discours.

Les variations des discours, des structures et des arguments avancés reflètent l'objectif et le public visés par les textes ; elles témoignent aussi de la façon dont le politique s'empare de cette thématique qui n'est pas (plus) seulement scientifique.

5.2. *La réalité des pratiques face aux discours sur le partage*

Les différentes paroles élaborées autour du partage des données, paroles politiques, institutionnelles ou militantes, sont-elles toujours pertinentes lorsqu'elles se trouvent confrontées aux réalités du terrain ? L'analyse comparée des discours et des pratiques permet de faire le constat d'un double écart : écart d'une part entre les discours politiques surplombants d'injonction, d'incitation et les pratiques effectives montrant que le stockage et l'archivage des données sont le premier problème rencontré par les chercheurs, et que la notion de partage recouvre une réalité complexe qui conduit à des positions différenciées ; écart d'autre part dans les discours des chercheurs eux-mêmes, entre le plus souvent un accord de principe, qui se fonde sur différents arguments (reconnaissance des valeurs du libre accès, visibilité de la recherche, accélération et dynamisme de la science, etc.) [SER et al., 2017, p. 60], et d'un autre côté des réticences plus ou moins fortes, en prise directe avec les contraintes du terrain de recherche.

La question du partage des données comme construction politique se heurte à un ensemble de freins de différentes natures, issus à la fois du contexte concret de recherche, mais aussi de représentations

mentales des chercheurs sur les enjeux et les risques potentiels de l'ouverture. Les entretiens menés avec les chercheurs ont permis de préciser notre enquête statistique et d'identifier plusieurs raisons expliquant les réticences au partage [SER et al. 17, p. 61]. Ce sont tout d'abord des raisons juridiques assez évidentes (cf. 4.2), liées directement à la nature des données et découlant notamment du processus d'appropriation affective (cf. 4.1) qu'opère le chercheur sur les données qu'il collecte, utilise et produit. On retrouve fréquemment la crainte du plagiat en cas de diffusion, la peur de se voir dépossédé de données qui ont fait l'objet d'un fort investissement personnel, et c'est là un argument assez répandu, qui révèle de fait une méconnaissance de la réalité complexe de la propriété intellectuelle des données. À ces raisons juridiques s'ajoutent souvent des raisons pratiques, en particulier le manque de temps, dans un contexte où les tâches administratives remplies par les chercheurs se superposent aux activités de recherche proprement dites, et où les tâches liées à l'organisation de la diffusion des données apparaissent comme une contrainte supplémentaire. Mais ce sont sans doute les raisons scientifiques qui apportent les éléments les plus intéressants et non des moindres au débat, comme on l'a vu dans la partie 3.2. Enfin, le manque de connaissances et de compétences, notamment dans la description des jeux de données, constitue pour le chercheur un obstacle notable au partage. Si le chercheur se trouve placé au centre du dispositif, il ne peut toutefois assumer seul les opérations afférentes au partage, qui ne peuvent se faire qu'avec la collaboration de plusieurs acteurs. Tout l'enjeu d'une politique des données de recherche sera sans doute de tenir compte des freins existants pour les réduire et de favoriser la convergence et l'interaction des différents acteurs.

La création de données de recherche structurées et interopérables est le résultat d'un processus global composé de plusieurs étapes interdépendantes, que sont la documentation, l'analyse, le stockage et l'archivage, permettant d'aboutir à la diffusion des données. Arrivant en fin du processus, l'ouverture des données est conditionnée par l'exécution effective des étapes précédentes. Autrement dit, tant que les premiers maillons de la chaîne de traitement des données ne sont pas opérants, le partage ne peut être réalisé dans de bonnes conditions. Or, plusieurs enquêtes portant sur les pratiques des chercheurs [PRO 15 ; SER et al. 17] montrent que, si l'on se place dans une optique d'ouverture, la situation concernant la conservation et la description des données de recherche est encore loin d'être pleinement satisfaisante. En effet, le stockage des données est globalement effectué de manière individuelle et locale, sur des supports non sécurisés. Quant à la question de la documentation, le constat est également mitigé : environ un chercheur sur dix a recours à des standards de métadonnées pour décrire ses données [SER et al. 17, p. 56]. La quasi-absence de protocoles systématiques d'archivage, qui eux seuls sont en mesure de garantir un accès fiable et documenté à ces données, est une des raisons permettant d'expliquer que les pratiques de partages soient encore assez limitées à l'heure actuelle.

Dans son discours du 4 juillet 2018 annonçant la mise en place du Plan national pour la science ouverte, la ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation, Frédérique Vidal, établit que « la première démarche est de structurer et conserver les données, en préalable à leur ouverture » [VID 18]. Cette reconnaissance de la donnée de recherche comme résultat d'un processus apparaît ainsi bien intégrée dans le cadre de la politique scientifique nationale. Mais si la complexité inhérente à la démarche d'ouverture des données est désormais instituée, la cause invoquée pour justifier les lacunes en matière de partage peut être discutée. Celle-ci se trouve identifiée dans la suite du discours : « Cela [la structuration et la conservation des données] se fait très bien dans environ un cinquième de la communauté scientifique, mais le reste est largement laissé à l'appréciation individuelle et aux hasards de la vie des clés USB, des vols d'ordinateurs portables dans les coffres de voitures ou des défaillances des disques durs individuels », puis « le problème n'est pas principalement technique, il est d'abord, et fondamentalement, humain » [VID 18]. Ainsi, l'absence de stratégies de documentation et d'archivage des données de recherche se retrouve, plus ou moins directement, attribuée à « l'appréciation individuelle », aux « hasards de la vie » et de manière générale à l'« humain ». L'insuffisance du partage, telle qu'elle est mise en évidence dans ce discours, est envisagée ici comme une addition de petites défaillances personnelles. Or, comme le souligne Joachim

Schöpfel, il conviendrait de rappeler que « l'absence de motivation et/ou des compétences des chercheurs eux-mêmes » ne constituent pas l'obstacle le plus important à une bonne gestion des données et que « la réduction psychologique et l'attribution culpabilisante d'une responsabilité individuelle ou collective semble plutôt l'arbre qui cache la forêt, et cette forêt, c'est l'absence de ressources informatiques et humaines sur un campus en SHS, en particulier dans les laboratoires universitaires (équipes d'accueil) » [SCH 18, p. 25].

Afin d'impulser le partage des données, le M.E.S.R.I., à travers son Plan national pour la science ouverte, affirme une volonté d'agir sur les pratiques (« transformer les pratiques scientifiques », « généraliser les pratiques quotidiennes de la science ouverte ») [VID 18], mais prévoit également de mettre à disposition des chercheurs des moyens financiers concrets. En effet, il est annoncé qu'« un appel FLASH de l'ANR sera publié prochainement pour accélérer la structuration, la citation et l'ouverture des données de la recherche des équipes françaises » [VID 18]. Mis en place depuis 2010, les appels à projets FLASH ont été conçus pour financer des projets dans un délai court afin de « favoriser la production de résultats scientifiques inédits en lien avec un événement dont l'ampleur et la fréquence sont exceptionnelles[9] ». La dimension « exceptionnelle », telle qu'elle est mise en exergue ici, met bien en évidence le caractère prioritaire accordé par le ministère, et plus largement par la politique nationale, aux données de recherche. Toutefois, au-delà du fait que l'on puisse trouver curieux de recourir à un programme utilisé jusqu'ici dans des situations de catastrophes naturelles (plan FLASH Haïti en 2010 et plan FLASH Japon en 2011), un des principaux écueils de ce type de financement est qu'il ne permet pas d'irriguer l'ensemble des unités de recherche du territoire – qui auraient pourtant toutes des besoins en la matière – mais seulement d'attribuer des dotations ponctuelles, limitées dans le temps et réservées à quelques-uns. Or, comme cela a déjà pu être démontré, la préparation des données de recherche à l'ouverture nécessite la mise en place de protocoles qui s'inscrivent dans une temporalité longue et qui ne peuvent se réduire à un travail réalisé dans l'urgence et/ou sur une période de courte durée. Ainsi, bien que les moyens proposés semblent ambitieux sur la forme, un doute subsiste quant à leur potentielle efficacité à long terme, dans la mesure où ils n'ont pas été conçus pour permettre de développer l'ouverture de manière égale à l'échelle nationale.

Le mouvement pour la science ouverte participe au renforcement d'une tendance globalisée à l'ouverture des données publiques. Cette ouverture se fonde notamment sur une injonction sociale et politique à la transparence. Considérée tantôt comme une exigence scientifique et une garantie d'éthique dans un contexte de recherche, tantôt comme un vecteur de renouveau démocratique, la préoccupation actuelle pour la transparence apparaît également comme un révélateur d'une certaine « mise en crise du régime de vérité » [ROU 13]. Les données collectées par les chercheurs, auparavant privées et non visibles, sont désormais soumises à l'injonction de l'ouverture : « il faut rendre publique la boîte noire » [SER et al. 17, p. 43]. Par ailleurs, il est intéressant de noter que le discours politique n'a commencé à s'emparer de la question des données de recherche qu'à partir du moment où elles se sont justement transformées en « données ». Les sources imprimées constituaient un impensé ; elles deviennent une préoccupation politique avec le développement de la numérisation. Ce constat semble prouver la valeur extra-scientifique accordée aux données de recherche. En effet, comme le note Hans Dillaerts, ces données sont désormais comprises dans le contexte d'une « économie de la connaissance », développée à l'échelle européenne et visant à accélérer la production, la valorisation et l'exploitation des avancées scientifiques. S'appuyant sur une rhétorique commune de la croissance, de l'innovation et de la collaboration, l'économie de la connaissance a pour objectif de créer de la valeur à partir de la recherche scientifique : « les données publiques et scientifiques doivent être abondantes et pouvoir circuler le plus librement et rapidement possible », à l'image des capitaux échangés sur les marchés financiers [DIL 17]. Si les données (personnelles, sociales, publiques, etc.) constituent aujourd'hui à l'échelle mondiale un enjeu crucial autour duquel se tissent de complexes rapports de pouvoir économiques, politiques et sociétaux, les données scientifiques apparaissent alors de plus en plus amenées à jouer un rôle stratégique majeur.

[9] Source : http://www.agence-nationale-recherche.fr/suivi-bilan/editions-2013-et-anterieures/recherches-exploratoires-et-emergentes/flash/

En guise de conclusion

Comme tous les groupes professionnels, les chercheurs sont souvent la cible d'un faisceau de diverses injonctions au changement : les « discours de l'adaptation » (adaptation aux évolutions techniques, au développement durable, à la transition écologique, aux évolutions sociétales, etc.) sont multiples, prégnants, parfois contradictoires. Dans le domaine de la science, qui ne saurait échapper à cette vulgate du changement et de l'adaptation permanente, plusieurs injonctions très fortes sont apparues ces dernières années :

– la nécessité de la communication : *be visible or vanish* serait le nouveau mot d'ordre, incitant les chercheurs à se rendre visibles, notamment sur les réseaux sociaux ;

– l'ouverture de la science : nous en avons abondamment parlé, et le discours sur la « Science Ouverte » est en passe de devenir le nouveau mantra de la communauté scientifique ;

– une science plus fiable et plus intègre : la régulation éthique en cours s'impose avec de plus en plus de force dans la recherche et poussera les chercheurs à améliorer leurs pratiques de recherche, pour renforcer l'intégrité scientifique et l'éthique de la recherche ;

– la protection de l'innovation : les chercheurs sont incités, dans plusieurs disciplines, à soutenir l'innovation, à protéger les inventions, à développer la valorisation de la recherche ;

– et ces quatre types de « discours injonctifs » s'ajoutent à l'injonction historique, désormais classique, qui gouverne encore l'activité des chercheurs, le célèbre *publish or perish*, qui s'accompagne d'un mode d'évaluation quantitative, de plus en plus dénoncé mais toujours dominant.

Ce bref récapitulatif des grands « discours injonctifs », qui ne saurait être exhaustif, permet de voir un certain nombre de contradictions, dans lesquelles se débattent aujourd'hui les chercheurs : par exemple, le développement de l'intégrité scientifique et des « bonnes pratiques » impliquerait une sortie progressive du modèle de la publication à tout prix et de l'évaluation purement quantitative, générateur de toutes sortes d'inconduites scientifiques. De même que les discours sur l'ouverture des données et de la science d'un côté et ceux sur la protection des inventions (par les brevets notamment) de l'autre peuvent apparaître parfois contradictoires, voire relever d'une « injonction paradoxale » [DIL 18]. Enfin, entre la visibilité à tout prix sur les réseaux sociaux, qui peut vite céder à la pression de l'immédiateté d'internet, et la *slow science*, également prônée pour développer une science plus intègre, on pressent que peuvent surgir de nombreuses contradictions.

Dans ce maelström de discours, d'injonctions, de recommandations, d'obligations, qui entoure ou s'abat sur le chercheur, on peut comprendre que celui-ci peut parfois réagir en se repliant sur ce qu'il maîtrise le mieux : sa pratique scientifique, disciplinaire, forgée au cours de ses années de formation et de ses travaux de recherche. Les incitations à l'ouverture et au partage des données de recherche, qui sont l'un des thèmes de notre article, sont donc à replacer dans ce contexte global, dans cet écart entre la réalité quotidienne, parfois la pesanteur des pratiques, individuelles ou collectives, et cet environnement fait d'un ensemble d'énoncés, plus ou moins prescriptifs.

À ce stade, deux observations nous paraissent particulièrement importantes :

– d'une part, éviter à tout prix les discours militants ou injonctifs sur le partage des données, mais plutôt prendre en compte la réalité des pratiques, des besoins, leur diversité, leur complexité ;

– d'autre part, compte tenu de l'extraordinaire hétérogénéité de ces pratiques et de ces besoins, développer une politique des données à géométrie variable, et non pas uniforme, s'imposant à tous les chercheurs. Un point clé est à souligner : la nécessité de trouver le bon niveau d'une politique des données de recherche : l'équipe de recherche, le laboratoire, l'UFR, l'établissement, le regroupement d'établissements ?

Plusieurs niveaux d'intervention politique restent nécessaires. Celui du projet de recherche (notamment pour les projets ANR ou européens), mis en avant par Joachim Schöpfel dans sa dernière enquête [SCH 18], semble particulièrement pertinent, à l'heure de l'obligation des plans de gestion de

données instaurée prochainement par l'ANR. Les obligations institutionnelles ou juridiques restent le meilleur moyen de faire changer les pratiques, et un champ d'action important va s'ouvrir pour tous les acteurs impliqués dans le mouvement pour la Science Ouverte.

Mais les obligations ne suffisent pas à modifier en profondeur les pratiques si elles ne s'accompagnent pas d'un changement de culture. L'un des grands enjeux à venir reste le développement d'une véritable culture des données, *i.e.* d'une élévation significative des compétences (en tant que savoirs, savoir-faire et savoir être) multiples, requises par cette nouvelle culture des données. L'enjeu de la formation à cette culture est crucial, pour un changement en profondeur des pratiques et une adaptation « maîtrisée », critique, aux évolutions nécessaires.

Bibliographie

[ABI 12] ABITEBOUL S., « Sciences des données : de la logique du premier ordre à la Toile : Leçon inaugurale prononcée le jeudi 8 mars 2012. Chaire d'Informatique et sciences numériques ». In *Sciences des données : de la logique du premier ordre à la Toile : Leçon inaugurale prononcée le jeudi 8 mars 2012*. Leçons inaugurales, Collège de France, Paris, 2013. Disp. sur : http://books.openedition.org/cdf/529

[ANR 18] AGENCE NATIONALE DE LA RECHERCHE, « Plan d'action 2019 », Paris, 2018. Disp. sur : http://www.agence-nationale-recherche.fr/financer-votre-projet/plan-d-action-2019/

[BRO 18] BROSSAUD C., « Conditions d'émergence et enjeux des communs scientifiques à partir d'une expérimentation lyonnaise », *tic&société*, vol. 12, n° 1, p. 201-228, 2018. Disp. sur : https://journals.openedition.org/ticetsociete/2435

[CNR 18] CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE, « Préfiguration du 9e programme cadre : contribution du CNRS », Paris, mars 2018. Disp. sur : http://www.cnrs.fr/sites/default/files/download-file/cnrs-fp9-vf.pdf

[COM 18a] COMMISSION EUROPEENNE, Direction générale des réseaux de communication, du contenu et des technologies, « Proposition de Directive du Parlement européen et du Conseil concernant la réutilisation des informations du secteur public (refonte) », Bruxelles, avril 2018, Pub. L. No. 2018/0111/COD (2018). Disp. sur : https://eur-lex.europa.eu/legal-content/FR/TXT/?qid=1540373138782&uri=CELEX:52018PC0234

[COM 18b] COMMISSION EUROPEENNE, « Recommandation (UE) 2018/790 de la Commission du 25 avril 2018 relative à l'accès aux informations scientifiques et à leur conservation », Bruxelles, avril 2018. Disp. sur : https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32018H0790&from=EN.

[COR 17] CORNU M., ORSI F., ROCHFELD J. (dir.), *Dictionnaire des biens communs*, Presses Universitaires de France, Paris, 2017.

[DEC 03] « Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales », Max Planck Gessellschaft, Berlin, 2003. Disp. sur : https://openaccess.mpg.de/68042/BerlinDeclaration_wsis_fr.pdf

[DIL 17] DILLAERTS H., « 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*, vol. 14, n° 1, p. 39-54, 2017. Disp. sur : https://www.cairn.info/revue-communication-et-management-2017-1-page-39.htm

[DIL 18] DILLAERTS H., « L'Open Access et les chercheurs au cœur d'injonctions paradoxales », Séminaire NUMEREV-MSH-SUD, Montpellier, 16 octobre 2018. Disp. sur : http://www.mshsud.tv/spip.php?article870

[FIN 14] FING (Fédération Internet Nouvelle Génération), BRUGIERE A., « Développer une culture des données au sein des organisations », Congrès *Société Informatique de France*, Poitiers, février 2014. Disp. sur: https://fr.slideshare.net/slidesharefing/dvelopper-la-mdiation-aux-donnes-dans-les-entreprises

[FLO 05] FLORIDI L., « Semantic Conceptions of Information », *Stanford Encyclopedia of Philosophy*, 2005. Disp. sur : http://plato.stanford.edu/entries/information-semantic

[FRA 17] FRANCE, « Position préliminaire de la France sur le 9ème PCRI - Note des autorités françaises », Paris, 2017. Disp. sur : http://cache.media.education.gouv.fr/file/2017/78/4/NAF-position_FR_FP9_859784.pdf

[FRA 18] FRANCE, Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, « Plan national pour la science ouverte », Paris, 2018. Disp. sur : http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf

[GAI 14] GAILLARD R., « De l'Open Data à l'Open Research Data : quelle(s) politique(s) pour les données de recherche ? », ENSSIB, Lyon, 2014. Disp. sur : https://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche

[GIN 14] GINGRAS Y., *Les dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie*, Raisons d'agir, Paris, 2014.

[LAT 01] LATOUR B., *L'espoir de Pandore. Pour une version réaliste de l'activité scientifique*, La Découverte, Paris, 2001.

[LEL 10] LELEU-MERVIEL S., « Le sens aux interstices, émergence de reliances complexes », Colloque international francophone *Complexité 2010*, Lille, 2010. Disp. sur : https://hal.archives-ouvertes.fr/hal-00526508

[MAU 16a] MAUREL L., « Les universités françaises et l'Open Data après la loi numérique », *S.I.Lex. Carnet de veille et de réflexion d'un juriste et bibliothécaire*, 2016. Disp. sur : https://scinfolex.com/2016/11/01/les-universites-francaises-et-lopen-data-apres-la-loi-numerique/

[MAU 16b] MAUREL L., « Quel statut pour les données de la recherche après la loi numérique ? », *S.I.Lex. Carnet de veille et de réflexion d'un juriste et bibliothécaire*, 2016. Disp. sur : https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/

[MIL 12] MILLERAND F., « La science en réseau. Les gestionnaires d'information "invisibles" dans la production d'une base de données scientifiques », *Revue d'anthropologie des connaissances*, vol. 6, n° 1, p. 163-190, 2012. Disp. sur : https://www.cairn.info/revue-anthropologie-des-connaissances-2012-1-page-163.html

[OCD 07] OCDE, « Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics », 2007, Disp. sur : http://www.oecd.org/fr/sti/sci-tech/principesetlignesdirectricesdelocdepour laccesauxdonneesdelarecherchefinanceesurfondspublics.htm

[PRO 15] PROST H., SCHÖPFEL J., *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3 : Rapport final*, Rapport de recherche, Université Lille 3, 2015. Disp. sur : http://hal.univ-lille3.fr/hal-01198379

[ROS 10] ROSA H., *Accélération. Une critique sociale du temps*, La Découverte, Paris, 2010.

[ROU 13] ROUVROY A., BERNS T., « Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation ? », *Réseaux*, vol. 1, n° 177, p. 163-196, 2013. Disp. sur : https://www.cairn.info/article.php?ID_ARTICLE=RES_177_0163

[SCH 18] SCHÖPFEL J., *Vers une culture de la donnée en SHS : Une étude à l'Université de Lille*, Rapport de recherche, Université de Lille, 2018. Disp. sur: https://hal.archives-ouvertes.fr/hal-01846849

[SER et al. 17] SERRES A., MALINGRE M.L., MIGNON M., PIERRE C., COLLET D., *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, novembre 2017. Disp. sur: https://hal.archives-ouvertes.fr/hal-01635186

[STE 17] STENGERS I., JAMES W., *Une autre science est possible ! Manifeste pour un ralentissement des sciences*, La Découverte, Paris, 2017.

[TEB 17] TEBOUL B., « Vers une philosophie de la donnée », *Paris Innovation Review*, 10 avril 2017. Disp. sur : http://parisinnovationreview.com/article/vers-une-philosophie-de-la-donnee

[VID 18] VIDAL F., « Discours de présentation du Plan national pour la science ouverte », Congrès LIBER, Lille, 4 juillet 2018. Disp. sur : http://m.enseignementsup-recherche.gouv.fr/cid132531/plan-national-pour-la-science-ouverte-discours-de-frederique-vidal.html

[VIT 18] VITALI-ROSATI M., « À quoi servent les publications scientifiques ? », *La vie de la recherche scientifique*, n° 412, p. 19-22, 2018. Disp. sur : https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/21016/PublicationsScientifiques_Vitali-Rosati.pdf?sequence=1&isAllowed=y

[ZIN 07] ZINS C., « Conceptual Approaches for Defining Data, Information, and Knowledge », *Journal of the American Society for Information Science and Technology*, vol. 58, n° 4, p. 479-493, 2007. Disp. sur : http://www.success.co.il/is/dik.html