Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation : étude de cas pour l'archive HAL

Marilou Pain

Mémoire de stage / septembre 2016

Diplôme national de master

Domaine - sciences humaines et sociales

Mention – sciences de l'information et des bibliothèques

Spécialité – sciences de l'information et des bibliothèques et information scientifique et technique

Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation : étude de cas pour l'archive HAL

Marilou Pain

Chérifa Boukacem-Zeghmouri
Maître de conférence en Sciences de l'Information et de la Communication – Université Lyon 1

Remerciements

Je remercie vivement Chérifa Boukacem pour son enthousiasme vis-à-vis de ce sujet, son accompagnement sans faille durant ce stage et également pendant la rédaction de ce mémoire.

Merci ensuite à Christine Berthaud et Bénédicte Kuntziger pour leur accueil au CCSD et leurs conseils avisés. Je remercie également Agnès, Baptiste, Bruno, Florence, Gala, Isabelle, Kévin, Maxime, Raphaël, Sarah et Yannick pour leur patience, leur présence et les réponses à mes nombreuses questions.

*Merci à Aaron Schwartz, Alexandra Elbakyan ainsi qu'à tous les défenseurs de l'*open access *sans qui ce travail n'aurait pas été possible.*

En dernier lieu, je tiens à remercier chaleureusement ma mère relectrice de toujours, mon père pour ses encouragements ainsi que mes colocataires pour leur soutien inconditionnel dans les moments de procrastination et particulièrement Mathias, toujours présent pour ajouter une couche de complexité à une réflexion.

Résumé :

L'archive ouverte nationale et pluridisciplinaire HAL héberge aujourd'hui des données de la recherche ainsi que des données supplémentaires sous la forme d'annexes. Afin de tenter de définir des orientations pour cette infrastructure, ce mémoire présente un état de l'art des différents acteurs et enjeux qui gravitent autour de la thématique des données de la recherche. Ensuite, il s'attache à décrire les différents services mis en œuvre par les entrepôts de données de la recherche ainsi que les défis auxquels ils doivent répondre. Enfin, est proposée une étude exploratoire des données supplémentaires hébergées par HAL, qui cherche à identifier quelles communautés scientifiques utilisent ce service et sous quelles formes.

Descripteurs : données de la recherche, libre accès, voie verte, entrepôt de données, entrepôt institutionnel, archive ouverte, information scientifique et technique

Abstract :

Currently, the French multidisciplinary open archive HAL hosts research data and supplementary materials as annexes. In an attempt to define guidelines for this infrastructure, this thesis presents a state of the art of the stakeholders and issues linked to research data. Then it attempts to describe the various services implemented by the research data repositories as well as the challenges they meet. Finally, it presents an exploratory study of the additional data hosted by HAL, which seeks to identify the scientific communities who are using this service.

Keywords : research data, open access, green road, data repository, institutional repository, open archive, scientific and technical information

Sommaire

Sigles et abréviations

AMUE	Agence de Mutualisation des Universités et des Établissements
ANR	Agence Nationale de la Recherche
BSN	Bibliothèque Scientifique Numérique
CC	Creative Commons
CCSD	Centre pour la Communication Scientifique Directe
CINES	Centre Informatique National de L'enseignement Supérieur
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CNRS	Centre National de la Recherche Scientifique
CODATA	International Council for Science : Committee on Data for Science and Technology
CORDONUM	COmité d'ORientation Numérique
COUPERIN	Consortium unifié des établissements universitaires et de recherche pour l'accès aux publications numériques
CSPLA	Conseil supérieur de la propriété littéraire et artistique
DADVSI	Loi relative au droit d'auteur et aux droits voisins dans la société de l'information
DARIAH-EU	Digital Research infrastructure for the Arts and humanities
DCC	Data Curation Center
DMP (PGD)	Data Management Plan (Plan de Gestion des Données)
DOI	Digital Object Identifier
EER	Espace Européen de la Recherche
HAL	Hyper Articles en Ligne
HDR	Habilitation à Diriger les Recherches
HEFCE	Higher Education Funding Council for England
IdRef	Identifiants et référentiels pour l'Enseignement supérieur et la Recherche
IFREMER	Institut Français de Recherche pour l'Exploitation de la Mer
INRA	Institut National de la Recherche Agronomique
INRIA	Institut National de Recherche en Informatique et en Automatique
INIST	Institut de l'Information Scientifique et Technique
INSERM	Institut National de la Santé et de la Recherche Médicale
IST	Information Scientifique et Technique
JISC	Joint Information Systems Comittee
LERU	League of European Research Universities
NOW	Netherlands Organisation for Scientific Research
OCDE	Organisation de Coopération et de Développement Économiques

ODbL	Open Database License
OpenAIRE	Open Access Infrastructure for Research in Europe
OpenDOAR	The Directory of Open Access Repositories
RNSR	Répertoire National des Structures de Recherche
ROAR	Registry of Open Access Repositories
SHS	Sciences Humaines et Sociales
STM	Science Technique Médical
TDM	Text and Data Mining
TGIR	Très Grande Infrastructure de recherche

INTRODUCTION

« Go forth and replicate![1] »

Le titre de cet éditorial de la revue *Nature* est une ode à la reproduction de la recherche. La production d'articles reproduisant ou réadaptant des recherches existantes est un domaine en émergence, qui manque de visibilité malgré le lancement de quelques revues dédiées. Tout comme la vérification approfondie des résultats de la recherche, cette démarche nécessite pour les auteurs l'ouverture de leurs dossiers de travail et le partage d'éléments de leur recherche dont leurs données, éléments qui restent traditionnellement dans les combles des ordinateurs.

L'écosystème de la publication scientifique semble ainsi tendre vers une diversification de ses objets où l'article, épuré de certains de ses composants, serait entouré d'éléments méthodologiques, d'actes de conférences, de vidéos ou encore de ses données, comme on peut le voir sur cette figure.

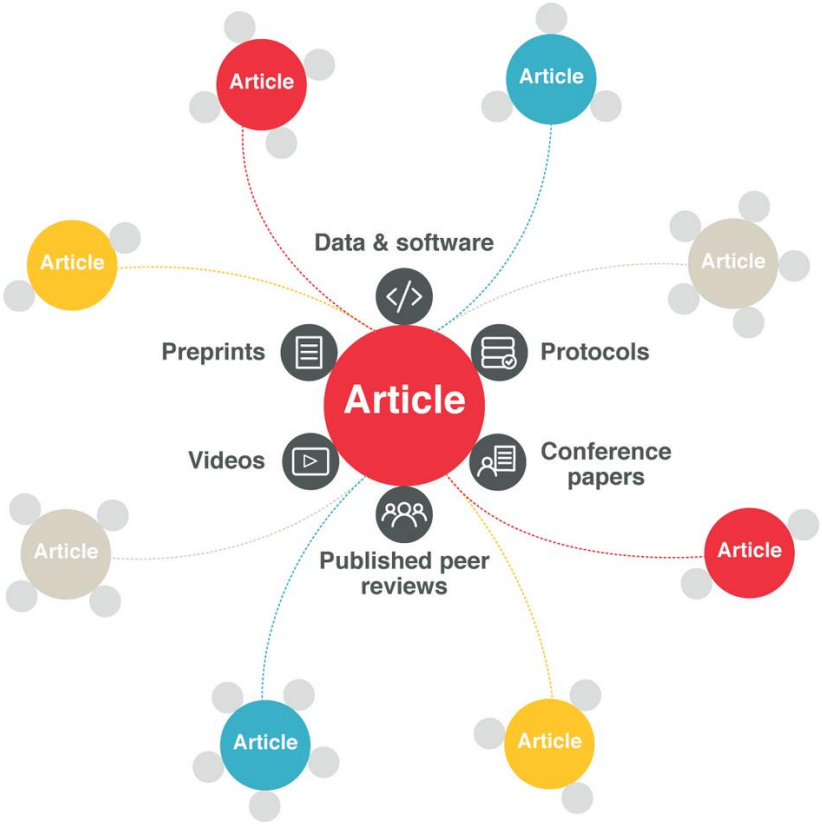


Figure 1 : Un nexus d'articles selon Crossref[2]

Avec cette image, Crossref présente l'élément fondateur de ces associations : l'identification unique des éléments et la description de leur relation, permettant leur référencement ainsi que leur connexion à d'autres publications. La création d'un écosystème de ce type semble ouvrir des perspectives enthousiasmantes pour

[1] « Go forth and replicate », *Nature* vol. 536 [En ligne]. 25 août 2016. Disponible sur http://www.nature.com/polopoly_fs/1.20473!/menu/main/topColumns/topLeftColumn/pdf/536373a.pdf

[2] Lin J., « The article nexus: linking publications to associated research outputs », The art of persistence, Crossref. [En ligne]. 25 août 2016. Disponible sur : http://blog.crossref.org/2016/08/the-article-nexus.html

la recherche, notamment si cela s'accompagne d'un effort de liaison des ressources et de partage de l'information. Ainsi, dans le contexte de l'*open science*, cette diversification pourrait pousser les scientifiques à partager des éléments qui étaient auparavant diffusés de façon très restreinte. En effet, la réutilisation ou même la simple lecture alimentant une réflexion sont conditionnées par l'accès aux matériaux de la recherche. Les données sont tout particulièrement concernées par ce mouvement d'ouverture de la science qui mobilise également des problématiques politiques et sociales, résumées par B. Fecher et S. Friesike[3] en cinq écoles de pensée[4].

Les éléments concernés seraient constitués de la société, qui suppose que la recherche doit être rendue au public. On peut rapprocher cette école de la seconde, celle dite « Démocratique », selon laquelle le savoir est distribué inégalement au lieu d'être librement accessible à tous. Une troisième école serait constituée des métriques cherchant des alternatives aux mesures d'évaluation actuelles. Viendrait ensuite l'école pragmatique, supposant que la création des savoirs pourrait être plus efficace si les chercheurs travaillaient ensemble.

Car jusqu'à présent, obtenir les données de recherche d'un collègue afin de vérifier, comparer ou reproduire un résultat s'avérait plus que complexe[5]. Le libre accès aux données[6] est un des éléments qui permettrait la reproductibilité des recherches scientifiques évoquées précédemment. Il s'agit d'un sujet brûlant dans la littérature de l'Information Scientifique et Technique, abordé dans des dizaines d'articles de revues, d'articles de blogs, de conférences ou encore de recommandations institutionnelles ou éditoriales. Ce sujet pose des questions épistémologiques épineuses comme celle de la définition de ce que peut être une donnée ou encore celle de l'influence qu'aura leur diffusion sur les paradigmes scientifiques. Si on peut établir la donnée de recherche comme un élément de la recherche ayant servi à établir des résultats dans notre contexte, leur définition, comme le dit C. Borgman, est un sujet de recherche à part entière d'une grande richesse épistémologique[7]. Le terme de données brutes, par exemple, est régulièrement remis en question, notamment du fait de la subjectivité inhérente à la collecte ou production des dites données et artefacts.

Au-delà de ces interrogations, la diffusion des données afin de permettre leur partage s'accompagne de défis techniques. Car pour référencer, stocker ou chercher ces données en contexte numérique, il faut créer de nouvelles plates-formes ou améliorer des plates-formes existantes. C'est pourquoi elles ont leur place comme cinquième école de pensée dans la vision de *l'open science* de B. Fecher et S. Friesike. Comme indiqué dans le schéma ci-dessous, elles ont pour rôle la création d'outils permettant l'accès à l'IST et, dans le cas présent, aux données de recherche[8].

---

[3] Fecher B., Friesike S. « Open Science: One Term, Five Schools of Thought ». In : Bartling S., Friesike S., *Opening Science*, [En ligne]. Springer International Publishing. 2014. Disponible sur : http://link.springer.com/10.1007/978-3-319-00026-8

[4] *Ibid.*, voir annexe 1 p. 72

[5] Voir à ce sujet la vidéo parodique créée par la NYU Health Science Library https://www.youtube.com/watch?v=N2zK3sAtr-4.

[6] Ou *Open Research Data*

[7] Borgman C. L. *Big data, little data, no data*. Cambridge, Massachusetts : The MIT Press, 2015. p. 19

[8] Bartling S., *op. cit.*

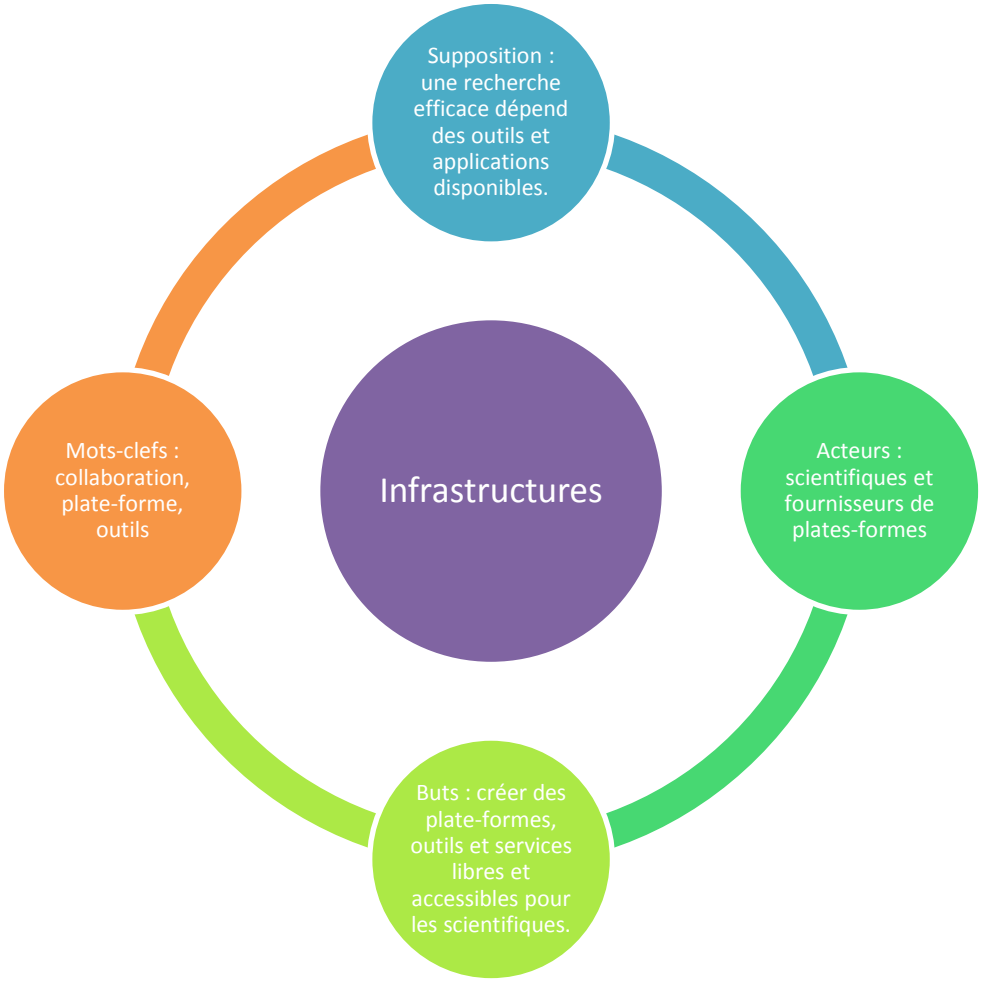


Figure 2 : Les infrastructures comme élément d'ouverture des sciences[9]

Ces infrastructures devront ainsi héberger des données de recherche, les référencer, les archiver, les diffuser et les partager. Ces services dépassent largement les capacités des bases de données locales des laboratoires et sont proposés par les entrepôts de données. Ceux-ci sont multiples, tant de manière quantitative que qualitative, leur tutelle peut ainsi être une entreprise à but lucratif, une organisation à but non lucratif ou encore une institution publique, ils peuvent avoir choisi un modèle en accès libre ou restreint. Ces entrepôts peuvent également faire partie d'une archive ouverte, rendant les données librement consultables et réutilisables.

L'empreinte du libre accès est ainsi particulièrement forte dans ce sujet. Car si l'*open access* s'est d'abord concentré sur les publications, plus aisées à appréhender et immédiatement visibles comme éléments primordiaux de la recherche, les données sont aujourd'hui au cœur de ce mouvement. La vision du partage des données de la recherche comme un bienfait incontestable pour la science, la société et l'économie[10], telle que la propose le Plan d'action

[9] Fecher B., 2014, *op. cit.*, p. 10

[10] « *Si toutes les publications scientifiques étaient consultables en ligne gratuitement et si les données de recherches étaient mieux utilisées, chacun pourrait en profiter au mieux et plus vite. Les médecins pourraient par exemple s'informer des dernières avancées en matière de traitement, les entrepreneurs seraient en mesure d'adapter plus rapidement leurs produits aux innovations scientifiques et les enseignants pourraient enrichir leurs cours à la lumière de la science.* » Source : http://francais.eu2016.nl/a-la-une/actualites/2016/04/05/plan-d%E2%80%99action-europeen-pour-la-science-ouverte

d'Amsterdam sur l'innovation en matière de science ouverte[11], semble néanmoins simpliste, gommant les nombreuses et parfois âpres négociations entre les différents protagonistes[12]. Car il s'agit bien d'un enjeu actuel, autour duquel de multiples acteurs trouvent des intérêts, parfois contradictoires, provoquant ainsi ruptures et mariages.

Le partage de l'Information Scientifique et Technique est un domaine que j'ai abordé en licence professionnelle à l'Université Rennes 2, à travers les problématiques *d'open access* en bibliothèque. Mes cours d'anthropologie des savoirs de première année de master à Lille 3ont renforcé l'intérêt que je pouvais y porter en y ajoutant une dimension historique et épistémologique. Mon stage d'archivistique au sein de l'Université de Nantes a alors plus que confirmé mon intérêt pour les échanges de connaissances liés au processus scientifique. C'est pourquoi les données de recherche m'ont parues une problématique intéressante pour ce stage de seconde année. Je me suis ainsi rendue à des rencontres, tout en démarchant des lieux de stage, afin de travailler sur cette thématique.

C'est le Centre pour la Communication Scientifique Directe qui m'a accueillie durant plusieurs mois afin d'interroger le thème des données de la recherche, aussi bien dans une perspective théorique, par le biais de la littérature, que dans une perspective pratique, en étudiant concrètement le contenu de l'archive ouverte HAL. En effet, même si le CCSD maintient plusieurs services à la recherche, mon activité a été essentiellement concentrée sur HAL. Les résultats intermédiaires de l'étude exploratoire ont été présentés à l'Université Humboldt de Berlin dans le cadre d'un séminaire, me permettant ainsi de découvrir le travail de présentation des résultats en langue étrangère, face à un public qui s'est avéré particulièrement intéressé par la démarche de HAL, et avec lequel j'ai pu interagir sur des questionnements communs.

Le fil conducteur de ce stage se constitue des enjeux liés aux données de la recherche et de la façon dont les entrepôts cherchent à y répondre. Les archives ouvertes se dotent aujourd'hui de services d'accueil des données de la recherche. Dans quelle mesure une archive nationale et multidisciplinaire telle que HAL doit-elle se doter de ce genre de service ?

La première partie de ce mémoire est vouée à la description des activités menées durant cette période de stage. Vient ensuite l'état de l'art réalisé sur les données de la recherche à partir de la littérature scientifique et professionnelle française et internationale, afin de dresser un portrait sémantique, institutionnel et scientifique de la question des données de recherche. La troisième partie développe l'étude exploratoire menée sur l'archive ouverte HAL ainsi que les préconisations émises suite à ces réflexions, l'ensemble permet de dégager des scénarios pour le CCSD dans cet écosystème de la recherche et de ses données.

[11] Zaken M. van B. « Amsterdam Call for Action on Open Science - Publicatie - EU2016.nl ». 2016. Disponible sur : http://www.eu2016.nl/documenten/rapporten/2016/04/04/amsterdam-call-for-action-on-open-science

[12] Chartron G. « Stratégie, politique et reformulation de l'open access », *Revue française des sciences de l'information et de la communication* [En ligne], 2016, mis en ligne le 24 mars 2016. Disponible sur : http://rfsic.revues.org/1836

ENVIRONNEMENT DE STAGE

1. LE CENTRE POUR LA COMMUNICATION SCIENTIFIQUE DIRECTE ET SES SERVICES

L'unité mixte de service CCSD a vu le jour en 2000 à l'initiative du CNRS et a depuis 2014 trois tutelles : le CNRS, l'Université de Lyon, qui représente le Ministère de l'Enseignement supérieur et de la Recherche, ainsi que l'INRIA. Dès le départ, l'objectif était la création d'une archive ouverte scientifique française, HAL, sur le modèle d'arXiv. L'équipe de celle-ci a d'ailleurs conseillé le CCSD lors de la mise en place de HAL. La problématique de départ était de concilier l'approche d'arXiv, marquée par la rigueur scientifique et l'archivage pérenne des articles, avec une approche plus institutionnelle, notamment en termes d'identification de la production scientifique. La dimension internationale de HAL a ainsi toujours été présente, se développant dans le temps, de par le reversement initial de ses dépôts à arXiv, puis à Pubmed Central à partir de 2005. De plus, l'archive est également moissonnée par RepEc et s'inscrit dans des projets internationaux tels que DARIAH-EU[13] ou encore OpenAIRE[14]. Le CCSD fait ainsi preuve d'une volonté d'interconnexion avec les bases existantes, selon les besoins formulés par les différentes communautés scientifiques utilisatrices de HAL.

Définie comme une « *archive ouverte pluridisciplinaire [...] destinée au dépôt et à la diffusion d'articles scientifiques de niveau recherche, publiés ou non, et de thèses, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.* »[15], elle promeut l'auto-archivage en encourageant les scientifiques à déposer eux-mêmes leurs publications. Celles-ci seront ensuite référencées par les bases de données administratives. Un partenariat avec le CINES permet l'archivage pérenne des données depuis 2006[16].

Quant à la teneur exacte de ces dépôts, elle doit être similaire à ce qu'on peut attendre d'une revue à comité de lecture. Ainsi, les publications déposées dans HAL répondent à une typologie variée, depuis l'article à la thèse, en passant par le rapport ou le document de travail. Une modération est effectuée par le CCSD. Il ne s'agit pas d'une évaluation scientifique, mais plutôt d'une vérification.

En effet, les métadonnées renseignées doivent décrire de façon exacte le document. Un soin particulier est accordé aux auteurs, disciplines, revues et structures de recherche (institutions, laboratoires etc.) ainsi qu'aux projets ANR et projets européens. Ces données sont indexées dans les référentiels AURéHAL, afin d'assurer une exploitation optimale des données. Par exemple, le référencement des disciplines Physique et Mathématiques est un miroir des disciplines référencées dans arXiv et permet ainsi de reverser des dépôts d'une archive à l'autre plus aisément. D'ailleurs, les disciplines représentées dans HAL sont nombreuses. Car l'archive s'est toujours positionnée comme multidisciplinaire, même si sa création a

[13] http://www.dariah.eu/

[14] https://www.openaire.eu/

[15] https://hal.archives-ouvertes.fr/

[16] « Vos fichiers lisibles dans 30 ans ? », blog du CCSD. [En ligne] 31 janvier 2013. Disponible sur : https://blog.ccsd.cnrs.fr/2013/01/vos-fichiers-lisibles-dans-30-ans/

été lancée par le physicien Franck Laloë et développée par des informaticiens de la physique.

La convention inter-établissements ou plus exactement la « Convention de partenariat en faveur des archives ouvertes et de la plate-forme mutualisée HAL » a été signée le 2 avril 2013 à l'Académie des sciences par L'AMUE, la CPU (Conférence des Présidents d'Universités), la CGE (Conférence des Grandes Écoles) ainsi que vingt-deux établissements : l'ANDRA, l'ANR, la BNF, Le BRGM, la CDEFI, le CEA, le CEE, le CIRAD, le CNRS, le CSTB, l'IFPEN, l'IFFSTAR, l'INED, l'INERIS, l'INRA, l'INRIA, l'INSERM, l'INVS, l'IRD, l'IRSN, l'IRSTEA et l'Institut Pasteur[17]. Cet aboutissement du travail de la BSN[18] conforte alors HAL comme une véritable infrastructure nationale, qui hébergera d'autres archives institutionnelles ainsi que leur contenu. Ces établissements ont ainsi le choix entre le reversement dans HAL du contenu de leur archive ouverte[19], ou bien la création d'un portail d'établissement, sous-ensemble de HAL permettant d'identifier l'ensemble de leur production scientifique grâce aux métadonnées d'affiliation et au référentiel des structures de recherche.

Cependant, le rapport de Serge Bauin, publié en 2014 peu avant la version 3 de HAL, pointe plusieurs difficultés auxquelles l'archive doit faire face et esquisse des pistes d'amélioration. En effet, l'archive fonctionne avec un « *sous-effectif chronique* »[20], notamment pour une structure à vocation nationale. C'est toujours le cas puisque le CCSD comptabilise aujourd'hui quatorze équivalents temps pleins dont sept en contrat à durée déterminée. Néanmoins, l'inscription de HAL à la feuille de route des infrastructures de recherche du Ministère[21] s'avère encourageante à ce propos.

Le rapport attire aussi l'attention sur le travail à réaliser pour faciliter l'utilisation de l'archive, on peut noter l'importance à ce sujet du groupe utilisateur constitué, entre autres, des administrateurs de HAL mais aussi de la plate-forme de support utilisateur qui permet de recueillir de nombreux avis et conseils. D'autant plus que HAL n'est pas le seul service maintenu par le CCSD. Il existe ainsi Episciences[22] qui héberge et organise le *peer-reviewing* de revues en libre-accès, alimentées par des articles déposés dans des archives ouvertes et non publiés par ailleurs. Le CCSD maintient de même la plate-forme d'organisation de conférences, colloques et *workshops* SciencesConf[23]. L'attention accordée à la facilité d'utilisation doit aussi stimuler les échanges entre ces différentes plates-formes et créer une véritable synergie entre les différents services offerts par l'unité[24].

---

[17] Convention de partenariat en faveur des archives ouvertes et de la plateforme mutualisée HAL. [En ligne] 02 avril 2013. Disponible sur : http://cache.media.enseignementsup-recherche.gouv.fr/file/HAL/93/3/01_Convention_HAL_246933.pdf

[18] http://www.bibliothequescientifiquenumerique.fr/

[19] Une solution choisie par exemple à Sciences-Po : https://spire.sciencespo.fr/web/

[20] Bauin S., « L'*open access* à moyen terme : une feuille de route pour HAL ». DIST/CNRS. [En ligne] septembre 2014. Disponible sur : http://www.enssib.fr/bibliotheque-numerique/notices/64775-l-open-access-a-moyen-terme-une-feuille-de-route-pour-hal-hyper-articles-en-ligne

[21] « HAL dans la feuille de route des infrastructures de recherche du Ministère », blog du CCSD. [En ligne] 07 avril 2016. Disponible sur : https://blog.ccsd.cnrs.fr/2016/04/hal-dans-la-feuille-de-route-des-infrastructures-de-recherche-du-ministere/

[22] http://episciences.org/

[23] https://www.sciencesconf.org/

[24] Bauin S., *op. cit.*

En tant qu'archive ouverte en auto-archivage, HAL s'adresse en même temps à des institutions et à des communautés de recherche. Serge Bauin encourage la séparation, nominative et fonctionnelle, des services proposés par HAL : communication entre chercheurs, fonction bibliographique et archivage. Il s'interroge également sur l'existence des notices bibliographiques qu'il qualifie de « métadonnées sans les données » et dont la présence dans une archive ouverte s'apparenterait peut être plus à de la gestion administrative qu'à de l'*open access*. De même, la possibilité de définir un embargo sur certains dépôts peut être intéressante pour certaines disciplines mais devrait selon lui s'accompagner de la possibilité de demander des tirés-à-part[25], ce qui est possible aujourd'hui pour les portails.

Cette approche utilisateur ainsi que l'animation du réseau de professionnels sont inscrites dans la feuille de route 2016-2020[26] du CCSD, tout comme la nécessité de positionner AURéHAL dans le réseau des référentiels en cours de construction au niveau national (IdRef, RNSR, …). Cette feuille de route a été construite avec l'aide du Comité Scientifique et Technique et fait état d'une volonté de prise en compte des données de recherche, introduites par les recommandations du CORDONUM. Le quatrième objectif de cette feuille de route est ainsi d'étendre HAL et ses services aux données de la recherche.

2. SUPPORT UTILISATEUR ET ALIMENTATION DE LA BASE DE CONNAISSANCES

C'est pourquoi un des objectifs de ce stage était la rédaction d'un état de l'art autour des données de la recherche et de leurs entrepôts. Celui-ci a été produit grâce à une étude de la littérature internationale portant sur les données de la recherche. Les références récoltées lors des recherches bibliographiques seront à terme partagées en interne mais également de façon publique.

En tant qu'assistante documentaliste pour le CCSD, une de mes activités a été la pratique du support utilisateur au sein de l'équipe pôle support, constituée de quatre personnes pour l'ensemble des services du CCSD[27]. Cette équipe traite via une plate-forme spécifique les demandes des utilisateurs souhaitant des évolutions ou rencontrant des difficultés dans l'utilisation des services. Ces demandes ou tickets se présentent sous la forme d'échanges par mail ou d'échanges téléphoniques. Sur un mois, j'estime le traitement des demandes concernant HAL à environ vingt par jour ouvré et par équivalent temps plein.

Une autre de mes missions était la capitalisation du savoir empirique des informaticiens et documentalistes travaillant sur HAL afin de le retranscrire dans une base de connaissances interne de type wiki. En effet, les savoirs-faire de l'équipe du CCSD étaient pour la plupart des connaissances tacites ou singulières, qui n'avaient pas encore été partagées formellement de manière collective.

Concernant les données de la recherche, quatre demandes recherche ont été effectuées sur le support entre avril et juillet 2016. Deux d'entre elles étaient le fait d'un chercheur isolé souhaitant déposer ses données dans HAL tout en

[25] Bauin S., *op. cit.*, p. 14

[26] « Feuille de route du CCSD 2016-2020 », Blog du CCSD. [En ligne]. 20 juin 2016. Disponible sur : https://blog.ccsd.cnrs.fr/2016/06/feuille-de-route-du-ccsd-2016-2020/

[27] https://www.ccsd.cnrs.fr/pdf/organigramme.pdf

s'interrogeant sur le meilleur moyen de le faire et les formats de fichiers acceptés par HAL. La question était alors d'ordre technique.

Les deux autres demandes, dont l'une d'entre elles a également été relayée sur la liste de diffusion halinfo[28], étaient d'ordre institutionnel, l'enjeu était de savoir si HAL allait développer une politique vis-à-vis des données de la cherche. Elles étaient le fait d'individus représentant leur institution et cherchant à s'informer sur la politique du CCSD et les éventuels développements de HAL. La sollicitation éventuelle de HAL par des chercheurs, les formats que HAL pourrait accepter mais surtout la possibilité de lier les publications réalisées dans HAL avec des données déposées ailleurs étaient les principales interrogations formulées.

La rédaction de réponses aux demandes des utilisateurs de même que la réalisation de procédures de support internes m'ont offert une connaissance approfondie de HAL ainsi qu'un lien direct avec ses différents publics : chercheurs, doctorants, documentalistes, bibliothécaires ou encore informaticiens.. Ce contact avec leurs besoins et leurs attentes, en particulier en matière de données a contribué à l'alimentation de cette étude.

[28] Liste de diffusion des utilisateurs de HAL.

PARTIE 1 : ÉTAT DE L'ART : LES DONNEES DE LA RECHERCHE ET LEURS ENJEUX

1. DEFINITIONS

1.1. Les données de la recherche

On ne peut réaliser un référencement effectif des définitions des données de la recherche, d'une part parce qu'il est difficile de définir quelque chose qui est différent d'un domaine scientifique à l'autre, voire d'un objet de recherche à un autre ; d'autre part parce que cette thématique reste innovante, les consensus de définition n'ont donc pas encore été établis.

On peut postuler qu'une donnée est la description élémentaire d'une réalité ou d'un fait qui sert de point d'appui à un raisonnement ou à une recherche. Dans le contexte de la recherche scientifique, les données de recherche ont étés définies par le Bureau de la gestion et du budget du gouvernement fédéral américain (*Office of Management and budget*) dans une circulaire amandée en 1999 :

> *« Les données de la recherche sont des enregistrements factuels (chiffres, textes, images et sons) utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires à la validation des résultats de recherche »*[29]

On retrouve généralement cette notion de validation des résultats dans les multiples définitions des données de recherche produites par les universités, organismes de recherche et professionnels de l'IST. Ainsi, pour l'université d'Edinburgh :

> *« Les données de la recherche, à la différence des autres types d'information, sont collectées, observée ou créées à des fins d'analyse pour produire des résultats de recherche inédits. »*[30]

L'université de Bristol précise quant à elle que les données de recherche doivent être :

> *« communicables, interprétables et adaptées à un traitement souvent informatisé. »*[31]

Le Livre blanc du CNRS réalisé en 2016 établit les données de la recherche comme « Bien Commun » nécessaire à la recherche, notamment par le biais du TDM[32]. Il reprend ces différentes définitions et décline les données de l'IST en

[29] OMB, Circular A-110 (Uniform Administrative Requirements for grants and agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations [En ligne]. Office of Management and Budget. Disponible sur https://www.whitehouse.gov/omb/circulars_a110. Citée dans Gaillard R. « De l'Open data à l'Open research data: quelle (s) politique (s) pour les données de recherche? » [En ligne]. Enssib, 2014. Disponible sur <http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

[30] Edinburgh University Data Library Research Data Management Handbook, Edinburgh University Library. [En ligne]. 2011. Disponible sur : http://www.docs.is.ed.ac.uk/docs/data-library/EUDL_RDM_Handbook.pdf

[31] Gaillard R., 2014, *op. cit.*

[32] Direction de l'Information Scientifique et Technique - CNRS. Livre blanc — Une Science ouverte dans une République numérique [En ligne]. Marseille : OpenEdition Press, 2016. (Laboratoire d'idées). Disponible sur : http://books.openedition.org/oep/1548 p. 32

deux catégories : les publications (résultats de la recherche) et les données de recherche, celles-ci étant produites durant le processus de recherche et "*ayant servi à établir ces résultats*". On pourrait ajouter à ces deux catégories d'information les données méthodologiques, parfois séparées des publications.

Il est possible de différencier les données de recherche selon plusieurs typologies. La première serait de les distinguer selon le stade de leur cycle de vie, représenté dans la figure ci-dessous. Ainsi, on aurait les données préliminaires ou données préparatoires (exclues dans un contexte de diffusion) ; les données brutes (données acquises lors du processus de recherche et déjà potentiellement traitées) ; et enfin les données traitées et analysées : c'est-à-dire ayant subi une transformation telle qu'il n'est plus possible d'accéder aux données brutes, un graphique par exemple.

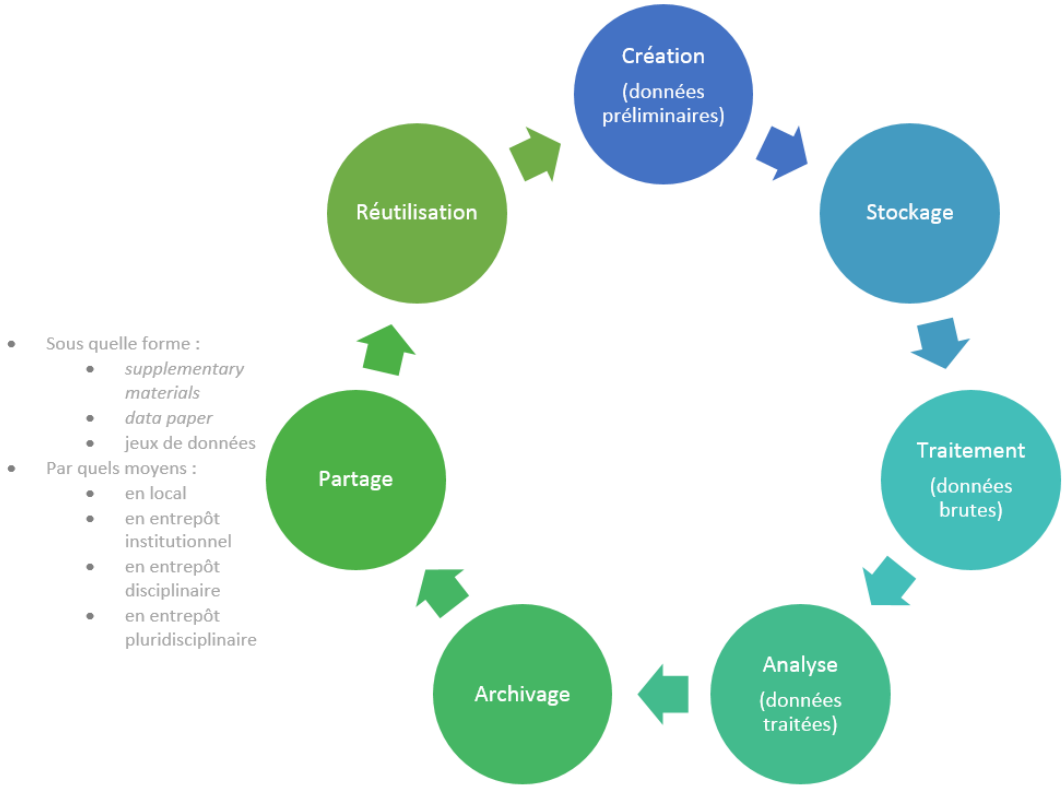


Figure 3 : Le cycle de vie des données de recherche[33]

Ensuite, on peut chercher à distinguer les données de recherche selon leur méthode d'obtention[34]. On a alors d'une part les données d'observation, relatives à un objet. Ces données d'observation peuvent elles-mêmes être descriptives, c'est-à-dire uniques et collectées par des individus ou des machines, ou bien expérimentales, c'est-à-dire récoltées en changeant les conditions de réalisation d'un phénomène pour en étudier les étapes.

D'autre part, on trouve les données traitées ou dérivées qui se composent de données retravaillées pour une meilleure compréhension ou une normalisation ou

[33] Thessen A. E., Patterson D. J. « Data issues in the life sciences ». Zookeys [En ligne]. 28 novembre 2011.n°150, p. 15-51. Disponible sur : http://dx.doi.org/10.3897/zookeys.150.1766

[34] Gaillard R., 2014, *Op. cit.*, p.17

bien des données de simulation issues de combinaison de modèles mathématiques ou informatiques.

Enfin, il semble vain de tenter de distinguer les données de recherche selon leur nature physique : leur étendue est infinie car variant selon le paradigme choisi. Chaque communauté scientifique aura ainsi ses propres limites et sa propre définition, selon ses méthodes, ses objets d'étude, etc. Cependant, dans le cadre de cette étude portant en partie sur les entrepôts digitaux de données, on peut caractériser les données de recherche comme des objets numériques.

Elles peuvent être rassemblées sous la forme de jeux de données ou *data sets*[35], c'est à dire des collections, le plus souvent numériques et liées par un thème ou une catégorie. Le sujet du jeu de données reflète alors ce qui est à l'étude dans la recherche. Ces données et jeux de données seront décrits par des métadonnées puis stockés, référencés et diffusées dans des entrepôts ou archives scientifiques. Les nombreux infrastructures et standards environnant ces données provoquent la création de nouveaux métiers (*data scientist, data librarian, data curator)* et de savoirs-faire informatiques, mais également documentaire, qui naviguent entre le théorique et l'empirique.

1.2. Les entrepôts de données

Mis en place dans les années 1990, une myriade d'entrepôts de données en *open access* est aujourd'hui à la disposition des chercheurs. Le Registry of Open Access Repository (ROAR)[36] comptabilise dans le monde vingt-neuf entrepôts dont le type de dépôt est « *Research data* », ce qui semble particulièrement sous-estimé. OpenDOAR[37] a choisi une approche différente puisque les entrepôts ne peuvent y être que multidisciplinaires, disciplinaires, gouvernementaux ou institutionnels. C'est donc par le type de contenu que ce répertoire fait la distinction entre entrepôt classique et entrepôt de données. Le type de contenu « *Datasets* » permet de trouver cent soixante-huit entrepôts, ce qui semble encore fort peu.

Le Re3data, en revanche, est consacré entièrement aux données de recherche et recensait plus de mille cinq-cents entrepôts en avril 2016. Ces chiffres doivent cependant être considérés avec précaution, ce site fonctionnant sur la base de suggestions. Les détenteurs ou utilisateurs d'entrepôts de recherche n'ont pas tous connaissance de l'existence de ce répertoire. Il n'est donc pas représentatif de l'ensemble des entrepôts de données de recherche au niveau mondial. Il est fort probable que ce chiffre soit également sous-dimensionné. Cependant, ces différences pourraient également être représentatives de la diversité des types d'entrepôts existants.

Le projet européen OpenAIRE moissonne des entrepôts en Open Access, des archives de la recherche ainsi que des journaux supportant l'Open Access. Il s'agit d'un dispositif H2020 visant à accompagner les politiques de dépôt en accès libre de la Commission Européenne et du Conseil Européen de la Recherche (ERC). Il définit un entrepôt de données de recherche de la façon suivante.

[35] http://dictionary.casrai.org/Dataset

[36] http://roar.eprints.org/

[37] http://www.opendoar.org/index.html

« Un entrepôt de données est une archive numérique qui collecte et diffuse des jeux de données et leurs métadonnées. Un grand nombre d'entrepôts de données acceptent également des publications et permet de lier les publications afférentes. »[38]

Un entrepôt dispose donc d'une notion d'archivage pérenne de ses données. De plus, il peut être constitué de plusieurs entrepôts, qu'il moissonne. Dans le contexte de *l'open access* son qualificatif « ouvert » fait qu'il doit pouvoir être sujet à des requêtes[39] afin d'accompagner la recherche. La différence entre une archive ouverte telle que HAL et un entrepôt pourrait se situer, en plus de l'aspect voie verte, au niveau des autres services proposés, tels que la mise à disposition de CV pour les chercheurs par exemple.

Les entrepôts de données de recherche conditionnent les dépôts selon trois modèles (avec parfois des frais de publication). Le premier consiste à accepter tous les dépôts, c'est notamment le cas de Zenodo ou Figshare. Ils peuvent également choisir de n'accepter des jeux de données que s'ils sont en lien avec une publication, comme dans le modèle choisi par Dryad[40]. Enfin, ils peuvent accepter uniquement certains types de données, en se concentrant sur une discipline, un domaine de recherche ou encore un projet particulier.

D'après l'étude commanditée par Wiley en 2014, 26% des données déposées le seraient dans un entrepôt institutionnel, 19% dans un entrepôt disciplinaire et seulement 6% dans des entrepôts généralistes tels que Figshare ou Dryad[41].

On peut distinguer les entrepôts institutionnels, thématiques, disciplinaires ou multidisciplinaires. Certains entrepôts sont à la fois disciplinaires et liés à des projets. C'est notamment le cas de PANGAEA[42]. Dans le cadre des entrepôts disciplinaires, l'avancement des disciplines à ce sujet est disparate. La problématique est en développement pour les SHS, la Physique, la Chimie, les Mathématiques et l'Informatique.

Certaines sous-disciplines ou inter-disciplines ont cependant des entrepôts bien identifiés, comme le COD[43] pour la cristallographie qui a également développé un format de métadonnées et des modèles.

Les sciences de l'univers ont quant à elles instauré une période d'embargo d'un an avant de diffuser leurs données. L'International Virtual Observatory Alliance (IVOA) regroupe des observatoires afin de définir des standards pour ces données[44].

---

[38] https://www.openaire.eu/support/faq

[39] Austin C. C. et al. « Key components of data publishing: using current best practices to develop a reference model for data publishing ». International Journal on Digital Libraries [En ligne]. 20 juin 2016. Disponible sur : http://dx.doi.org/10.1007/s00799-016-0178-2

[40] Il est indiqué sur la page d'accueil que les jeux de données doivent illustrer des publications librement réutilisables, citables et référencées. Source : http://datadryad.org/

[41] Ferguson L., « How and why researchers share data (and why they don't) » Wiley Exchanges. [En ligne]. 03 novrembre 2014. Disponible sur : http://hub.wiley.com/community/exchanges/discover/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont?referrer=exchanges

[42] https://www.pangaea.de/

[43] Crystallography Open Database : http://www.crystallography.net/cod/

[44] Dans lequel le Centre de Données astronomiques de Strasbourg http://cdsweb.u-strasbg.fr/index-fr.gml est impliqué par exemple.

En Biologie certaines revues comme *Nature* conditionnent la publication d'article par la diffusion des données de recherche liées[45]. Ce critère est probablement rendu possible grâce aux nombreux entrepôts spécifiques à la biologie. La revue peut ainsi imposer des entrepôts par domaine de recherche spécifique[46] ou bien, à défaut, des entrepôts généralistes comme Dryad. La recherche biomédicale fait quant à elle face à la réticence des éditeurs ainsi qu'à des réserves vis-à-vis de l'anonymisation des patients mais peut tout de même procéder au dépôt de données. Par exemple, le UK le National Institute of Health dépose dans PubMed Central.

Ces disciplines ou inter-disciplines ont ainsi une réelle culture des données de recherche, développant ces problématiques depuis plusieurs années. La littérature produite par les professionnels de la recherche et de l'IST au sujet des données de la recherche et de leurs entrepôts est ainsi en pleine ébullition. On constate une multiplication des journées d'études, *workshops* et séminaires[47] mais aussi le développement d'outils d'accompagnement, notamment à propos des plans de gestion des données[48]. Les professionnels de l'IST disposent en effet de compétences en matière de gestion de l'information qui leur font voir les données de la recherche comme un nouvel objet documentaire et comme l'occasion de jouer leur rôle de médiateur entre les multiples acteurs de cette thématique[49]. Les bibliothèques par exemple ont été signataires des principales déclarations du mouvement *open access* (Berlin, Budapest, Bethesda) et cherchent actuellement à se renouveler afin de faire partie du continuum de la recherche, de son processus de valeur[50].

2. ENJEUX POLITIQUES, ORGANISATIONNELS ET SOCIETAUX DES DONNEES DE LA RECHERCHE

2.1. Les acteurs des données de la recherche

Si on peut distinguer plusieurs acteurs prenant place à différents moments du cycle de vie des données de recherche, les chercheurs et communautés scientifiques sont au centre de cette problématique à la fois en tant que producteurs et en tant qu'utilisateurs. Les acteurs de cet écosystème sont les suivants :

- chercheurs et communautés scientifiques ;

[45] « *A condition of publication in a Nature journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.* » Source : http://www.nature.com/authors/policies/availability.html

[46] Voir annexe 2, p. 73

[47] Par exemple la journée d'étude « *Données de la recherche: enjeux, pratiques et services* » qui a eu lieu à Nice en juin 2016 https://frama.link/DonneesNiceJuin2016 ou encore la journée d'étude « *Données de la recherche: quel rôle pour la documentation ?* » organisée lors du 43ème congrès de l'ADBU en 2013 dont les vidéos sont en ligne http://www.canalc2.tv/video/12164.

[48] Voir par exemple les tutoriels de l'INIST www.inist.fr/?Donnees-de-la-recherche et notamment celui portant sur le libre accès aux résultats de recherche dans le cadre du projet H2020. Ou encore le page Web de la NASA consacrée aux PGD http://www.nasa.gov/open/researchaccess/data-mgmt

[49] Délémontez R., Boukacem-Zeghmouri C. « Données de la recherche: entre discours, réalités et valeur ». *I2D–Information, données & documents* [En ligne]. 2015. Vol. 53, n°4, p. 56–57. Disponible sur : http://www.cairn.info/resume.php?ID_ARTICLE=I2D_154_0056

[50] Boukacem-Zeghmouri C. *Mutations dans la sous-filière de la revue scientifique dans les domaines STM : une analyse par les industries culturelles* [en ligne]. 2015. Disponible sur : http://archivesic.ccsd.cnrs.fr/tel-01281524

- financeurs (par exemple l'ANR dans le cas français ou H2020 pour l'Europe) : une agence de financement peut être un état, un ministère de tutelle, une agence d'évaluation, une institution supranationale (OCDE) ;
- établissement d'enseignement supérieur & de recherche (Universités, Grandes Écoles, EPST...) ;
- éditeurs (Elsevier, Springer ou encore Wiley, la problématique des données de recherche concerne plutôt le pan STM de ces groupes éditoriaux pour l'instant) ;
- nouveaux intermédiaires (entreprises privées industrielles ou start-up se positionnant sur l'IST)[51] ;
- médiateurs : bibliothèques et centres de documentation académiques qui, face à la réorganisation de l'économie numérique, essayent de se positionner en faisant évoluer leurs missions.

Le sujet des données de la recherche semble avant tout avoir été lancé par les agences de financement et les éditeurs, de plus en plus explicites quant à leurs attentes pour ces matériaux[52]. Ces acteurs, privés ou publics, conditionnent à la fois la pérennité économique et la bonne diffusion des recherches.

2.1.1. Une incitation à déposer de plus en plus vive de la part des financeurs et tutelles

L'Union Européenne a rendu obligatoire, dans le programme européen pour la recherche et l'innovation H2020[53], la diffusion des données de recherche nécessaire aux résultats pour les projets validés par le programme et sur la base du volontariat[54]. Les données doivent ainsi s'accompagner d'un plan de gestion des données, précisant ce que le bénéficiaire d'H2020 a choisi de diffuser en justifiant les exclusions. Le livrable est ainsi constitué de la publication, de ses données et du plan de gestion des données (PGD)[55]. Cet exemple récent est représentatif de l'évolution des demandes des agences de financement en matière de données de la recherche.

En France, l'ANR ne semble pas avoir encore développé de recommandations particulières à ce sujet. Cependant, les organismes de recherche français se penchent sur le sujet, comme on a pu le voir avec le Livre Blanc du CNRS. L'INRA[56] et l'INRIA, par exemple, ont organisé des journées d'études et des formations sur le sujet. Des dispositifs d'accompagnement des déposants ont également été proposés.

---

[51] Boukacem-Zeghmouri C. « Nouveaux intermédiaires de l'information, nouvelles logiques de captation de la valeur ». *I2D–Information, données & documents* [En ligne]. 2015. Vol. 53, n°4, p. 34–35. Disponible sur : http://www.cairn.info/resume.php?ID_ARTICLE=I2D_154_0034

[52] CODATA-ICSTI Task Group on Data Citation Standards and Practices. « Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data ». Data Science Journal [En ligne]. 8 septembre 2013. Vol. 12, n°0. Disponible sur : http://dx.doi.org/10.2481/dsj.OSOM13-043

[53] Horizon 2020 : http://www.horizon2020.gouv.fr/

[54] Seuls certains domaines font partie de ce projet pilote : http://www.sussex.ac.uk/library/researchdatamanagement/create/biddingforfunding/horizon2020

[55] Ou *data management plan* (DMP) en anglais

[56] Blanc I., Gaspin C., Hologne O., Partage des données de larechercheSéminaire de lancement de la mise en œuvre de la politique Inra, INRA. [En ligne]. 2013. Disponible sur : https://www6.inra.fr/reseau-in-ovive/content/download/3209/32325/version/1/file/Accueil_atelier_IN-OVIVE.pdf

Pour ce qui est des universités, elles sont nombreuses à s'être dotées de recommandations et de programmes portant sur les données de recherche dans le monde, parfois doublées de politiques nationales. Ces politiques peuvent parfois viser les données administratives de ces institutions plus que les données produites par leurs scientifiques. D'ailleurs, la réutilisation des données administratives pour la recherche n'est pas forcément approfondie[57]. Nous nous intéresserons ici aux politiques orientées strictement vers la recherche, en lien avec les politiques d'*open access* et non d'*open data* administrative.

L'Université chinoise de Hong-Kong, par exemple, héberge l'Universities Service Centre for China Studies et le Databank for Chinese Studies[58], où les scientifiques peuvent déposer leurs données de recherche et les partager. Il ne s'agit pas ici de données administratives. En revanche, ce partage semble relatif :

> *« It is forbidden to duplicate, transfer, or export datasets to third parties; duplication is restricted to the user's personal use for the purposes of backup or statistical analysis.»*[59]

Ainsi, si cet entrepôt permet bien le partage de données de la recherche entre scientifiques, il conditionne fortement la réutilisation des données et ne s'inscrit pas dans une perspective d'*open access.*

À l'inverse, les universités situées aux États-Unis doivent satisfaire les conditions de la circulaire de l'OMB citée précédemment. Chaque institution doit conserver tous les documents produits dans le cadre d'un financement, dont les données de recherche[60]. Et l'OMB ouvre une porte certaine sur la réutilisation des données grâce à ce paragraphe :

> *« The Federal awarding agency(ies) reserve a royalty-free, nonexclusive and irrevocable right to reproduce, publish, or otherwise use the work for Federal purposes, and to authorize others to do so. »* [61]

Ajoutons que les politiques d'*open access* nationales, dans le contexte actuel, se dotent de volets sur les données de la recherche. Plusieurs pays ont également des institutions de référence établies à ce sujet. Ainsi, une interconnexion représentative de l'écosystème de la recherche peut être constatée entre politique universitaire et mise en œuvre d'outils nationaux.

L'Australie a ainsi créé en 2008 l'Australian National Data Service[62], une collaboration entre l'Université Nationale Australienne, la Monash University et le CSIRO, qui permet aux chercheurs australiens de référencer leurs données de recherche. Elles sont ensuite indexées par les moteurs de recherche et potentiellement réutilisées. Il s'agit donc d'un référencement des données, avec des liens vers leur plate-forme de stockage.

Le Royaume-Uni est aujourd'hui en train de suivre cette même voie. Le JISC[63], agence de conseil pour la recherche anglaise à but non lucratif, très active

---

[57] Thessen A., 2011, *op. cit.*, p. 18

[58] http://www.usc.cuhk.edu.hk/Eng/AboutDCS.aspx

[59] Databank for Chinese Studies, « Data usage policies » : http://www.usc.cuhk.edu.hk/Eng/UsagePolicy.aspx

[60] Thessen A., 2011, *op. cit.*, p. 18

[61] OMB, 1999, *op. cit.*, p. 17

[62] http://www.ands.org.au/, maintenu par le Research Data Australia https://researchdata.ands.org.au/

[63] Joint Information Systems Committee : https://www.jisc.ac.uk/

sur la question des données de recherche, y a créé le Digital Curation Center[64], situé à Edinburgh. Cette même agence annonçait ainsi en 2015 le projet Research Data Recovery Service[65], mis en place en 2013 par le UK data service[66] et le DCC. Il s'agit également d'un catalogue de données renvoyant vers les sites hébergeurs. L'objectif pour l'instant est « d'agréger les métadonnées des données de recherche conservées dans les universités du Royaume-Uni et dans les *data centers* nationaux et disciplinaires »[67]. À terme, les institutions envisagent un hébergement de données par ce service, en cas de défaillance pour une discipline par exemple.

Les universités du Royaume-Uni travaillent en profondeur le sujet des données de la recherche. Elles ont fait partie des premières universités à éditer des recommandations en termes de gestion de données et à proposer des services d'hébergement et de diffusion de données en libre accès[68]. On peut citer par exemple Edinburgh, dont la politique de gestion des données de la recherche a été validée en 2011[69], mais aussi Oxford[70], Exeter[71] ou encore Bristol[72]. Elles restent particulièrement actives puisqu'elles ont diffusé le Concordat on Open Research Data[73] le 28 juillet 2016, avec le HEFCE[74], Research Councils UK[75] et l'agence de financement Wellcome[76]. Ce document cherche à s'assurer de la bonne accessibilité des données de recherche générées par la recherche anglaise, il demande ainsi la justification de toute rétention de données, que le chercheur producteur des données ait la priorité en termes de réutilisation et rappelle qu'un soin particulier doit être accordé à l'aspect légal et éthique des données.

Le Royaume-Uni est donc parmi les pays les plus avancés en matière de politique de données de la recherche, tant dans ses organismes de recherche qu'au niveau national. On peut d'ailleurs noter qu'il s'est doté dès le 1er juin 2014 d'une exception au droit d'auteur pour la fouille de texte et de données[77].

Les Pays-Bas ont également une politique des données qui s'affirme d'année en année. L'Organisation néerlandaise pour la recherche scientifique (NWO) annonçait fin juillet 2016 que tous les appels à projet qu'elle lancera à partir du 1er octobre 2016 devront contenir un plan de gestion des données garantissant leur bon stockage de même qu'une réutilisation potentielle. Si à la candidature, cela ne consistera qu'à répondre à quelques questions, si le financement du projet est

[64] http://www.dcc.ac.uk/

[65] http://ckan.data.alpha.jisc.ac.uk/dataset

[66] https://www.ukdataservice.ac.uk/

[67] https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery

[68] Directions for research data management in UK universities, Cambridge [en ligne]. Novembre 2014. Disponible sur : https://www.jisc.ac.uk/events/directions-for-research-data-management-in-uk-universities-06-nov-2014#resources

[69] http://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy

[70] http://researchdata.ox.ac.uk/

[71] http://www.exeter.ac.uk/research/openaccess/data/

[72] https://data.blogs.ilrt.org/

[73] Concordat on Open Research Data launched, Research Councils UK [En ligne]. 28 juillet 2016. Disponible sur http://www.rcuk.ac.uk/media/news/160728/

[74] Higher Education Funding Council for England : http://www.hefce.ac.uk/

[75] http://www.rcuk.ac.uk/

[76] https://wellcome.ac.uk

[77] https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception

retenu, le chercheur devra alors fournir un véritable PGD. La NWO annonce d'ailleurs le suivi de la pérennisation et de l'accessibilité de ces données à l'issue du projet[78]. Les Pays-Bas avaient d'ailleurs mis en place dès 2005 le DANS[79], un service de référencement et de stockage de données, fournissant également aux chercheurs des conseils en termes d'entrepôt pour leurs *data sets*.

Les Etats mettent ainsi en place des politiques liées aux données de la recherche, soit par le biais de services nationaux, soit à travers leurs universités. Cet intérêt de la part des institutions politiques des pays du Nord ne doit pas être ignoré. La recherche est en effet un vecteur d'innovation et ainsi un élément non négligeable de l'économie. Ce qui peut partiellement expliquer l'implication des d'Etats dans cette problématique, cherchant tout comme les autres acteurs économiques à normaliser les données de la recherche afin de les rendre exploitables.

Ainsi, les agences de financement développent une demande forte en termes de dépôt des données de recherche qu'elles financent. Ces incitations au dépôt peuvent s'accompagner de recommandations sur les modalités afférentes, même si cette pratique ne s'est pas encore généralisée. C'est pourquoi, en réponse à ces demandes qui influencent les besoins des chercheurs, les universités et, de façon plus générale, les établissements d'enseignement supérieur et de recherche mettent en place des politiques de gestion des données de la recherche, se concentrant particulièrement sur la production de DMP. Les organismes de recherche et d'accompagnement à la recherche nationaux semblent souhaiter la mise en œuvre d'outils d'accompagnement ainsi que de plates-formes nationales de références sur les données de la recherche. Cette ébullition institutionnelle autour des données cherche à inciter les scientifiques au dépôt de données et est en partie liée au développement de nouveaux modèles éditoriaux.

2.1.2. De nouveaux modèles éditoriaux

En effet, si les agences de financement et tutelles recommandent le dépôt des données, cette demande trouve un écho particulièrement fort dans le domaine éditorial. Les éditeurs, en tant qu'acteurs de la recherche, se sont positionnés rapidement sur le sujet et demandent de plus en plus régulièrement le dépôt des données parallèle à la publication. Certains en ont même fait une obligation et ont ainsi vu l'essor de nouveaux modèles éditoriaux mettant en valeur les données de recherche.

L'approche traditionnelle de l'édition était de demander l'intégration des données à l'article[80]. Ce sont les modèles d'article classiques, comme par exemple le modèle IMRED[81]. Les données s'avèrent alors quasiment impossible à réutiliser.

On a ensuite vu l'apparition des *supplementary materials*. Ces « matériaux » étaient largement acceptés dans les revues en 2009, marquant ainsi le début de la

[78] NWO stimulates optimal access to research data, NWO. [En ligne] 20 juillet 2016. Disponible sur : http://www.nwo.nl/en/news-and-events/news/2016/nwo-introduces-data-management-protocol-in-all-calls.html

[79] Data Archiving and Networked Service : https://dans.knaw.nl/en

[80] Candela L. et al. « Data journals: A survey ». Journal of the Association for Information Science and Technology [En ligne]. septembre 2015. Vol. 66, n°9, p. 1747 1762. Disponible sur : http://dx.doi.org/10.1002/asi.23358

[81] Où l'article s'organise de la façon suivante : Objet-Introduction, Matériel et méthode, Résultats, Discussions, Conclusions. Voir par exemple à ce sujet la fiche « Article scientifique » du CIRAD http://coop-ist.cirad.fr/aide-a-la-publication/rediger/article-scientifique

séparation entre résultats et données. Ils sont caractérisés comme nécessaires à la compréhension de l'article ou faits pour aller plus loin, et peuvent être de forme variée : textes comprenant des informations additionnelles, tableurs, figures, vidéos ou même bases de données. On peut aussi y trouver des informations méthodologiques. Ainsi, les *supplementary materials* peuvent contenir des données de recherche, plus ou moins transformées et offrent un moyen au chercheur de diffuser une partie de ses données directement avec la publication. Ce dernier élément est également le principal défaut de ce format, car les données ne peuvent être retrouvées indépendamment de la publication.

Un format entièrement dédié aux donnés, les *data papers*[82], a alors vu le jour. Ils peuvent éventuellement être réunis dans des *data journals*. Contrairement aux *supplementary materials*, les données de recherche y seront décrites par un article spécifique constitué de métadonnées détaillées. Le lecteur aura ici accès à un historique plus précis des données. Celles-ci devront être accessibles via l'article ou via un lien pérenne. Ce format permettrait d'encourager la pratique du *peer-reviewing* pour les données de recherche et serait une occasion de donner du crédit non plus seulement à ceux qui analysent les données mais également à ceux qui les collectent et les préparent[83]. Cependant, ce dernier aspect n'est pas réservé aux *data papers* puisque les métadonnées d'un dépôt en entrepôt permettrait cette même reconnaissance. On peut également s'interroger sur le temps nécessaire à la production d'un article de ce genre si les données ont déjà été déposées dans un entrepôt et correctement décrites, étant donné l'aspect chronophage de ce dépôt.

Quoi qu'il en soit, la diffusion des données de recherche permettrait d'augmenter la citation des publications qui y sont liées[84]. Les éditeurs se sont emparés de cette thématique qui, si ce postulat est vérifié, est tout à leur avantage. C'est pourquoi de nouveaux services éditoriaux ont vu le jour ces dernières années. Thomson Reuters a créé une nouvelle métrique, dédiée aux données liées aux articles indexés par le Web of Science, le Data Citation Index[85].

Elsevier, en partenariat avec le consortium de services dédiés aux Sciences de l'Information DataCite[86] ainsi que plusieurs entrepôts de données[87], a réalisé le DataBase Linking Tool[88] pour lier les publications de ScienceDirect[89] avec des jeux de données à travers des identifiants tels que les DOI. De plus, ce groupe éditorial a annoncé[90], le 24 août 2016, la création d'Elsevier Data Search[91]. Il s'agit

---

[82] « 1 - Qu'est-ce qu'un data paper ? / Data paper / Rédiger / Aide à la publication - Coopérer en information scientifique et technique - Cirad ». [En ligne]. Disponible sur : http://coop-ist.cirad.fr/aide-a-la-publication/rediger/data-paper/1-qu-est-ce-qu-un-data-paper

[83] Chavan V., Penev L. « The data paper: a mechanism to incentivize data publishing in biodiversity science ». BMC Bioinformatics [En ligne]. 2011. Vol. 12, n°15, p. 1 12. Disponible sur : http://dx.doi.org/10.1186/1471-2105-12-S15-S2

[84] Piwowar H. A., Day R. S., Fridsma D. B. « Sharing Detailed Research Data Is Associated with Increased Citation Rate ». *PLOS ONE* [En ligne]. 21 mars 2007. Vol. 2, n°3, p. e308. Disponible sur : http://dx.doi.org/10.1371/journal.pone.0000308

[85] http://wokinfo.com/products_tools/multidisciplinary/dci/

[86] https://www.datacite.org/

[87] La liste des entrepôts supportés est accessible sur cette page Web https://www.elsevier.com/books-and-journals/content-innovation/data-base-linking/supported-data-repositories

[88] https://www.elsevier.com/books-and-journals/content-innovation/data-base-linking

[89] http://www.sciencedirect.com/

[90] Defeo C., Introducing Elsevier DataSearch [En ligne]. Mendeley blog, 24 août 2016. Disponible sur : https://blog.mendeley.com/2016/08/24/introducing-elsevier-datasearch/

d'un moteur de recherche pour les données indexant des données issues de plates-formes Elsevier ou en *open access*. En juin[92], cela concernait une partie des données associées aux publications de ScienceDirect et arXiv ainsi que les entrepôts NeuroElectro, Dryad, PetDB, ICPSR, Harvard Dataverse et TRC[93].

Enfin, on peut noter que les éditeurs se lancent également dans des projets d'entrepôts de données de recherche en tant que tel, puisque l'entrepôt de données privé Figshare[94] est un produit de Digital Science, entreprise gérée par le groupe d'édition Holtzbrinck[95]. Ce groupe détient également l'éditeur Springer qui a lancé des recommandations pour les données[96].

Les recommandations éditoriales et nouvelles politiques cherchent à standardiser les pratiques en matière de diffusion des données de la recherche, notamment en travaillant sur le lien entre publication et données ainsi que les procédures de *peer-reviewing* dans un objectif annoncé d'encourager le partage des données afin d'améliorer la reproductibilité de la recherche. *Springer Nature* a ainsi mis en place un service s'information dédié, le Research Data Support helpdesk[97].

En conclusion, les tutelles, les agences de financement ainsi que les plus grands groupes d'édition scientifique ont pour beaucoup développé des services ou recommandations liés aux données de la recherche, la plupart du temps en partenariat avec des entrepôts de données en *open access*. Les chercheurs doivent donc s'adapter à ces nouvelles exigences et seront amenées à s'informer et se former au sujet des données, tant dans les aspects techniques que législatifs. C'est à ce moment qu'ils se tournent vers les institutions de recherche locales, nationales, internationales ou disciplinaires, structures censées pouvoir apporter des réponses à ces questionnement, d'où l'urgence de développer des politiques institutionnelles claires à ce sujet. Par exemple pour orienter les chercheurs vers des outils tels que le DMP online UK[98] créé par le DCC ou le DMP Tool de l'Université de Californie[99], pour lesquels nous ne disposons pas pour l'instant d'équivalent français.

2.2. Les modalités de la collaboration entre chercheurs et professionnels de l'IST

Dans cet univers, de plus en plus d'articles de la littérature scientifique et professionnelle sont consacrés aux données de recherche et rédigés, le plus souvent, par des professionnels de l'IST (au sens bibliothécaires, documentalistes

---

[91] https://datasearch.elsevier.com/

[92] https://datasearch.elsevier.com/faq

[93] ThemoML at NIST Thermodynamic Research Center : http://trc.nist.gov/

[94] https://figshare.com/

[95] https://www.digital-science.com/about-us/

[96] Nature I. H. of D. P. at S. « Promoting research data sharing at Springer Nature ». In : BioMed Central blog [En ligne]. 2016. Disponible sur : http://blogs.biomedcentral.com/bmcblog/2016/07/05/promoting-research-data-sharing-springer-nature/

[97] http://www.springernature.com/gp/group/data-policy/helpdesk

[98] https://dmponline.dcc.ac.uk/

[99] https://dmp.cdlib.org/

et archivistes), des biologistes, des linguistes et parfois des informaticiens. Les données de recherche semblent être un domaine où les professionnels de l'IST pourraient réaliser leur mission d'accompagnement auprès des chercheurs. Toutefois, cela ne va pas sans poser plusieurs questions aux différents acteurs. On constate d'une part une forme de crispation des scientifiques vis-à-vis de ces personnes, perçues comme extérieures au monde de la recherche, qui viendraient ausculter leurs recherches sans avoir les compétences informatiques nécessaires[100]. D'autre part, une crispation de certains professionnels de l'IST vis-à-vis de nouvelles compétences à acquérir est présente, notamment en France.

Alors quel peut-être le rôle de ces derniers dans les projets de gestion des données de la recherche ? Ils sont organisateurs de nombre des journées d'études consacrées au sujet et écrivent abondamment à ce propos, tout en se divisant en deux catégories.

On peut ainsi observer un courant, porté par des doubles profils tels que les bibliothécaires informaticiens[101], comme dans cet article de la revue *Computers in Libraries* où le directeur de la Library and Information Resources Institute for Research on Labor and Employment de l'Université de Berkeley, Terence K. Huwe parle du « nouveau rôle » de *data scientist*[102]. Il identifie notamment la tendance qu'a l'analyse de données à s'immiscer dans tous les corps de métiers, y compris dans celui des professionnels de l'IST. Ceux-ci joueraient trois rôles potentiels dans les projets intégrant une dimension données : un rôle de conservation (organisation et préservation), de data visualisation ou de curateur de données.

Pour autant, comme l'évoque un rapport de l'ICSU, les données de la recherche et leur gestion ne peuvent plus se restreindre à une tâche pour du personnel mal formé ou comme une routine bâclée lors du dépôt de projets de recherche par les scientifiques[103]. La seconde catégorie est plus modérée vis-à-vis de ces évolutions professionnelles. Dans le numéro 289 d'Archimag[104] consacré au *Big Data*, la directrice de l'INTD et professeure titulaire de la chaire ingénierie documentaire au CNAM, Gislaine Chartron, rappelle que ce sont avant tout les ingénieurs informaticiens qui ont les compétences pour gérer des jeux et bases de données. Lorsqu'elle évoque le rôle que peuvent tenir les professionnels de l'IST, il s'agit plutôt d'agir en amont des projets de gestion de données pour l'identification, la qualification, la classification et la mise en place de plans de gestion, des compétences mal appréhendées par les chercheurs. En aval, les professionnels de l'IST pourraient travailler à la meilleure intelligibilité des données, notamment pour leur diffusion ou visualisation[105].

---

[100] Boukacem-Zeghmouri C., Data management in a French university: attitudes, incentives and policy, 06 juillet 2016, Berlin, Université Humboldt de Berlin.

[101] Des profils de plus en plus courants, comme le montre l'initiative Code4lib http://code4lib.org/ une plate-forme de diffusion d'offres d'emploi dédiées.

[102] Huwe T. K., « Your New Role as a Data Scientist », *Computers in Libraries.* Avril 2016, p. 23

[103] ICSU, Scientific Data and Information: A Report of the CSPR Assessment Panel, 2004 p. 9 http://www.icsu.org/publications/reports-and-reviews/priority-area-assessment-on-scientific-data-and-information-2004/PAA_Data_and_Information_report.pdf

[104] Chartron G. « Ghislaine Chartron : "Je ne transformerai pas mes étudiants en data scientists" ». *Archimag* [En ligne]. 2015, n°289. Disponible sur : http://www.archimag.com/veille-documentation/2015/11/26/ghislaine-chartron-transformer-etudiants-data-scientists

[105] Monino J.-L., Sedkaoui S. *Big Data, Open Data et valorisation des données*. London, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : ISTE Éditions, 2016

La Suisse accueille par exemple une bonne initiative en matière d'accompagnement des chercheurs à la gestion des données de la recherche par des professionnels de l'IST. La bibliothèque de l'EPFL[106] est ainsi intervenue à la journée « Les données de la recherche : où en est-on aujourd'hui ? » de l'Open Access Week Lyon 2015[107]. Cette bibliothèque propose ainsi un service d'accompagnement aux chercheurs de l'institution avec une charge de travail de deux équivalents temps plein et demi. Sollicités par les chercheurs, ces professionnels de l'IST abordent la charge de travail importante de ce genre de projet, notamment en termes d'acquisition de compétences, mais aussi l'apport positif de cette nouvelle activité, en termes de d'image pour la bibliothèque ainsi que pour la recherche elle-même grâce aux outils de *data management* développés. Ces nouveaux métiers demandent ainsi la création de nouvelles formations, à l'image du master Master Medas : MEgaDonnées et Analyse Sociale[108] ouvert cette année par l'INTD-CNAM. Les formations existantes devront s'adapter afin de proposer des cursus en accord avec ces évolutions professionnelles.

Bibliothécaires et documentalistes pourraient ainsi acquérir ces compétences tout en mobilisant leurs outils traditionnels d'organisation de l'information, de référencement et de recherche d'information afin de permettre l'amélioration des liens dans les universités et structures de recherche entre laboratoires, services informatiques et administrations[109]. D'autant plus qu'ils sont particulièrement investis dans les organismes nationaux et internationaux travaillant actuellement sur les données comme le CODATA[110], la Research Data Alliance[111], l'ICSU World Data System[112] ou encore le DCC.

2.3. Impacts socio-culturels de la libération des données

La diffusion en *open access* des données de recherche amène également des questionnements d'ordre socio-culturel. Ainsi, au-delà des questionnements éthiques sur l'anonymisation, déjà abordés dans les enjeux scientifiques, et des débats épistémologiques, on peut s'interroger sur l'impact de l'ouverture des données de recherche.

Les médias, en s'emparant du sujet plus vaste du *big data*, sont prompts à présenter les données comme solution à de nombreux problèmes. Cependant, les données ne sont pas représentatives d'une vérité mais mouvantes, subjectives et surtout doivent être interprétées. Les chercheurs peuvent être vus comme des interprètes des données dont l'imagination est alimentée par ses paradigmes disciplinaires[113].

---

[106] École Polytechnique Fédérale de Lausanne : http://library.epfl.ch/cms/lang/fr/pid/119191

[107] Captations vidéo de la journée http://www.bibliotheque-diderot.fr/les-journees-open-access-275727.kjsp

[108] http://intd.cnam.fr/ouverture-du-master-medas-en-professionnalisation-megadonnees-et-analyse-sociale-804366.kjsp

[109] http://www.ariadne.ac.uk/issue70/cox-et-al

[110] http://www.codata.org/

[111] https://rd-alliance.org/node

[112] https://www.icsu-wds.org/

[113] Gitelman L., « Notes for the Upcoming Collection "Raw Data" is an Oxymoron », 23 juillet 2011, https://files.nyu.edu/lg91/public

« L'utilisation de données numériques publiées, comme l'utilisation de publications, dépend de la capacité à identifier, authentifier, localiser, accéder et interpréter ces données. »[114]

La décontextualisation est une des craintes de la communauté scientifique envers la diffusion des données, qui pourraient perdre en intelligibilité, notamment dans le cadre de recherches qualitatives. Cette décontextualisation amène également la question de la mésinterprétation des données de recherche. Un pan de la recherche réfractaire au partage, ou s'interrogeant simplement sur ses modalités, pointe régulièrement du doigt les compétences nécessaires à la compréhension de données de recherche. Il s'agit à la fois de compétences matérielles (l'accès à certains logiciels ou certaines machines pourrait être essentiel pour interpréter les données) mais également de compétences analytiques.

Un des arguments avancés, à l'inverse, par les défenseurs de l'ouverture des données est une sorte de retour à la société. En effet, le financement de la recherche étant partiellement réalisé par les citoyens, ceux-ci pourraient avoir *de facto* un droit d'accès aux recherches produites et donc aux données. Pour autant, le risque de mésinterprétation reste réel et le développement d'une science citoyenne parallèle à l'ouverture des données doit s'accompagner d'une médiation de la part des chercheurs et des professionnels de l'IST.

La diffusion des données de recherche pourrait induire des avancées considérables pour la société dans son ensemble, par exemple en matière de santé publique. Le circuit traditionnel de la recherche est particulièrement long. Or, lors d'une urgence sanitaire, on ne peut pas se permettre d'attendre des années la publication des données à travers les articles de revues. Ainsi, l'accès facilité aux données par l'ouverture pourrait permettre à la recherche bio-médicale une plus grande efficacité[115].

Cependant, le partage des données de recherche doit également être questionné dans une optique de développement durable. Stocker massivement des données dans des centre de données semble aujourd'hui peu onéreux mais s'avère coûteux en énergie et donc coûteux écologiquement mais aussi économiquement sur le long terme[116]. On peut donc s'interroger sur la pertinence du stockage exhaustif, une tentation forte qui pourrait s'avérer mal à propos à la fois scientifiquement et socialement.

3. ENJEUX SCIENTIFIQUES

« Les chercheurs ont besoin pour réaliser leurs travaux de recherche d'un accès libre et gratuit sous forme numérique à l'ensemble des données de la Science composées :

- *des résultats scientifiques en ce compris les résultats publiés par un éditeur scientifique ;*

[114] CODATA-ICSTI Task Group on Data Citation Standards and Practices. *op.cit.*, p. 22

[115] Chretien J.-P., Rivers C. M., Johansson M. A. « Make Data Sharing Routine to Prepare for Public Health Emergencies ». *PLOS Med* [En ligne]. août 2016. Vol. 13, n°8, p. e1002109. Disponible sur : http://dx.doi.org/10.1371/journal.pmed.1002109

[116] Les data centers sur Plaine Commune, Agence Locale de l'Energie et du Climat de Plaine Commune [En ligne]. Août 2013. Disponible sur : http://www.alec-plaineco.org/wp-content/uploads/2013/10/ALEC-Plaine-Commune-2013-Les-data-centers-sur-Plaine-Commune.pdf

- *des données de la recherche entendues comme les données ayant servi à établir ces résultats. »*[117]

La problématique du partage des données de la recherche rejoint les débats plus généraux du *big data* et de l'*open data*, c'est à dire de la mise à disposition massive de jeux de données pouvant être exploités scientifiquement. Plusieurs enjeux scientifiques peuvent être identifiés dans ce contexte.

3.1. La recherche par les données, quelles limites et quels changements épistémologiques ?

Dans son rapport sur le *Big data*[118], François Ewald estime, en parlant des données de façon générale, que le développement de *l'open data* amènerait une révolution épistémologique, notamment liée aux modèles de prédiction. Notre rapport à la connaissance deviendrait quantitatif, la donnée ne serait plus un produit de la connaissance mais son matériau. Dans une certaine mesure, la perspective de pouvoir conserver l'ensemble des données produites par nos sociétés réveillerait le vieux rêve de la connaissance universelle.

Dominique Cotte, chercheur en SIC au laboratoire Geriico, exprimait ainsi une crainte, lors de la journée Humanités numériques et données ouvertes[119] organisée à Lyon en mai 2016, crainte de l'abandon du modèle hypothético-déductif. Certains chercheurs, redoutent ainsi que l'étude des données par les algorithmes prenne le pas sur l'analyse qualitative et qu'on ne s'interroge plus que, par exemple, sur ce que les gens font et non pourquoi[120].

De plus, on remarque que les éditeurs, les agences de financement et les tutelles s'expriment aujourd'hui de façon régulière sur le sujet[121]. À quel point cette problématique de diffusion des données de la recherche répond-elle à un besoin des chercheurs ? Ne reflète-t-elle pas plutôt un nouvel intérêt économique ou un nouvel outil d'évaluation ? D'autant plus que des services statistiques liés au nombre de téléchargement ou de consultation des jeux de données sont implantés. Il faut ainsi mettre en regard les besoins de l'institution et les besoins des chercheurs. Les données deviennent requises lors des demandes de financement et le projet risque d'être mal évalué si elles sont mal traitées ou si le PDG n'a pas été suivi. Une des craintes des chercheurs est ainsi que l'évaluation glisse vers une prise en compte des données de recherche[122].

Le contexte ambivalent de la recherche actuelle pousse les chercheurs à vulgariser et à diffuser leurs matériaux, tout en ne valorisant que la publication d'articles dans des revues scientifiques à comité de lecture lors de l'évaluation. Une des craintes des chercheurs est donc d'ordre temporel. Le partage des données

---

[117] Direction de l'Information Scientifique et Technique – CNRS,2016, *op. cit.*, p. 18

[118] Ewald F., Assurance, prévention, prédiction... dans l'univers du Big Data. [En ligne] 2014. Disponible sur : http://www.institut-montparnasse.fr/wp-content/files/Collection_recherches_n_4.pdf

[119] https://hnlyon2016.sciencesconf.org/

[120] Anderson C. « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete » WIRED. [En ligne] . Disponible sur : http://www.wired.com/2008/06/pb-theory/

[121] Borgman C. « The Conundrum of Sharing Research Data ». Journal of the American Society for Information Science and Technology. [En ligne]. 2012. Vol. 63, n°6, Disponible sur : http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1869155 pp. 3-4

[122] Boukacem-Zeghmouri C., 2016, *op. cit.*, p. 26

risque de prendre le pas sur la publication, alors même que cette activité n'est pas valorisée.

D'autre part, les chercheurs craignent également que la diffusion des données de recherche entre en compte dans l'évaluation. On pourrait alors voir apparaître des logiques de standardisation des pratiques de la recherche. C. Boukacem évoquait ainsi le « nouveau chercheur global » au séminaire ELICO « Observer les dynamiques socio-économiques de la publication scientifique : approches qualitative et bibliométrique ». Ce modèle du chercheur publie sans douter et sans faire d'erreur, de façon normalisée et sans prise de risque, l'important n'est plus le projet mais le décrochement du financement. Il risque de toucher essentiellement les jeunes chercheurs, plus précaires, qui pourraient mettre en place des stratégies de publication des données extrêmement normées afin de ne pas avoir à justifier leur façon de faire[123].

Malgré ces écueils, la diffusion des données de recherche peut également participer au renouvellement des méthodes de recherche. Le traitement de jeux de données massifs par le biais de la fouille de texte et de données n'est pas à rejeter. Le Livre blanc du CNRS le présente comme un enjeu majeur de la science contemporaine[124].

3.2. Un encouragement à la vérification, réutilisation et à la création

Les données de recherche sont marquées par le phénomène de la longue traîne. Une partie des corpus de données est bien organisé, référencé et accessible par la publication. En revanche, la plupart de ces données reste dans l'ombre, comme le montre le schéma ci-dessous. Dans l'étude menée par *Science* en 2011, 7,6% des chercheurs interrogés déclaraient ainsi que leurs données étaient archivées dans un entrepôt. À l'inverse, 88,7% de l'échantillon gardait ses données sur son disque dur personnel ou dans les locaux de son laboratoire[125].

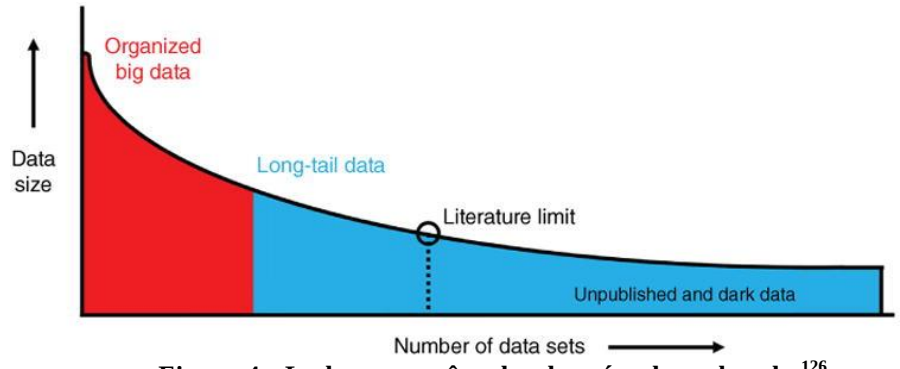


Figure 4 : La longue traîne des données de recherche[126]

[123] Observer les dynamiques socio-économiques de la publication scientifique : approches qualitative et bibliométrique, Séminaire ELICO. Le programme est disponible sur : http://www.elico-recherche.eu/actualites/observer-les-dynamiques-socio-economiques-de-la-publication-scientifique-approches-qualitative-bibliometrique-vendredi-17-juin-2016-lyon-1

[124] Direction de l'Information Scientifique et Technique – CNRS, *op. cit.*, p. 18

[125] Staff Science. « Challenges and Opportunities ». *Science* [En ligne]. 11 février 2011. Vol. 331, n°6018, p. 692 693. Disponible sur : http://dx.doi.org/10.1126/science.331.6018.692

[126] Ferguson A. R. et al. « Big data from small data: data-sharing in the "long tail" of neuroscience ». *Nature Neuroscience* [En ligne]. 28 octobre 2014. Vol. 17, n°11, p. 1442 1447. Disponible sur : http://dx.doi.org/10.1038/nn.3838

Les chercheurs reconnaissent que le partage de leurs données permettrait d'augmenter la visibilité de leur travail[127] et éviter l'éparpillement des efforts[128]. Les défenseurs du dépôt des données avancent également que les chercheurs pourraient réutiliser ces données pour reproduire les recherches, valider les résultats ou créer de nouvelles recherches. Néanmoins, les scientifiques restent très partagés vis-à-vis de la diffusion des données et que leurs avis dépend des disciplines tout autant que des opinions personnelles et individuelles. Certains chercheurs, en particulier en SHS semblent ainsi ouvert au partage des données mais n'utilisent pas celles qui ont été produites par d'autres à d'autre fin que de la citation[129].

Ce n'est pour l'instant pas le cas de tous les domaines scientifiques, certaines disciplines sont plus acculturées aux données de recherche et à leur utilisation. La Biologie, par exemple, a produit des études utilisant des données de recherche issues d'entrepôts en *open access* avec succès[130]. En étudiant la citation des données issues de l'entrepôt Dryad dans plusieurs bases bibliographiques (Google Scholar, Web of Science, Scopus) He et Nahar ont montré qu'il s'agissait le plus souvent d'auto-citation[131], notamment afin d'appuyer les résultats et démontrer ainsi la qualité de l'étude.

Le partage de ces données pourrait également pousser les scientifiques à la collaboration[132]. Les techniques utilisées pour le traitement des données telles que la *data vizualisation* ou les bases de données offrent l'opportunité d'un rapprochement entre des chercheurs de disciplines diverses. Ces collaborations peuvent se dérouler au sein de projets de recherche traditionnels aussi bien que dans une visée plus épistémologique ou lors d'études bibliométriques.

Pour conclure, si le partage de données pose de nombreuses questions épistémologiques ou éthiques, il a un intérêt certain pour la recherche. Cela nécessite néanmoins des infrastructures pour stocker ces données et les diffuser. Ce sont ces plates-formes qui permettront aux chercheurs de gagner du crédit par leurs données et de réutiliser des jeux existants[133].

---

[127] Ferguson A. R., 2014, *op. cit.*, p. 32

[128] Wallis J. C., Rolando E., Borgman C. L. « If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology ». *PLOS ONE* [En ligne]. juil 2013. Vol. 8, n°7, p. e67332. Disponible sur : http://dx.doi.org/10.1371/journal.pone.0067332

[129] Cabrera F. « Les données de la recherche en Sciences humaines et sociales: enjeux et pratiques ». Disponible sur : http://hal-obspm.ccsd.cnrs.fr/mem_01128394/document p. 89

[130] He L., Nahar V., Lewandowski D. « Reuse of scientific data in academic publications: an investigation of Dryad Digital Repository ». *Aslib Journal of Information Management* [En ligne]. 2016. Vol. 68, n°4,. Disponible sur : http://www.emeraldinsight.com/doi/abs/10.1108/AJIM-01-2016-0008

[131] *Ibid.*

[132] Kenall A., Harold S., Foote C. « An open future for ecological and evolutionary data? ». *BMC ecology*. [En ligne] 2014. Vol. 14, n°1, p. 10. Disponible sur : http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-14-66

[133] Candela L., *op. cit.*, p. 25

4. LES QUESTIONS JURIDIQUES[134]

L'ouverture des données de recherche cristallise les conflits d'intérêts entre d'un côté les chercheurs, les militants de l'*open access*, ainsi que certaines institutions favorables à cette diffusion comme le CSPLA, l'INRA ou le consortium COUPERIN et de l'autre côté, les éditeurs. Dans ce contexte, les institutions transnationales, c'est particulièrement le cas de la Commission Européenne, de même que des États, recherchent aujourd'hui des compromis pour permettre cette ouverture.

4.1. Le droit français

L'IST est aujourd'hui encadrée en France par trois couches de droit. Le droit d'auteur, selon l'Article L.111 du Code de la Propriété Intellectuelle[135], y intègre un champ très large depuis les textes à la musique en passant par les vidéos, les photographies etc. Cependant, les idées, les faits, les données ou les découvertes y restent « libres de parcours ». La production scientifique est spécifique au sens où certains de ses acteurs sont des agents publics. En principe, leur droit d'auteur est affaibli[136] car limité au droit de paternité, le reste étant automatiquement cédé à l'État. Pour autant, la loi ménage une exception qui permet aux professeurs et aux chercheurs de rester titulaires des droits sur leurs créations (articles, ouvrages, cours, rapports, etc.) même s'ils restent cessibles à des éditeurs. Les différentes clauses des contrats d'édition et des financements de recherche, liées à la confidentialité par exemple, viennent d'ailleurs complexifier encore le sujet.

La seconde couche de droit appliquée sur les données de recherche est le droit des bases de données. Plus récent, il date des années 1990 et est une transposition française de lois européennes. On y trouve la définition légale d'une base de données :

> *« On entend par base de données un recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen. »*[137]

La loi ne définit pas en revanche la donnée. Trois couches de protection peuvent ainsi être appliquées sur les bases de données. Tout d'abord, la structure de la base qui peut se protéger avec le droit d'auteur si elle est originale. Vient ensuite le contenu de la base de données, avec le droit *sui generis*, très particulier, qui n'est pas un droit de la création mais protégeant l'investissement. Celui-ci n'est applicable que quinze ans à partir de l'investissement pour empêcher les extractions et diffusions de contenu. Ce droit peut appartenir à une personne

---

[134] Une partie du contenu de cette partie est une synthèse de l'atelier *Aspects juridiques, éthique et droit*, assuré par Lionel Maurel de l'Université Lumière, suivi dans le cadre des journées Humanités Numériques et Données Ouvertes. Le PDF de la présentation est consultable sur cette page : https://hnlyon2016.sciencesconf.org/conference/hnlyon2016/pages/Atelier_Maurel_L._Aspects_juridiques.pdf

[135] *Code de la propriété intellectuelle - Article L112-3 | Legifrance* [En ligne]. Disponible sur : https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006278879&cidTexte=LEGITEXT000006069414

[136] LOI n° 2006-961 du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information, *Legifrance*. [En ligne]. Disponible sur : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350&dateTexte=&categorieLien=id

[137] *Code de la propriété intellectuelle - Article L112-3 | Legifrance, op. cit.*

morale et concerne la recherche, car c'est la personne ayant réalisé l'investissement qui en bénéficie. Enfin, sont protégés les éléments constitutifs du contenu. Des licences spéciales existent afin de couvrir toutes les couches de façon cohérente, comme l'ODbL[138].

Or, comme le rappelle le Livre blanc du CNRS, l'IST numérique est la plupart du temps accessible sur des bases de données institutionnelles, des épi-revues ou des bases de données en *open access*[109]. Ce sont donc les producteurs de ces bases qui disposent du droit *sui generis* et peuvent en interdire « toute extraction qualitativement ou quantitativement substantielle »[139].

Pour finir, la dernière couche de droit pouvant s'appliquer pour les données de la recherche est le droit des données publiques, auquel la Loi république numérique s'intéressera particulièrement. Il s'attache à tout ce que les administrations produisent dans le public. Ce n'est pas un droit de propriété intellectuelle mais un droit de réutilisation reconnu au profit des citoyens, pouvant toutefois être encadré et conditionné par l'administration (notamment au paiement d'une redevance). La Directive 2003/98 du 17 novembre 2003[140], transposée en droit français par l'ordonnance du 6 juin 2005 relative à la liberté d'accès aux documents administratifs et à la réutilisation des données publiques, disposait d'un régime dérogatoire pour certaines administrations, dont les organismes de recherche, ce qui devrait changer avec la future loi Lemaire.

Pour conclure, si les agents publics sont censés céder leurs droits sur leurs créations à leur administration de tutelle, les enseignants chercheurs conservent pleinement ce droit de paternité. Les différentes couches de droit font que la plupart du temps, les données et les métadonnées appartiennent à l'institution qui produit la base en collectant les informations tandis que les contenus appartiennent à l'individu.

4.1. La loi pour une République Numérique

Les administrations seront obligées, dans une certaine mesure qui reste pour l'instant floue, de mettre en ligne et de rendre librement et gratuitement réutilisables les bases de données qu'elles produisent et les informations essentielles qu'elles détiennent, listées dans un Répertoire d'Information Publiques (RIP), en faisant primer la vie privée sur leur diffusion. Le Livre blanc du CNRS recommande ainsi l'accompagnement de l'ouverture des données de la recherche d'un cadre éthique solide[141].

L'impact de cette loi sur les données de la recherche reste encore obscur. La transposition de la directive européenne sur la réutilisation des informations du secteur public supprimerait l'exception concernant les données. Une administration ayant un droit des bases de données ne pourra l'opposer à une demande de réutilisation, sauf si un tiers, comme un chercheur dispose également d'un droit sur la base.

---

[138] http://opendatacommons.org/licenses/odbl/

[139] *Code de la propriété intellectuelle - Article L112-3 | Legifrance, op. cit,* p. 34

[140] DIRECTIVE 2003/98/CE DU PARLEMENT EUROPÉEN ET DU CONSEIL du 17 novembre 2003 concernant la réutilisation des informations du secteur public. [En ligne] Disponible sur : http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:FR:PDF

[141] Direction de l'Information Scientifique et Technique – CNRS, *op. cit.*, p. 18

*« II. - 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. »*
« III. - 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. »

Tout serait donc réutilisable sauf si une couche de droit s'est appliquée avant et un garde-fou est appliqué aux éditeurs avec le paragraphe III. Cela semble crucial à la vue du développement de la recherche par fouille de textes et de données (TDM) qui ne bénéficie pas pour l'instant de statut légal propre[142]. Le gouvernement ne souhaitant pas réduire le droit d'auteur avec des exceptions, le cauchemar juridique évoqué par Rémi Gaillard en citant Serge Bauin[143] semble amené à perdurer.

[142] Direction de l'Information Scientifique et Technique – CNRS, *op. cit.*, p. 18

[143] Gaillard R. *op.cit.*, p. 17

PARTIE 2 : LES ENTREPOTS DE DONNEES : DES DEFIS TECHNIQUES, DOCUMENTAIRES ET SCIENTIFIQUES

1. LES ENTREPOTS FRANÇAIS

Le répertoire d'entrepôts de données de recherche re3data[144] recensait en juin 2016 soixante-dix-huit entrepôts français. La plupart proposent un accès non restreint à la base de données étaient à but non lucratif et disciplinaires. Cependant, comme on l'a évoqué plus haut, le re3data peut aussi bien recenser de simples bases de données de recherche propres à un laboratoire que des entrepôts multidisciplinaires transnationaux.

Néanmoins, l'offre française semble s'étoffer en matière d'entrepôts institutionnels. Ceux-ci restent la plupart du temps restreints à un type de données ou bien représentant seulement les disciplines présentes dans l'institution de laquelle ils dépendent.

On peut citer en Sciences Humaines et Sociales le Réseau Quetelet[145], l'équipement DIMESHS[146] de Sciences-Po, ou encore NAKALA, mis en œuvre par la TGIR Huma-Num et offrant un service de stockage sécurisé avec identifiant pérenne et accès interopérable avec OAI-PMH mais sans moteur de recherche, site Web d'éditorialisation scientifique ou dispositif d'enrichissement de données.

Les archives ouvertes françaises développent également ce type de service. Certaines n'ont pas développé d'offre d'hébergement mais annoncent une politique volontaire en termes de données de recherche. C'est le cas de ProdInra[147], maintenue par l'INRA, qui réalise un cahier des charges des entrepôts de confiance[148] afin d'accompagner ses chercheurs dans le dépôt de données. Actuellement, les scientifiques de l'INRA sont orientés vers l'archive européenne Zenodo.

D'autres ont développé des services de stockage et de référencement. C'est notamment le cas d'ArchiMer[149], archive institutionnelle de l'IFREMER. Cette institution a mis en œuvre l'entrepôt de données de la recherche Sextant et la possibilité d'attribuer des DOI aux éléments qui y sont hébergés. Or, Archimer permet d'associer à une publication des données via le DOI. Il est donc possible de lier les données présentes sur Sextant ou un autre entrepôt avec les publications d'Archimer à partir du moment où ces données ont été dotées d'un identifiant DOI[150].

[144] http://www.re3data.org/

[145] http://www.reseau-quetelet.cnrs.fr/spip/

[146] http://www.sciencespo.fr/dime-shs/

[147] http://prodinra.inra.fr/?locale=fr

[148] https://www6.inra.fr/datapartage/

[149] http://archimer.ifremer.fr/

[150] Source et quelques exemples : http://archimer.ifremer.fr/html/association-publications-donnees-de-la-recherche.htm

HAL se situe plutôt dans la première catégorie. L'archive nationale héberge en effet des données de recherche sur le mode des *supplementary materials* depuis plusieurs années, mais sans pour autant avoir de politique formelle à ce propos.

L'article d'Assante, Candella, Castelli et Tani paru cette année identifie huit critères pouvant accompagner l'étude d'entrepôts de données de la recherche[151]. Cinq d'entre eux interviennent lors de la préparation et de la publication des données. Cela concerne le format des fichiers, leur documentation avec les métadonnées, la licence choisie pour le contenu, les coûts de publication éventuels et enfin la validation ou modération des données. Les trois derniers critères conditionnent la vie des données suite à leur publication. C'est-à-dire la disponibilité des données, leur accessibilité ainsi que la possibilité de les citer. La partie suivante reprend ces différents critères afin d'étudier les différents enjeux à l'œuvre pour les entrepôts de données de la recherche en *open access*. Les entrepôts cités sont pour la plupart multi-disciplinaires et acceptent soit tous les dépôts, soit ceux liés à des publications ou enfin en se concentrant sur une thématique particulière. Ce dernier modèle, par exemple, est celui d'arXiv, qui accepte le dépôt de données de recherche s'ils sont en corrélation directe avec la publication, elle-même conditionnée par les disciplines acceptées dans l'archive.

2. LA PREPARATION ET PUBLICATION DES DONNEES

C'est au chercheur d'estimer à quel moment de son cycle de vie la donnée doit être diffusée. Cela doit être envisagé dès la création pour qu'un enrichissement ait lieu lors du traitement par le biais des métadonnées. Alors les données pourront être déposées et conservées dans l'optique d'une future diffusion et réutilisation.

Trois éléments sont à identifier lors d'un dépôt de données : les données elles-mêmes, les métadonnées les décrivant et éventuellement la publication qui s'y rattache (*paper* et/ou *data paper*). Il convient de différencier les entrepôts réalisant un simple référencement de ceux stockant les données. Les entrepôts cités ici hébergent les données, les fichiers y sont donc stockés.

2.1. Les formats utilisés

Les entrepôts doivent faire face à une diversité de formats à l'image de la diversité de la recherche. Si certains entrepôts disciplinaires peuvent se permettre de restreindre les formats acceptés, les entrepôts multi-disciplinaires en sont incapables. Certains néanmoins prodiguent des conseils vis-à-vis des types de format de fichier à utiliser, notamment pour des raisons de pérennité[152]. C'est notamment le cas de 4TU.Centre for Research Data qui fournit même une liste des « formats préférés » [153].

On voit ainsi que les entrepôts cherchent à orienter les scientifiques vers les formats non-propriétaires[154] permettant des ponts inter-disciplinaires, la pérennité

[151] Assante M., Candela L., Castelli D., Tani A. « Are Scientific Data Repositories Coping with Research Data Publishing? ». *Data Science Journal* [En ligne]. 26 avril 2016. Vol. 15,. Disponible sur : http://dx.doi.org/10.5334/dsj-2016-006

[152] *Ibid.*, pp. 6-7

[153] http://researchdata.4tu.nl/en/publishing-research/data-description-and-formats/

[154] Chez Dryad par exemple : http://datadryad.org/pages/filetypes

des données et dont les logiciels de consultation et d'édition sont accessibles à la communauté des utilisateurs. On retrouve alors des formats de type CSV, XML, JSON ou encore WAVE, par exemple.

Une limite de taille de fichier est néanmoins souvent établie. Parfois, cette taille conditionne l'application de coûts de publication, notamment chez Dryad en cas de dépassement d'une certaine limite. Ainsi, au-delà de 20GB, des frais supplémentaires sont demandés, d'une hauteur de 50$ à partir du dépassement puis pour chaque tranche de 10GB supplémentaire[155].

Ainsi, si les entrepôts ne peuvent visualiser de façon exhaustive les futurs formats de fichier auxquels ils devront faire face, ils peuvent toutefois sensibiliser les chercheurs à l'utilisation de certains d'entre eux. Néanmoins, cela rend difficile pour les entrepôts de mettre en place des services demandant une gestion des formats de fichier, comme de la visualisation de données[156]. Les scientifiques doivent être orientés vers des formats neutres et sensibilisés au rapport entre format de fichier, disponibilité des logiciels et visualisation ou ré-exploitation.

2.2. Métadonnées et description

La description des jeux de données est une des étapes les plus primordiales du dépôt de données de la recherche dans un entrepôt. Le DCC a défini plusieurs bénéfices pouvant être retirés de l'attention accordée aux métadonnées.

Les bénéfices immédiats d'une bonne description concernent le lecteur, qui peut localiser, valider les données et les utiliser. Le lien vers la publication d'origine, s'il existe, lui apporte une contextualisation ainsi que des informations méthodologiques. L'auteur quant à lui, bénéficiera du crédit de ses données ainsi que de leur impact sur d'autres publications si elles sont citées. À plus long terme, la description des données peut leur assurer une certaine pérennité, permettre de lutter contre le plagiat et rendre la recherche des jeux de données plus simple. L'auteur pourra également mesurer son impact sur la recherche et en tirer de la reconnaissance[157].

Concernant la documentation des données, aucune ne semble requérir de *data paper*. CSIRO et Dryad recommandent néanmoins l'accompagnement des dépôts de données d'un document de type *Read-Me*. L'inconvénient de ce type de document descripteur, tout comme les *data papers*, réside dans leur format, intelligible par l'humain mais pas par la machine.

Les métadonnées semblent donc cruciales. Les données ont besoin d'être décrites, environnées afin d'avoir une valeur scientifique et être comprise. C'est également ce qui permet aux données d'être retrouvées, grâce à l'indexation par les moteurs de recherche ou l'interopérabilité assurant un moissonnage par d'autres plates-formes.

Les métadonnées ont ainsi un aspect descriptif. Elles doivent fournir des informations sur le jeu de données et être lisibles aussi bien par des chercheurs d'une même communauté disciplinaire que, potentiellement, par ceux d'autres

[155] http://datadryad.org/pages/payment

[156] Assante M., 2016, *op. cit.*, p. 38

[157] Ball A., Duke M., « Data Citation and Linking | Digital Curation Centre », DCC [En ligne]. 19 juillet 2011, version mise à jour le 21 juin 2012 Disponible sur : http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking

communautés. Pour utiliser un exemple issu des Sciences de la Vie, une séquence ADN d'une bactérie marine collectée devrait être accompagnée de métadonnées sur la localisation de la prise d'échantillon (latitude, longitude, profondeur), sur les paramètres environnementaux, la collecte (quelle collecte ? Quand ? Avec quels outils?) ainsi que d'un identifiant[158] selon la version utilisée. On a alors des métadonnées descriptives, variant selon le type de données et la discipline, un identifiant pérenne mais aussi des données méthodologiques.

Les entrepôts de données de recherche fournissent généralement à leurs déposants la possibilité d'utiliser un identifiant. Il peut s'agir d'une URL pérenne ou d'un DOI. Un grand nombre d'entrepôts proposent un système d'identification qui leur est propre[159].

L'identification des données avec un identifiant unique et pérenne permettra la citation, d'autant plus qu'il peut être unique ou par version. Car la longueur du cycle de vie des données de recherche ainsi que leur dynamisme doivent attirer l'attention sur leur traçabilité, à la fois par les traitements effectués et que par le *versionning*.

Il s'agit d'un des points faibles dans les entrepôts de données de la recherche qui ne permettent pas forcément au déposant de modifier les jeux de données en ajoutant une nouvelle version. Alors qu'il paraît primordial de dater les jeux de données, seulement une des plates-formes étudiées par Austin et. al. offre une méthode d'identification des versions systématique et permanente des jeux de données[160].

Des données méthodologiques peuvent également être présentes. Souvent peu représentées dans les publications, on voit aujourd'hui émerger des revues dédiées à la méthodologie de recherche et des projets de sensibilisation dans les laboratoires[161].

De nombreux standards existent pour la description de ces métadonnées. Ils peuvent être généraux, comme le DataCite Metadata Schema 3.1[162], créé spécifiquement pour les données de recherche et dont la version précédente[163] avait été traduite par l'INRA[164]. La version 3.1 propose 3 niveaux d'obligation dans le renseignement des métadonnées :

- *mandatory* (obligatoires),
- *recommended* (recommandées),
- *optional* (optionnelles).

[158] Thessen A., 2011, *op. cit.*, p. 18

[159] Austin C., 2016, *op cit.*, p. 20

[160] *Ibid.*

[161] Cabrera F., 2015, *op cit.*, p. 33

[162] La page officielle du format : https://schema.datacite.org/meta/kernel-3/. Un exemple au format XML : https://schema.datacite.org/meta/kernel-3/example/datacite-example-full-v3.1.xml

[163] La version précédente 2.1 : « DataCite Metadata Schema 2.1 », DataCite [En ligne]. Disponible sur : https://schema.datacite.org/meta/kernel-2.1/

[164] Traduction de la version 3.1 en 2015 par l'INRA : « Schéma de métadonnées datacite pour la publication et la citation des données de la recherche ». [En ligne]. Disponible sur : http://prodinra.inra.fr/?locale=fr#!ConsultNotice:326796

ID	Property	Obligation
1	Identifier (with type sub-property)	M
2	Creator (with name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M

Figure 5 : Les métadonnées obligatoires à la description d'un jeu de données selon le DataCite

Le DataCite recommande de fournir au moins les deux premiers types de métadonnées et est en partenariat avec la communauté Science and Metadata de la Dublin Core Initiative (DCMI) afin de produire une version du schéma dans un profil d'application Dublin Core.

Ces standards peuvent également être spécifiques. Le Royaume-Uni, par exemple, souhaite s'inspirer du schéma du DataCite afin de créer son propre standard[165]. De nombreux standards disciplinaires ont été mis en place, dont voici quelques exemples : Ecological Metadata Language (EML)[166], compatible avec le Dublin Core, OLAC (Open Language Archives Community)[167], CMDI (Component MetaData infrastructure)[168], Darwin Core[169] etc.

Cette démultiplication des standards disciplinaires ne produit pas pour autant de compétition réelle : les métadonnées des standards décrivant les données de recherche répondent souvent à un besoin spécifique tout en restant intelligible pour les autres systèmes, en particulier grâce à une approche d'emprunts des champs, on les retrouve donc partiellement d'un système à l'autre.

Cependant, elle est révélatrice de l'ignorance de l'existence de ces schémas plus ou moins satisfaisant. On recrée donc de nouveaux standards sans chercher à améliorer ou fusionner les existants. C'est d'ailleurs peut-être également le cas avec les entrepôts de données ? Plusieurs initiatives ont néanmoins été mises en œuvre pour documenter[170] et lister les standards de description des données de recherche[171]. La RDA[172] a ainsi développé le Metadata Directory[32] afin d'accompagner les chercheurs et les entrepôts dans leur choix de standards de métadonnées. Ce répertoire rassemble les standards, les extensions (variations des

[165] Brown S., Bruce R., Kernohan D., Directions for Research Data Management in UK Universities, JISC [En ligne] mars 2015. Disponible sur : http://repository.jisc.ac.uk/5951/4/JR0034_RDM_report_200315_v5.pdf

[166] https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

[167] http://www.language-archives.org/

[168] http://www.clarin.eu/content/component-metadata

[169] http://rs.tdwg.org/dwc/

[170] Ball A., Chen S., Greenberg J., Perez C., Jeffery K., Koskela R. « Building a Disciplinary Metadata Standards Directory ». *International Journal of Digital Curation* [En ligne]. 17 juin 2014. Vol. 9, n°1, p. 142 151. Disponible sur : http://dx.doi.org/10.2218/ijdc.v9i1.308

[171] Par exemple : McQuilton P., Gonzalez-Beltran A., Rocca-Serra P., Thurston M., Lister A., Maguire E., Sansone S.-A. « BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences ». *Database* [En ligne]. 1 janvier 2016. Vol. 2016, p. baw075. Disponible sur : http://dx.doi.org/10.1093/database/baw075

[172] Ball A., Greenberg J., Jeffery K., Koskela R. « RDA Metadata Standards Directory Working Group: Final Report ». [En ligne] 2016. Disponible sur : https://rd-alliance.org/system/files/MSDWG-Final-Report.pdf

standards), des outils et des cas d'usages pouvant être triés selon plusieurs catégories et par sous-domaines.

2.3. Les licences et les entrepôts de données

Actuellement, pour réutiliser un objet soumis au droit d'auteur, il convient d'en demander l'autorisation au titulaire, sauf s'il existe une exception législative. Pour ouvrir un objet juridiquement, il faut donc lui attribuer une licence. Un contenu sous CC, par exemple, peut être réutilisé sans autorisation à demander, tant qu'on reste dans le cadre définit par la licence. Les conditions d'exploitation d'un jeu de données de recherche doivent donc être décrites, afin que l'utilisateur puisse cerner les différents paramètres.

Les licences utilisées par les entrepôts de données dépendent de la thématique choisie et donc du contenu qui sera mis à disposition sur la plate-forme. Les entrepôts adaptent les licences à leurs utilisateurs sans pour autant suivre de standard particulier, en différenciant néanmoins la relation de l'entrepôt avec le déposant et l'utilisateur. Certains développent leur propre licence, comme le CSRIO[173]. Un grand nombre d'entre eux mettent à l'honneur les licences de type CC[174], c'était le cas de 22 entrepôts sur un échantillon de 34 lors de l'étude commanditée par le Research Data Canada[175]. Tous ces entrepôts ne pratiquent pas l'*open acces*, certains restreignent l'accès aux données ou leur réutilisation, comme on peut le constater sur re3data avec la présence de la catégorie « *Data access restrictions* ».

L'utilisation d'une licence ouverte neutralise le droit des bases de données et le choix, subtil, des couches à protéger est un véritable défi juridique. Or, les chercheurs sont la plupart du temps à mille lieux de ces connaissances juridiques. Les choix de licence existent avec de nombreuses variations. La licence CC propose par exemple quatre paramètres différents permettant un choix total entre six licences[35]. Pour la couche des données, comme la numérisation d'un livre ancien par exemple, l'auteur pourra choisir entre le domaine public Public Domain Dedication and License (PDDL), la licence CC-0 ou la Licence ouverte. Pour la couche des métadonnées entre une licence CC, CC-0, Licence ouverte ou OdbL[176]. La couche des données pourra être protégée par une des licences CC. S'il s'agit de logiciel, l'utilisateur pourra se tourner vers les licences GNU-GPL, BSD, CeCCIL, Apache etc.

L'impact de ces différentes licences est peu détaillé par les différents entrepôts. Le déposant doit pourtant s'assurer que les droits dont il dispose sur ses données sont compatibles avec la politique de l'entrepôt, des embargos peuvent cependant être définis[177]. La compatibilité des licences protégeant les données et les métadonnées par exemple n'est jamais abordée. La réutilisation des données de recherche peur pourtant dépendre de la capacité à réutiliser les informations qui

[173] https://data.csiro.au/dap/legal

[174] Un descriptif des six licences disponibles : http://creativecommons.fr/licences/

[175] Austin C., Brown S., Fong N., Humphrey C., Webster P. « Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements » IASSIST Quaterly (IQ). [En ligne]. 2015. Disponible sur : https://www.rdc-drc.ca/wp-content/uploads/Review-of-Research-Data-Repositories-2015.pdf

[176] http://opendatacommons.org/licenses/odbl/

[177] Assante M., 2016., *op. cit.*, p. 38

gravitent autour[178]. Ce domaine pourrait faire partie des compétences à acquérir pour les professionnels de l'IST, qui deviendraient ainsi en mesure d'accompagner les chercheurs dans ce dédale juridique.

2.4. Validation des dépôts et coûts de publication

L'étape de validation des dépôts ou de modération s'avère parfois complexe. Elle peut être constituée d'une simple vérification de la cohérence des métadonnées et du contenu des fichiers, comme c'est le cas sur HAL. Cependant, cette étape du dépôt peut parfois s'apparenter à du *peer-reviewing*, en poussant la vérification jusqu'à la cohérence des données et leur qualité scientifique. Certains entrepôts ne disposent d'aucune forme de modération, c'est notamment le cas de Figshare mais ils semblent être peu nombreux.

Processus complexe, la modération ne dispose pas encore de cadres aussi établis que le *peer-reviewing*. D'autant plus dans le cas des données de la recherche, produites au sein d'équipes multiples et évoluant au fil de la recherche. De plus, la qualité des données de recherche ne semble pas avoir été définie clairement pour l'instant. On pourrait alors oublier la qualité et n'envisager que la pertinence scientifique[179], mais cela demanderait un travail encore plus approfondi lors de la documentation des données.

Le conditionnement du dépôt par une transaction financière est une des entraves principales au dépôt de données de recherche dans les entrepôts. Les agences de financement recommandant le dépôt des données abordent d'ailleurs souvent cette question dans leur documentation, signifiant parfois que ces frais seront pris en compte en cas d'acceptation[180].

Dans une étude pré-publiée en 2015[181], les deux tiers des entrepôts analysés proposaient la gratuité du dépôt. Le dernier tiers en revanche avait développé une politique de frais. Ceux-ci peuvent être liés au dépôt en tant que tel, notamment selon sa taille, ou encore aux différents services proposés par l'entrepôt, comme la préparation des données ou leur préservation.

Dryad, par exemple, se rapproche du modèle doré de l'accès ouvert en réclamant des frais de publication des données. Dans le cas des publications, le modèle doré a su séduire un large public et se développe de façon exponentielle, avec la création de revues sur ce modèle par de grands groupes éditoriaux. Ce modèle va-t-il se généraliser de la même façon pour les entrepôts de données afin de pallier aux frais de stockage et de curation des données ?

On peut également s'interroger sur le développement de ces modèles et les inégalités qu'ils peuvent mettre en place entre des chercheurs favorisés par leur pays d'origine ou la taille de leur institution et des chercheurs plus isolés. Ces derniers pourront alors lire ou réutiliser les données, si elles sont diffusées en accès libre, mais il leur sera impossible de valoriser leur propre production. Certains entrepôts mettent cependant en place des solutions pour ce cas de figure, Dryad par exemple offre des dispenses de frais de publication des données pour les

[178] *Ibid.*

[179] *Ibid.*

[180] C'est par exemple le cas de la nouvelle politique de l'Organisation pour la Recherche Scientifique (NWO) aux Pays-Bas, *op. cit.*, p. 24

[181] Austin C., 2015, *op. cit.*, p. 41

chercheurs basés dans les pays classifiés par la Banque mondiale comme à faible revenu[182].

3. LA CONSULTATION, LE PARTAGE ET LA REUTILISATION

3.1. Accessibilité (disponibilité, sécurisation et accès)

Même pour des plates-formes en *open access*, l'accès aux données n'est pas forcément totalement libre. Certains entrepôts demandent l'inscription à leur plate-forme avant de pouvoir accéder au contenu[183], voire une participation financière pour accéder aux données[184]. Faire payer les institutions permet l'accès général aux données, sauf pour les chercheurs du Sud dont les institutions ne peuvent pas forcément financer cette charge.

De plus, certains jeux de données, en Astronomie par exemple, nécessitent des moyens importants tant analytiques que matériels, en raison de leur taille. Ces jeux de données pourront-ils vraiment être utilisés par d'autres chercheurs que ceux qui les ont produits ou les grandes infrastructures ? Encore une fois, s'ils sont réutilisés, il est probable que cela sera le fait, non pas de chercheurs isolés, mais de laboratoires de grande taille[185].

La diffusion des données de recherche questionne ainsi l'accès à la méthodologie de production des données, aux données elles-mêmes et aux logiciels permettant *a minima* de les consulter[186]. On peut donc s'interroger sur l'accès aux données pour les individus des pays les moins favorisés.

La mise à disposition d'entrepôts en accès ouvert ou d'archives ouvertes semble alors d'autant plus importante. La diffusion des données de recherche en s'inscrivant dans la voie verte de *l'open access* pourrait permettre de résoudre en partie ce problème[187]. La mise en place de ces infrastructures doit donc s'accompagner d'une perspective sociale et d'une prise en compte des barrières technologiques[188].

C'est également une notion à réinterroger pour les études qualitatives. Les chercheurs des Sciences Humaines et Sociales mettent l'accent sur le temps long d'utilisation de leurs données (parfois 10/15 ans) ainsi que sur l'importance de l'anonymisation. Cependant, ces techniques d'anonymisation peuvent être remises en question concernant les recherches qualitatives, fondamentalement

---

[182] http://datadryad.org/pages/payment

[183] C'est le cas de 78% de l'échantillon de l'étude du RDC. Austin C., *op. cit.*, p. 43

[184] Tene O., Polonetsky J. « Big Data for All: Privacy and User Control in theAge of Analytics ». Northwestern Journal of Technology and Intellectual Property [En ligne]. Vol. 11, n°5,. Disponible sur : http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1191&context=njtip pp.18-19

[185] Tene O., Polonetsky J., *op. cit.*, p. 43

[186] Béranger J. *Les Big data et l'éthique: le cas de la datasphère médicale*. London (UK), Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : ISTE editions, 2016, 2016. 313 p.

[187] Piron F.. « Éthique, développement durable et libre accès ». *I2D–Information, données & documents* [En ligne]. 2016. Vol. 53, n°1, p. 42-43. Disponible sur : < http://www.cairn.info/resume.php?ID_ARTICLE=I2D_154_0034 > (consulté le 1 septembre 2016)

[188] Atkins D. « Revolutionizing Science and Engineering Through Cyberinfrastructure » [En ligne]. 2003. Disponible sur : http://www.nsf.gov/cise/sci/reports/atkins.pdf pp. 10-12

intersubjectives. Elles prônent l'importance à la fois de l'interprétation et de l'hypothèse de départ, celle-ci conditionne les résultats, par exemple dans le cas d'un questionnaire, pour les SHS mais également pour les sciences médicales.

D'autant plus qu'un autre chercheur, n'ayant pas suivi les mêmes biais culturels, aurait peut être choisi des données différentes. C'est ce que Goodman et al. appelaient la reproductibilité inférentielle :

« inferential reproductibility » : « scientists might draw the same conclusions from different sets of studiesand data or could draw different conclusions from the same original data, sometimes even if they agree on the analytical results »[189]

La diffusion des données de recherche pourrait avoir un effet également sur les participants de l'étude, à la fois pour des problèmes d'anonymisation mais aussi sur le comportement des participants eux-mêmes[190]. Quels mécanismes psychiques vont se mettre en œuvre si le participant sait que les données seront diffusées ? La façon de répondre « pour faire plaisir » à la personne en position de supériorité (l'interviewer) s'en trouvera-t-elle décuplée ?

Même l'anonymisation pose question puisque ces procédures devront être conservées précieusement de façon détaillée et dépendront de la nature des données collectées[191]. Selon la granularité nécessaire, le temps consacré à cette activité pourrait être exponentiel tout en introduisant des erreurs ou des approximations, voire même en réduisant à néant l'intérêt de l'étude en dévoyant son sens. Enfin, même en prenant du temps et en évitant ces erreurs, les techniques d'anonymisation ne parviendraient peut-être pas à être opérationnelles. Les méthodes de recoupement statistiques permettent en effet de retrouver assez facilement des caractères uniques, de même que la présence d'extraits de *verbatim* offre la possibilité de reconnaître une personne.

Les conséquences de ces partages peuvent être réelles et parfois ignorées dans les publications d'articles ou de thèses. La diffusion des jeux de données devra donc s'accompagner d'un consentement de la part de la population étudiée ainsi que d'une réflexion approfondie, qui devrait être préalable à l'étude, en envisageant quel sera l'impact de cette diffusion.

Prendre conscience du nécessaire développement de recommandations pour ces procédures ainsi que du temps que pendrait leur application, peut-être plus long que celui de la rédaction de la publication. Il y a donc un risque que les participants boudent les études mais aussi que les chercheurs ignorent les revues trop demandeuses en partage de données[192].

---

[189] Goodman S. N., Fanelli D., Ioannidis J. P. A., What does research reproducibility mean?, *Science Translational Medicine* [En ligne]. 01 juin 2016, Vol. 8, Issue 341, pp. 341. Disponible sur : http://stm.sciencemag.org/content/8/341/341ps12.full

[190] Tsai A. C., Kohrt B. A., Matthews L. T., Betancourt T. S., Lee J. K., Papachristos A. V., Weiser S. D., Dworkin S. L. « Promises and pitfalls of data sharing in qualitative research ». *Social Science et Medicine* [En ligne]. août 2016. Disponible sur : http://dx.doi.org/10.1016/j.socscimed.2016.08.004

[191] *Ibid.*

[192] Tsai A. C., 2016, *op. cit.*, p. 44

3.2. Citation de données et jeux de données

Des principes généraux pour la citation[193] ont été établis par le groupe de travail FORCE11. Ils mettent en avant les données de recherche comme des produits de recherche légitimes et citables dont la citation devrait faciliter le crédit. Une citation de données devrait inclure une méthode pérenne d'identification, reconnue par les communautés scientifique afin de faciliter l'accès aux données et compréhensible par les machines. Les éditeurs encouragent l'utilisation d'identifiants tels que les DOI (*Nature Biotechnology* les autorise depuis 2012[194]) ou l'identifiant donné par l'entrepôt[195] pour citer des données de recherche.

Deux modèles de citation des données de la recherche peuvent être identifiés[196] :

- la citation directe, qui renvoie vers l'entrepôt où sont stockées les données. C'est le modèle adopté par GenBank et valorisé dans des entrepôts thématiques généralistes tels que DataONE, Dryad… Il nécessite le dépôt en entrepôt, l'utilisation d'un format interopérable et la description par des métadonnées.
- La citation d'un *data paper* : les métadonnées nécessaires et le lien vers le jeu de données sont présents dans cette publication. C'est alors le *data paper* et non le jeu de données qui est cité. Ce modèle a été mis en place en géoscience avec la création de *data journals* tels que *Earth System Science Data* ou *Geoscience Data Journal*.

Cette seconde forme est celle dont Thomson Reuters et Elsevier se sont emparés avec la création du Data Citation Index et du DataBase Linking Tool. Ces citations des jeux de données de recherche pourraient ainsi être prises en compte dans l'évaluation de la recherche.

L'étude de Belter a montré que cette diffusion par la citation était internationale. Les modes de citation restent néanmoins variés, malgré le développement de standards de format de citation généralisé. Or, l'existence d'un format seul ne garantit pas la pertinence de la citation.

En matière de citation, le DataCite a produit une recommandation[197] sous la forme suivante :

Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

Concernant la citation de jeux de données dynamiques, trois approches sont décrites par le DataCite :

---

[193] « Joint Declaration of Data Citation Principles - FINAL ». In : FORCE11 [En ligne]. 2013. Disponible sur : https://www.force11.org/group/joint-declaration-data-citation-principles-final

[194] Assante M., 2016., *op. cit.*, p. 38

[195] CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013, *op. cit.*, p. 22

[196] Belter C. W. « Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets ». *PLoS One* [En ligne]. 26 mars 2014. Vol. 9, n°3. Disponible sur : http://dx.doi.org/10.1371/journal.pone.0092590

[197] Andro, M., Cocaud, S., Dzale Yeumo, W. E., Martin, A., Young, C., « Schéma de métadonnées datacite pour la publication et la citation des données de la recherche ». [En ligne]. Traduction de : DataCite Metadata Working Group, *DataCite Metadata Schema for the Publication and Citation of Research Data* (p. 44). 2015. Disponible sur : http://prodinra.inra.fr/record/326796

Citer une partie spécifique (la série de mises à jour effectuées au cours d'une période donnée ou à un secteur particulier du jeu de données)

Citer un instantané (copie du jeu de données entier à un moment précis)

Citer le jeu de données continuellement mis à jour, mais ajouter une date et une heure d'accès pour la citation.

Les options (a) et (b) citent des versions du jeu de données et nécessitent des identifiants uniques. Le choix (c) est controversée car : la citation ne mène pas à l'observation de la ressource telle qu'elle a été citée.

Les entrepôts de données peuvent ainsi proposer cette citation à copier-coller, un outil d'export de la citation, un lien HTML d'intégration ou encore une option de partage sur les réseaux sociaux. Les organisations internationales travaillant sur les données de la recherche travaillent actuellement à améliorer ces possibilités de citation, en particulier pour les jeux de données dynamiques.

Le rôle des entrepôts de données est ainsi d'assurer la stabilité des données, l'accès aux jeux de données et fournir un lien de citation aux auteurs, notamment grâce à un identifiant pérenne. Ces missions doivent s'accompagner de recommandations claires sur la citation des données et le cadre législatif qui les entoure[198]. Afin d'envisager si HAL peut tenir ces rôles auprès de la communauté scientifique française, une étude préliminaire a été réalisée sur les données supplémentaires aux dépôts dans HAL.

[198] Ball A., 2011, *op. cit.*, p. 41

PARTIE 3 : LES DONNEES DE RECHERCHE DANS HAL : ETUDE EXPLORATOIRE

Les entrepôts de données de recherche déjà nombreux proposent une offre variée et, pour certaines disciplines, satisfaisante. S'il n'y pas, pour l'instant, de profil existant dans la littérature du « chercheur déposant ses données », l'utilisation des données déposées peut être le fait de plusieurs profils : chercheur du champ scientifique concerné, chercheur d'une autre discipline ou citoyen lambda. On peut alors s'interroger sur la place du CCSD et plus précisément de HAL dans ce contexte.

En effet, HAL étant une archive ouverte pluridisciplinaire, institutionnelle et nationale, les déposants, l'auteur/e, ou la personne chargée des dépôts dans HAL pour le laboratoire, devront y créer un compte utilisateur avant de réaliser leur dépôt. Une fois validé, le dépôt est mis en ligne et accessible à tout internaute, avec ou sans compte utilisateur.

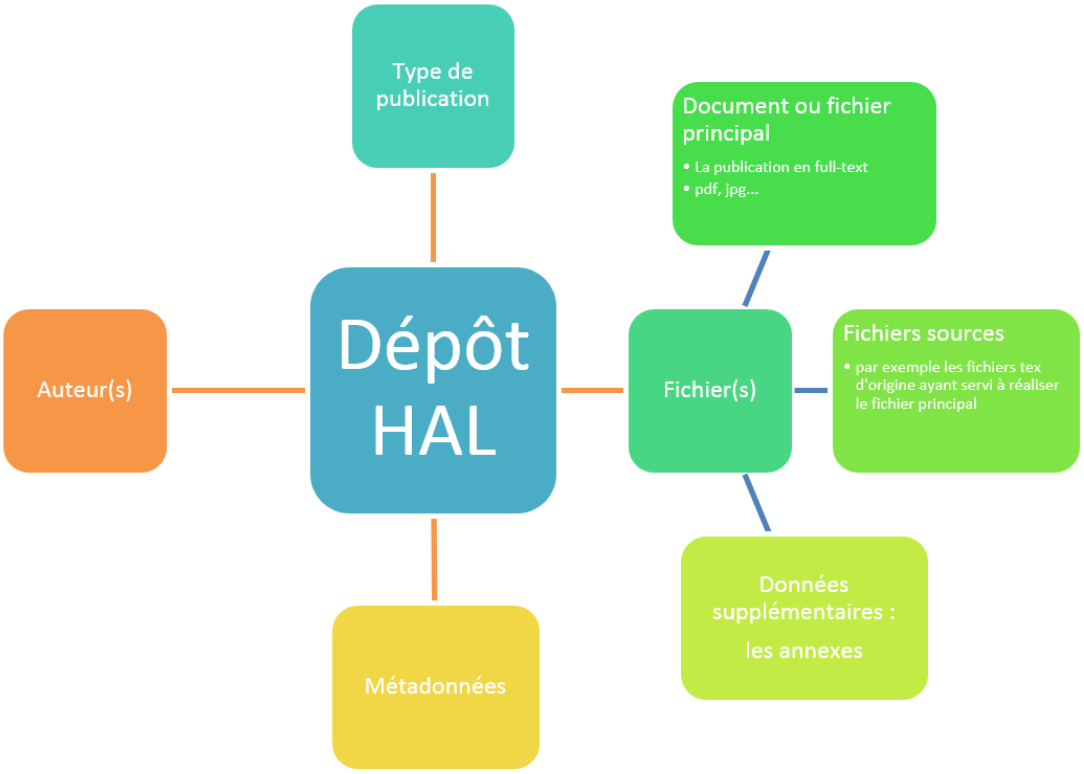


Figure 6 : les différentes étapes du dépôt dans HAL

Il est constitué de différents éléments représentés par les différentes étapes du dépôt qu'on peut observer dans le schéma ci-dessus. La présence de fichiers dans les dépôts est vivement souhaitée mais reste néanmoins non obligatoire, à part pour certains types de dépôts comme les thèses. Ensuite viennent les métadonnées de la publication qui varient également selon le type de publication. Parmi elles, on trouve par exemple les métadonnées auteur, c'est-à-dire le nom du ou des auteurs/es ainsi que leurs affiliations reliées au référentiel AURéHAL.

Dans le cas d'un dépôt avec plusieurs fichiers, il convient de définir leur format : document, fichiers sources ou données supplémentaires. C'est cette dernière catégorie qui nous intéresse tout particulièrement. En effet, cette étude vise à évaluer les usages que les chercheurs peuvent développer à partir des différentes possibilités techniques de HAL en termes de dépôts de données de la recherche.

1. METHODOLOGIE

Les données nécessaires à cette étude ont été extraites de HAL grâce à des requêtes SQL dans la base de données ainsi qu'à l'interrogation par les API, avec l'URL de base `http://api.archives-ouvertes.fr/`. La liste des différentes requêtes est accessible en annexe 3[199]. Les résultats ont ensuite été traités dans des feuilles de calcul grâce au logiciel Excel de la suite Office 2013. On peut noter que, grâce aux API, l'ensemble de HAL est interrogeable, toute personne intéressée peut ainsi reproduire les résultats ci-dessous.

Un type de document représente sa nature d'article, de chapitre d'ouvrage, de thèse ou encore de données de la recherche[200]. HAL dispose d'un type de document « Données de la recherche », clairement identifié lors du dépôt. Le fichier principal sera alors la donnée déposée, ou le jeu de données. Il ne s'agit pas des données étudiées ici.

L'étude se concentre sur les documents déposés en tant que données supplémentaires. Car on peut aussi réaliser un dépôt de données de recherche accompagné d'une publication comme fichier principal, sous un des autres types de document. Les données seront alors déposées comme données supplémentaires au dépôt, c'est à dire en tant qu'annexes. Enfin, on peut déposer seulement des données supplémentaires, sans document principal, dans une notice bibliographique. Ces deux derniers types de dépôts ont été traités de façon identique dans cette étude. Cette définition du statut des fichiers a lieu lors du dépôt ou par la suite, en ajoutant les données supplémentaires.

Dans quelle mesure des données de recherche sont-elles présentes dans ces annexes ? Ces données, si elles existent, sont-elles représentatives de phénomènes disciplinaires ? Pour pouvoir répondre à ces questions, les annexes ont été étudiées à la fois de manière quantitative par le biais des requêtes et de manière qualitative en allant consulter un échantillon aléatoire de vingt dépôts. La taille de cet échantillon ne permet pas de considérer ces résultats comme représentatifs de l'entièreté de HAL et de tous les cas de figure potentiels.

Cette étude visait tout particulièrement à déterminer quels sont les types de documents concernés par ces annexes et dans quelle mesure celles-ci sont constituées de données de la recherche. De même, elle interroge les formats de ces annexes, afin d'envisager quels types de fichiers constituent ce corpus. Ainsi, il s'avère que si des données de recherche sont présentes, il est possible de déterminer sous quels formats de fichier. Enfin, l'extraction des domaines donne un aperçu des disciplines concernées par le dépôt d'annexes. L'identification de ces pratiques pour les fichiers annexes servira de support à la réflexion sur le thème des données de la recherche dans HAL pour le CCSD.

[199] Voir annexe 3, p. 74

[200] Voir annexe 4, p. 75

2. RESULTATS DE L'ETUDE

2.1. Le type de document « Données de recherche »

Historiquement, c'est le portail MédiHAL, construit en 2010 avec le soutien du TGE Adonis, qui servait d'interface de dépôt non pas pour des données de recherche mais pour des images issues d'activités de recherche. Un portail est un sous-ensemble de l'archive qui rassemble généralement toutes les publications d'une même institution de recherche. En 2014, MédiHAL s'est étendu aux vidéos, sons et cartes et il est devenu possible de déposer des données de recherche dans ce portail depuis l'URL générique de HAL[201]. Les données de recherche ont ainsi intégré l'interface de HAL. Cette initiative ayant été lancée par des chercheurs en Sciences Humaines et Sociales, on trouve une majorité d'images de ces disciplines dans ce corpus de données.

Cependant, on observe une diversification des disciplines sans doute liée à celle des types de documents. Les mathématiques, de même que la physique, déposent des images, des sons et des vidéos. Dans ce dernier cas, les vidéos semblent avoir été déposées dans leur majorité par l'Institut Fourier. On peut d'ailleurs questionner leur nature de données de la recherche. Il s'agit pour la plupart de cours, à l'image de ce que l'on peut trouver dans JOVE[202]. Ces démonstrations mathématiques sont-elles des données de la recherche ? Il semble délicat de répondre à cette question sans mener des entretiens auprès de mathématiciens.

De même, l'offre de données de recherche dans HAL couvre-t-elle vraiment les besoins des communautés scientifiques ? Plusieurs différences peuvent être identifiées entre cette offre et les entrepôts multidisciplinaires de données de la recherche. En particulier, les dépôts sont ici limités à quatre types de médias seulement : images, vidéos, sons et cartes. D'autant plus que les formats utilisés pour diffuser et archiver ces données ne suffiraient peut-être pas à les exploiter en tant que données. On peut également s'interroger sur la visibilité de HAL en tant qu'archive de données de la recherche ? Le portail MédiHAL est-il bien identifié dans les communautés de recherche ? Le dépôt de données de la recherche est-il connu ? Enfin, un des avantages certains de l'utilisation du type de document « Données de la recherche » pour le dépôt de données est l'assurance d'un archivage pérenne auprès du CINES, ce qui n'est pas assuré dans le cas d'un dépôt en annexe, seuls les documents principaux étant pris en compte par le CINES.

2.2. Les annexes dans HAL : des données de recherche ?

Le premier constat de cette étude est la part minime des dépôts avec annexes dans HAL qui ne concernent que 1,2% de l'ensemble des dépôts. En comparant les dépôts avec annexes et les dépôts sans annexes, on constate donc que la différence de proportion est immense. Ces résultats sont ainsi à considérer avec précaution.

[201] Dépôt de données de recherche via l'interface, annexe

[202] Holmes J. « How Methods Videos Are Making Science Smarter ». *The New Yorker* [En ligne]. 28 août 2015. Disponible sur : http://www.newyorker.com/tech/elements/how-methods-videos-are-making-science-smarter

2.2.1. A l'intérieur des fichiers

L'échantillon de vingt publications avec annexes, s'est avéré varié. On y trouve des dépôts réalisés dans des portails aussi bien que dans HAL générique, des types de publication divers ainsi que des dates de dépôts étalées. Les annexes sont particulièrement différentes les unes des autres. Certains dépôts n'ont qu'une seule annexe mais la plupart en ont plusieurs. En termes de formats, on trouve des images, des vidéos, des sons, des fichiers textes, des feuilles de calcul ainsi que des présentations. La distribution selon le type de document[203] est la suivante.

TYPDOC	Nombre d'occurrences dans l'échantillon
ART	6
COMM	7
COUV	1
HDR	3
MEM	1
THESE	2

Figure 7 : Les différents types de document représentés dans l'échantillon

Un premier point intéressant est la présence de fichiers de type *supplementary materials* ou *read-me*, détaillant la méthodologie de recherche. Ainsi, des publications ont parfois été déposées selon un modèle particulièrement complet : le fichier principal avec ses métadonnées avec des données de recherche et la méthodologie en annexe.

Ensuite, cet échantillon proposait plusieurs présentations, notamment en relation avec les thèses. On peut supposer que thèses et HDR sont des types de publication où des présentations de soutenance seront régulièrement présentes.

Si des données peuvent être présentes à l'intérieur de ces fichiers, elles le sont de façon transformée et peu ré-exploitable. Ceci amène le dernier point. En effet, un nombre important de fichier PDF a été constatée. Parfois un fichier .PDF nommé « données » reprend divers graphiques ainsi qu'une feuille de calcul. Cependant, il s'agit là encore d'un format image où les données sont difficilement ré-exploitables. Enfin, de même que dans MédiHAL, les images, les vidéos et les sons sont-ils d'assez bonne qualité pour être réutilisables dans le cadre d'autres recherches ? Car, si l'on trouve bien des données sous forme de feuilles de calcul, la plupart d'entre elles se présentent de façon transformée. Les publications liées n'ont pas été étudiées en détail, nous ne savons donc pas si la méthodologie utilisée y était décrite.

On peut alors se demander quel type de donnée un entrepôt souhaite préserver. Car dans cet échantillon, il s'agit la plupart du temps d'illustrations par le biais d'images ou de graphiques. Finalement, peu de publications offraient les clefs de compréhension et d'extraction des données et donc des possibilités d'évaluation et de reproductibilité de ces recherches.

2.2.2. Les types de publications

HAL accepte en théorie seulement treize types de publication ou types de document dans son archive, en plus des quatre présentés pour MédiHAL. Dans les

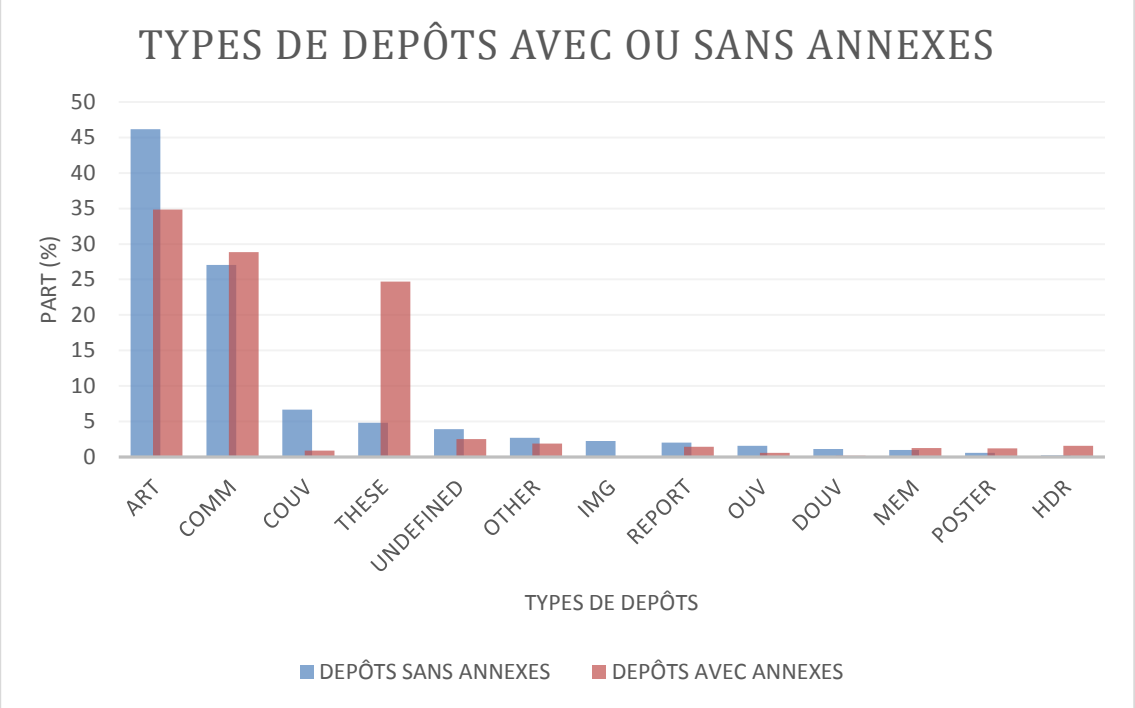
[203] Voir annexe 3, p. 74

faits, on trouve d'autres types de publication dans les portails, ceux-ci ayant pu éprouver le besoin de mettre à disposition un type de document particulier à leurs utilisateurs, par exemple les mémoires d'étudiants dans les portails Dumas et Memsic. Ces types de publication sont au nombre de 23 en tout (l'annexe 4 est un tableau récapitulatif de ceux-ci accompagnés de leur abréviation dans la base de données[204]).

En comparant la part en pourcentage des types de dépôts avec et sans annexes, plusieurs tendances émergent (voir illustration 6). Les articles dans des revues ainsi que les communications dans les congrès sont les publications les plus représentées dans HAL. Il n'est donc pas surprenant qu'elles soient celles comptabilisant le plus d'annexes. Cependant, on constate également que la représentation dans HAL ne garantit pas la présence d'annexes, puisque les chapitres d'ouvrage n'en ont que très peu. La présence d'annexes semble donc bien varier selon le type de publication.

Figure 8 : Les types de dépôts dans HAL sans annexes (bleu) et avec annexes (rouge)

Ainsi, les thèses et HDR sont les catégories où la présence d'annexes est la plus forte par rapport à leur représentation dans l'archive, notamment les thèses. D'après l'échantillon cité ci-dessus, on peut supposer que nombre de ces annexes se



composent de présentations de soutenances de thèses. On a également constaté dans l'étude qualitative que certains dépôts de thèse comportent des articles de revue en annexe. Il n'est pas possible de quantifier ce phénomène mais il est peut-être récurent dans le cas d'articles dont l'éditeur refuse le dépôt en archive ouverte. Le dépôt accompagnant la thèse permet donc de diffuser l'article malgré la politique éditoriale.

[204] Tableau des principaux types de document de HAL, voir annexe 4, p. 75

2.2.3. Une myriade de formats de fichiers

L'article écrit cette année à propos des entrepôts de données par Assante, Candella et Castelli[205] mettait en avant la diversité de formats de fichiers auxquels ces entrepôts doivent faire face. En effet, il semble impossible dans le cas d'un entrepôt multi-disciplinaire d'étudier de façon assez fine la typologie de données de chacune des disciplines qu'il pourra rencontrer afin d'estimer à quels formats il aura affaire.

EXTENSION	NOMBRE DE FICHIERS PAR EXTENSION
pdf	10632
jpg	3826
png	3394
xml	2061
ppt	1405
doc	1280
gif	1132
avi	973
mp3	881
wav	750
tif	571
xls	494
txt	470
docx	374
mov	341
tiff	324
mp4	281
zip	258

Figure 9 : 14 des extensions retrouvées le plus souvent dans les annexes de HAL sur un total de 230

Ce postulat se trouve confirmé ici puisque la base de données de HAL recense 230 extensions de fichiers différentes dans les annexes aux dépôts. Cependant, il convient de nuancer ce chiffre, sans doute au-dessus de la réalité. Car ces extensions de fichiers peuvent être erronées. On trouve par exemple un nombre important d'annexes dont l'extension est .0. Il s'agit probablement d'une erreur de nommage de fichier, que le système aura considéré comme une extension, le déposant ayant renseigné le nom de son fichier avec un point suivi

[205] Assante M., 2016., *op. cit.*, p. 38

d'un zéro. De même, certaines annexes sont très anciennes. La liste des formats de fichier acceptés aujourd'hui dans HAL est plus restreinte[206].

Ensuite, la présence importante des fichiers de type .pdf se confirme puisque c'est le type d'annexe le plus fréquent. On peut le considérer comme un format image peu ré-exploitable et se caractérise par le manque d'information dont nous pouvons disposer sur son contenu. Il en va de même pour les fichiers à l'extension .zip, ces archives pouvant être constituées de n'importe quel type de fichier.

Enfin, certaines de ces extensions pourraient être catégorisées selon le média auquel elles renvoient : images, vidéos, textes, codes sources etc. Il pourrait être intéressant de dater les dépôts d'annexes de type image, par exemple, afin de comparer leur courbe temporelle avec l'apparition du type de document image et de MédiHAL afin d'évaluer si certains chercheurs ont pu modifier leurs pratiques depuis.

Les annexes dans HAL sont donc bien représentatives de la diversité que l'on peut retrouver dans d'autres entrepôts pluridisciplinaires tels que Figshare où le contrôle sur le format de fichier est peu contraignant.

2.2.4. L'approche disciplinaire

Pour finir, une comparaison entre les dépôts avec annexes et les dépôts sans annexe dans HAL a été réalisée par le biais des disciplines scientifiques. Ainsi, comme le montre le tableau ci-dessous, si la communauté des Sciences Humaines et Sociales est la plus représentée dans HAL, la pratique du dépôt d'annexes n'y semble pas particulièrement répandue, de même que pour la Physique.

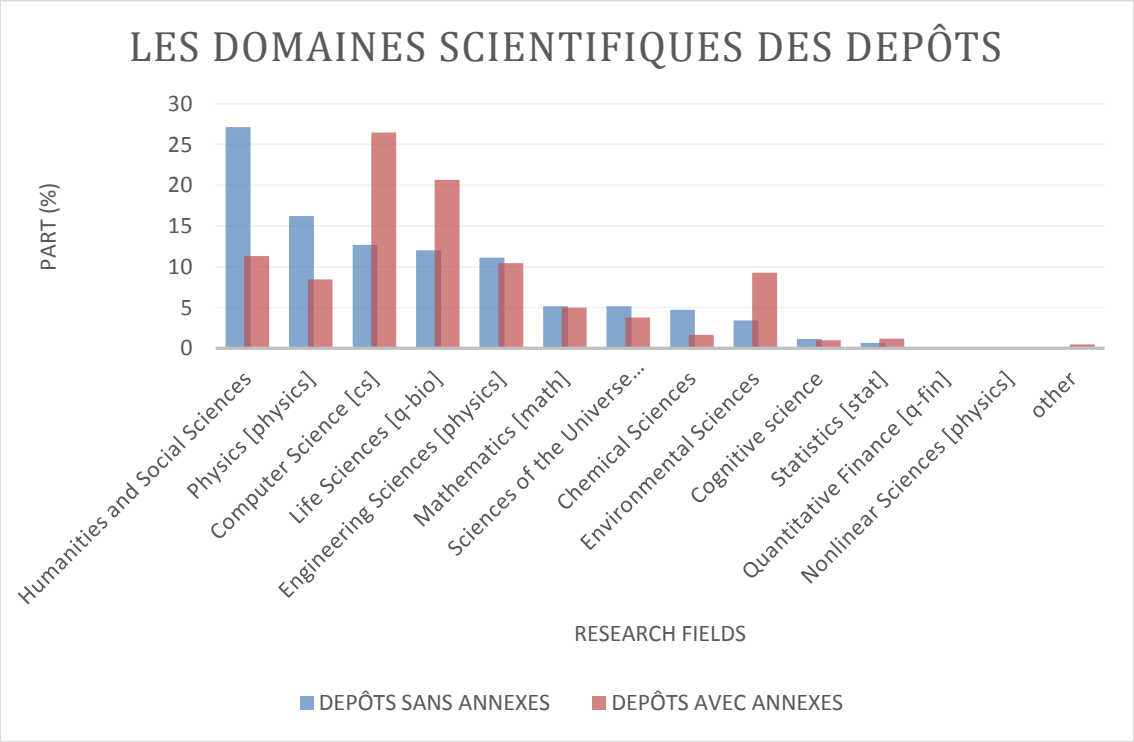


Figure 10 : Les champs disciplinaires des dépôts sans annexe (bleu) et des dépôts avec annexes (rouge) dans HAL

[206] Voir annexe 5, p. 75

À l'inverse, l'Informatique, les Sciences du Vivant ainsi que les Sciences de l'Environnement accompagnent une partie de leurs dépôts d'annexes et donc, dans une certaine mesure, de données de la recherche. On peut alors se demander si ces dépôts d'annexes ont été effectués dans le cadre d'une acculturation aux *supplementary mateirals* et aux données de la recherche. Les deux dernières disciplines citées possèdent toutes deux des entrepôts bien identifiés et les chercheurs qui les représentent mobiliseraient peut-être une volonté de lier leurs publications avec leurs données.

Cependant, cette étude n'a été menée que sur les domaines principaux de HAL, sans détailler les sous-domaines. Or, les pratiques en termes de données de recherche semblent être représentatives de communautés de recherche parfois réduites à une sous-discipline ou étendues à une inter-discipline particulière.

En détaillant, par exemple, les SHS, la catégorie disciplinaire de loin la plus vaste que l'on peut trouver dans HAL, on constate des disparités importantes entre les disciplines.

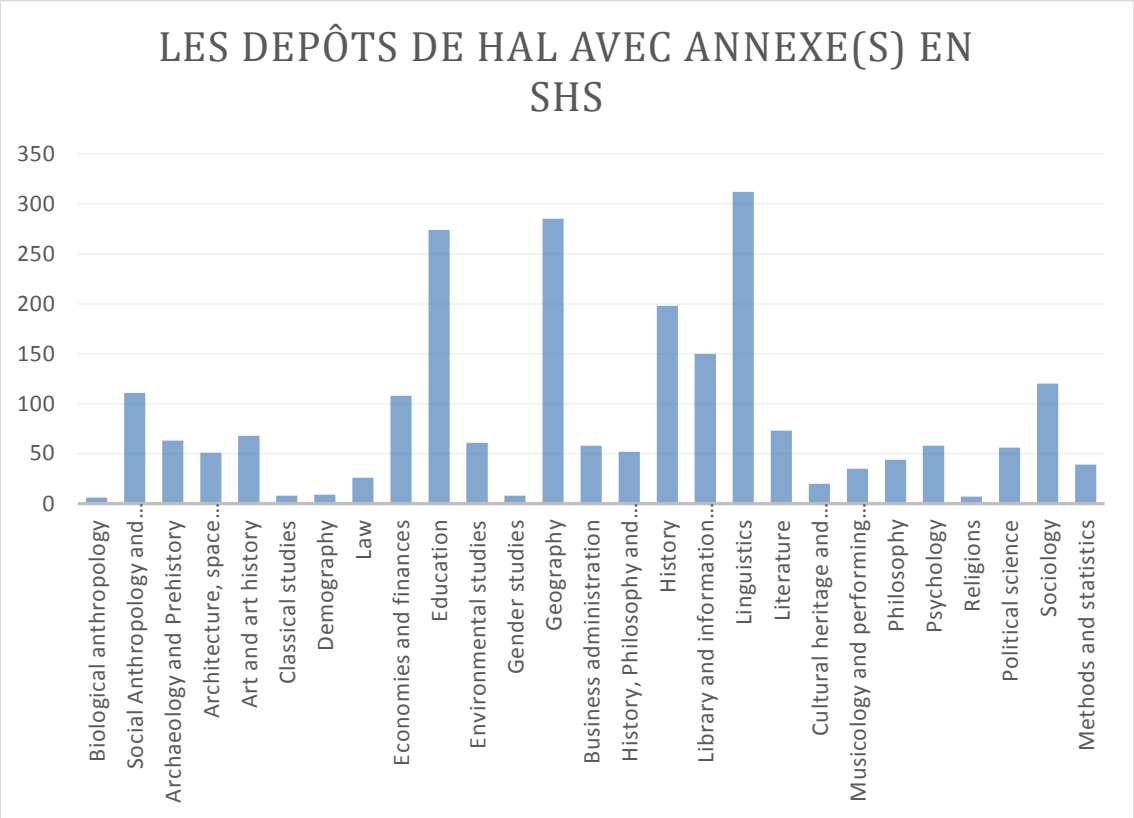


Figure 11 : Les dépôts avec annexes par sous-disciplines en SHS

Ainsi, en SHS, ce sont la Linguistique, les Sciences de l'éducation, la Géographie, l'Histoire et les Sciences de l'Information qui déposent le plus d'annexes. Il reste cependant délicat d'analyser ce résultat en l'absence de consultation des fichiers.

3. DES DONNEES SUPPLEMENTAIRES MULTIPLES ET NON STRUCTUREES

En conclusion, HAL comporte à la fois des documents identifiés comme données de recherche et des données de recherche sous la forme de *supplementary*

materials. Celles-ci représentent cependant une part minime des dépôts. Il n'est pas possible à l'heure actuelle d'évaluer leur nombre, notamment parce l'évaluation de ce qui est ou n'est pas une donnée de recherche demande une étude qualitative approfondie des fichiers. De même, une dimension chronologique n'a malheureusement pu être intégrée aux requêtes. Il aurait pourtant été intéressant de pouvoir comparer les évolutions dans le temps des dépôts dans MédiHAL par exemple.

Il faudrait ainsi étendre l'étude qualitative à un échantillon aléatoire plus large ou bien à un échantillon non aléatoire précis. Il paraît souhaitable d'étudier en profondeur une discipline ou un type de document particulier et notamment d'approfondir les Sciences Humaines et Sociales dont le domaine principal ne permet pas de percevoir les spécificités propres à chacune des disciplines. De façon générale, l'étude pourrait être poursuivie en mettant en parallèle les pratiques des différents représentants de ces disciplines, leur culture disciplinaire en termes de données et leur volonté propre par le biais de la littérature épistémologique et d'entretiens.

De même, il est impossible aujourd'hui de savoir si ces données identifiées comme données supplémentaires ont été réutilisées ou même consultées. Les statistiques de HAL ne donnent des informations que sur la consultation de la notice ou le téléchargement du document principal. Cette étude préliminaire sur les annexes déposées sur HAL a ainsi montré que l'archive ouverte française accueille bien des données de recherche selon le modèle des *supplementary materials*. Il ne s'agit donc pas de jeux de données structurés, référencés et archivés. Doit-on pour autant transformer HAL en archive ouverte de données de recherche ?

4. QUELLES EVOLUTIONS ENVISAGER POUR HAL ?

4.1. Renforcer les collaborations existantes et faire fructifier les échanges internationaux

Le CCSD est aujourd'hui impliqué dans les groupes de travail français sur les données de la recherche, notamment à travers la BSN[207]. Cependant, étant donné l'avancement de certains pays comme le Royaume-Uni ou les Pays-Bas sur cette problématique, il semble nécessaire de développer aujourd'hui une vision internationale, à l'image de la recherche elle-même multicolore. Lors de rencontres à l'étranger, la communauté de l'IST semble intéressée par HAL et sa future politique en matière de données de recherche, d'autant plus qu'il s'agit d'une infrastructure nationale. D'ailleurs, les deux possibilités qui s'offrent à HAL demandent toutes deux la mise en œuvre de partenariats forts avec d'autres structures, et notamment les entrepôts de données.

4.2. Favoriser le lien entre publication et données

Travailler sur le lien entre les différentes productions gravitant autour de la recherche, en particulier en liant publications et données pourrait être une des orientations du CCSD. HAL s'attache particulièrement à la diffusion ainsi qu'à la préservation des publications. Si la transformation en un entrepôt de données de

[207] http://www.bibliothequescientifiquenumerique.fr/bsn-10-donnees-de-la-recherche/

recherche national semble aujourd'hui irréalisable et pas forcément souhaitable, aux vues des moyens humains dont dispose le CCSD, lier ces publications avec leurs données pourrait s'avérer intéressant de plusieurs manières.

D'une part, cela correspond aux missions de valorisation et de communication scientifique de l'archive. Ensuite, le CCSD maintient un dialogue avec plusieurs organisations et institutions à l'international. Ainsi, les identifiants ORCID sont en cours d'implémentation dans le système pour permettre aux chercheurs de les lier avec leur IdHAL. Ce travail nécessite le maintien d'un dialogue avec ORCID. Il en va de même pour OpenAire.

Le développement d'un module liant les publications à leurs données serait une occasion pour faire éclore des partenariats avec les entrepôts de données de la recherche, favorisant ainsi la mise en œuvre de la feuille de route 2016-2020 du CCSD, notamment en termes de communication, dont deux des objectifs sont le développement de la communication institutionnelle et de la visibilité à l'international.

Les données pourraient ainsi apparaître sur la notice sous la forme d'un lien vers l'entrepôt de données. On pourrait également envisager la création de notices bibliographiques propres aux données. Le CCSD pourrait également mettre en place un service de conseil, afin d'orienter les usagers en jonglant entre les entrepôts avec lesquels le CCSD serait en partenariat, et donc avec la possibilité de lier les données, et les autres entrepôts, en prenant en compte, dans une certaine mesure, la discipline du chercheur déposant : quel sera l'entrepôt le plus adapté pour lui, sa publication et ses données ?

Avec ce dernier point, HAL entrerait dans un service de type conseil. Aller aussi loin dans la démarche ne semble pas réalisable aujourd'hui étant donné la surcharge de travail que cela produirait pour le support en termes de recherche d'information sur les entrepôts. Mais HAL s'est construite comme archive ouverte grâce à la mobilisation de communautés scientifiques. Cette démarche a produit une infrastructure sensible à ses utilisateurs, qui contribue à l'*open access*, aussi bien pour des communautés étendues que pour des sous-disciplines plus réduites. Or, les entrepôts de données se construisent également grâce à l'action des communautés scientifiques. Si la littérature a tendance à se focaliser sur les gros réservoirs de données de la recherche, il existe de nombreuses sources de petite taille, par exemple traitant d'un sujet hyperspécialisé, partagées seulement entre quelques personnes qui se sont prises en charge. HAL pourrait apporter son expertise à ces projets, en tant qu'outil national.

Si cette solution est retenue, l'orientation vers d'autres entrepôts demandera l'établissement d'une grille qualitative décrivant ce que l'on peut attendre comme qualité de service de la part de ces entrepôts.

4.3. HAL comme archive ouverte de données

Une seconde possibilité serait de doter HAL d'un volet entrepôt de données de la recherche. Cela nécessiterait un travail de typologie des données de recherche approfondi, de même que l'évaluation des besoins des communautés utilisateurs en termes de services. Il faudrait ainsi réaliser, une grille évaluative des entrepôts de données de la recherche existants et étudier les modèles mixtes disposant à la fois d'une archive ouverte de publications et d'un entrepôt de données.

Au-delà des études préalables à réaliser, la mise en œuvre du projet aurait des implications conséquentes pour le CCSD. En effet, si l'on reprend les huit critères évoqués précédemment pour les entrepôts de données, de nombreux changements et développements s'avéreraient nécessaires.

La liste des formats acceptés par HAL devrait évoluer afin de correspondre à la fois aux pratiques des déposants en matière de données et aux recommandations en matière de formats pérennes[208].

Concernant la description des données, des choix seraient à faire en termes de standards de métadonnées. Soit en mettant en place un standard unique, identique à celui du DataCite par exemple, soit en proposant plusieurs standards, qui pourraient s'adapter aux domaines et sous-domaines scientifiques.

HAL propose déjà plusieurs licences à ses utilisateurs pour leurs dépôts : CC, Licence ouverte, Domaine public ou Copyright (Tous droits réservés). Les premières sont déjà décrites dans la documentation[209] mais un travail de communication plus poussé autour du droit semble important afin de sensibiliser les chercheurs déposants et utilisateurs.

La publication dans HAL est aujourd'hui entièrement libre et gratuite. Cependant, la question des coûts de stockage pourra se poser à l'avenir, notamment dans le cas des vidéos. La limite de taille d'un fichier est aujourd'hui d'1 Go, certaines données peuvent être bien plus importantes que cela. Doit-on imposer une limite de taille et privilégier la gratuité ? C'est le modèle choisi par Zenodo, qui rappelle néanmoins que le stockage n'est pas gratuit et encourage les déposants les plus importants à réaliser des donations. Une autre possibilité serait de créer un modèle économique en demandant des frais de dépôt des données lorsque les jeux de données dépassent une certaine limite, à l'image de Dryad.

Se pose alors la question de la validation de ces données. Si un service d'hébergement est mis en œuvre pour les données, celles-ci devront être modérées. Cela demandera l'établissement d'un plan de formation pour les modérateurs ainsi que des procédures de modération différentes de celles existant aujourd'hui.

De plus, il faudrait assurer l'archivage pérenne des données de la recherche. Le CINES, partenaire du CCSD pour l'archivage des fichiers principaux de HAL, a déjà commencé à travailler sur ces questions, notamment à travers Huma-Num[210]. Il faudrait donc redéfinir ce partenariat.

On peut donc s'interroger sur la transformation de HAL en entrepôt de données, notamment lorsque l'on compare les moyens du CCSD et ceux que nécessiterait ce type de projet. De plus, des entrepôts généralistes tels que Zenodo existent déjà. Y a-t-il un intérêt à réaliser ce doublon ? Les communautés de recherche françaises ont-elles réellement besoin d'un entrepôt de données national ? D'autant plus que les recommandations des agences de moyens et éditeurs semblent encourager les chercheurs à déposer dans des entrepôts hyper-spécialisés. Le dépôt en entrepôt généraliste semble souvent envisagé comme un choix par défaut pour palier à un manque, notamment parce qu'il ne permet pas forcément de traiter en profondeur les besoins de chacune des disciplines. D'autant plus que HAL n'est pas forcément connu de tous les chercheurs. Or, il semble

[208] Le CINES, par exemple, maintien le valideur de format FACILE : https://facile.cines.fr/

[209] https://hal.archives-ouvertes.fr/page/telechargement-des-fichiers#informations

[210] https://www.cines.fr/archivage/donnees-scientifiques/

nécessaire d'établir une relation avec ces communautés avant de développer une offre de service aussi ambitieuse.

Ces deux scénarios ne sont pas incompatibles l'un avec l'autre et pourraient être développés aussi bien l'un que l'autre, séparément ou en parallèle. On pourrait également envisager d'héberger les données de disciplines orphelines, ne disposant pas d'entrepôt de données bien identifié. Cependant, la priorité serait de travailler au référencement des données de recherche déposées dans les entrepôts disciplinaires et généralistes.

CONCLUSION

Sujet complexe et passionnant, les données de la recherche, telles qu'elles ont été abordées au sein de ce stage ont présenté un certain nombre de difficultés. Les premières ont été de faire face au nombre de publications sur le sujet et de parvenir à prendre en compte les attentes de chacun des acteurs qui gravitent autour de cette problématique. Ainsi, dans le contexte de l'archive ouverte HAL, la réflexion doit s'engager entre la politique de l'État souverain, la recherche de visibilité des EPST, les objectifs scientifiques des chercheurs, les défis techniques soulevés par les informaticiens ou encore les compétences de description et d'organisation des professionnels de l'IST. Chacun pouvant être exécutant, décisionnaire, utilisateur ou encore financeur, voire combinant plusieurs de ces rôles.

Les données de recherche sont un des sujets brûlants du monde de l'IST. En parallèle à la littérature abondante sur le sujet, la diversité de nature des données de recherche ainsi que l'organisation de la recherche elle-même ont produit un foisonnement de formats, standards et lieux de stockage pour ces données. Il est indéniable que les solutions existantes, tant en termes d'entrepôts que de standards de métadonnées ou de citation ne sont pas parfaites. Le défrichage bibliographique et l'investissement dans les organismes nationaux et internationaux pour travailler à leur découverte et amélioration promet d'être long et harassant.

Aujourd'hui, les institutions semblent tentées de passer outre afin de créer chacune leur solution propre, qui serait en phase avec leurs publics. Cependant, se laisser convaincre par les sirènes du projet unique pourrait s'avérer tout aussi chronophage. D'autant plus que l'on peut s'interroger sur la pertinence de certaines focales. La recherche, cet écosystème éclaté et international, peut-elle vraiment être représentée par une institution ?

Le choix et la description des données sont les éléments cruciaux des projets portant sur les données de la recherche. D'autant plus qu'il s'agit des activités les plus gourmandes en temps à la fois pour les chercheurs et les spécialistes de l'IST. Ainsi, le choix des données, d'un format de description puis la mise en œuvre des métadonnées et la rédaction d'une documentation méthodologique semblent constituer des étapes obligatoires afin de réaliser une diffusion optimale.

HAL héberge déjà quelques données de recherche. La solution envisagée pour l'instant consisterait en la mise en place d'un service de référencement de données de recherche. HAL pourrait ainsi lier données de recherche et publication. Ces deux objectifs ne sont pas incompatibles avec l'hébergement de données de la recherche : il est possible de référencer des jeux de données dans HAL, de créer des liens avec les entrepôts où sont stockées les données et des publications stockées dans HAL, tout en intégrant à HAL une dimension entrepôt de données en *open access*. Cependant, cette solution nécessite de trouver un équilibre entre la facilitation du dépôt des données et le temps que demande leur bonne description[211] et ainsi une augmentation conséquente des moyens octroyés au CCSD. Il semble plus réaliste d'envisager pour l'instant un modèle similaire à

[211] GilPress. Data Scientists Spend Most of Their Time Cleaning Data, What's The Big Data? [En ligne]. What's The Big Data?. 1 mai 2016. Disponible sur : https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/

l’Australie ou au Royaume-Uni mêlant référencement de données et conseil aux chercheurs afin de les rediriger vers des entrepôts sûrs.

Ainsi, il pourrait s’avérer intéressant d'étudier plus en détail les offres nationales de ces deux pays mais également du Canada, des États-Unis ou du Pays-Bas. C'est à dire de poursuivre cette étude par une veille bibliographiques et surtout par le développement d'un dialogue approfondi avec les institutions, organisations, associations et consortium français, étrangers et trans-nationaux.

SOURCES

4TU.Centre for Research Data, [en ligne] http://researchdata.4tu.nl/en/publishing-research/data-description-and-formats/

ANDS - Australian Nationale Data Service, [en ligne] http://www.ands.org.au/

ARCHIMER - Archive Institutionnelle de l‟Institut Français de Recherche pour l'Exploitation de la Mer, [en ligne] http://archimer.ifremer.fr/

BLOG DU CCSD, [en ligne] https://blog.ccsd.cnrs.fr/

CARSAI, [en ligne] http://dictionary.casrai.org/Dataset

CINES - Centre Informatique National de l'enseignement Supérieur, [en ligne] https://www.cines.fr/

Creative Commons, [en ligne] http://creativecommons.fr/licences/

Darwin Core, [en ligne] http://rs.tdwg.org/dwc/

DataCite, [en ligne] https://www.datacite.org/

DataCite Schema, [en ligne] https://schema.datacite.org/meta/kernel-3/

Data Search, [en ligne] https://datasearch.elsevier.com/

Digital Science, [en ligne] https://www.digital-science.com/about-us/

Dryad, [en ligne] http://datadryad.org/

CIRSO, [en ligne] https://data.csiro.au/dap/legal

COD - Crystallography Open Database, [en ligne] http://www.crystallography.net/cod/

CODATA - International Council for Science : Committee on Data for Science and Technology, [en ligne] http://www.codata.org/

CLARIN, [en ligne] http://www.clarin.eu/content/component-metadata

DANS - Data Archiving and Networked Service, [en ligne] https://dans.knaw.nl/en

DARIAH-EU - Digital Research infrastructure for the Arts and humanities, [en ligne] http://www.dariah.eu /

Databank for Chinese Studies, « Data usage policies », [en ligne] http://www.usc.cuhk.edu.hk/Eng/UsagePolicy.aspx

DCC - Data Curation Center, [en ligne] http://www.dcc.ac.uk/

DIME-SHS Sciences Po, [en ligne] http://www.sciencespo.fr/dime-shs/

DMP Online, [en ligne] https://dmponline.dcc.ac.uk/

DMP Tool, [en ligne] https://dmp.cdlib.org/

Elsevier, [en ligne] https://www.elsevier.com/

EPFL - École Polytechnique Fédérale de Lausanne, [en ligne] http://library.epfl.ch/cms/lang/fr/pid/119191

Episciences, [en ligne] http://episciences.org/

FACILE - CINES, [en ligne] https://facile.cines.fr/

Figshare, [en ligne] https://figshare.com/

HAL - Archive pluridisciplinaire nationale, [en ligne] https://hal.archives-ouvertes.fr

HEFCE - Higher Education Funding Council for England, [en ligne] http://www.hefce.ac.uk/

ICSU World Data System, [en ligne] https://www.icsu-wds.org/

INIST-CNRS [en ligne], http://www.inist.fr

JISC - Joint Information Systems Committee, [en ligne] https://www.jisc.ac.uk/

NYU Health Sciences Library - YouTube, [en ligne] https://www.youtube.com/user/nyuhsl

ODbL - Open Data Commons Open Database License, [en ligne] http://opendatacommons.org/licenses/odbl/

OLAC - Open Language Archives Community, [en ligne] http://www.language-archives.org/

OpenAIRE - Open Access Infrastructure for Research in Europe, [en ligne] https://www.openaire.eu/

OpenDOAR - The Directory of Open Access Repositories, [en ligne] http://www.opendoar.org/index.html

PANGAEA, [en ligne] https://www.pangaea.de/

PRODINRA - Archive ouverte de l'Institut National de la Recherche Agronomique (INRA), [en ligne] http://prodinra.inra.fr/

Research Councils UK, [en ligne] http://www.rcuk.ac.uk/

RDA - Research Data Alliance, [en ligne] https://rd-alliance.org/node

Re3data, [en ligne] http://www.re3data.org/

Research Data Australia , [en ligne] https://researchdata.ands.org.au/

Reseau QUETELET, [en ligne] http://www.reseau-quetelet.cnrs.fr/spip/

ROAR - Registry of Open Access Repositories, [en ligne] http://roar.eprints.org/

Sciencesconf, [en ligne] https://www.sciencesconf.org/

Science Direct, [en ligne] http://www.sciencedirect.com/

Springer Nature, [en ligne] http://www.springernature.com/gp/group/data-policy/helpdesk

The Chinese university of Hong-Kong, [en ligne] http://www.usc.cuhk.edu.hk/Eng/AboutDCS.aspx

ThemoML at NIST Thermodynamic Research Center, [en ligne] http://trc.nist.gov/

UK Data Service, [en ligne] https://www.ukdataservice.ac.uk/

Wellcome, [en ligne] https://wellcome.ac.uk

BIBLIOGRAPHIE

Austin C., Brown S., Fong N., Humphrey C., Webster P. « Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements » IASSIST Quaterly (IQ) [en ligne]. 2015. Disponible sur : < https://www.rdc-drc.ca/wp-content/uploads/Review-of-Research-Data-Repositories-2015.pdf > (consulté le 22 juin 2016)

Austin C. C., Bloom T., Dallmeier-Tiessen S., Khodiyar V. K., Murphy F., Nurnberger A., Raymond L., Stockhause M., Tedds J., Vardigan M., Whyte A. « Key components of data publishing: using current best practices to develop a reference model for data publishing ». *International Journal on Digital Libraries* [en ligne]. 20 juin 2016. Disponible sur : < http://dx.doi.org/10.1007/s00799-016-0178-2 > (consulté le 22 juin 2016)

Ball A., Greenberg J., Jeffery K., Koskela R. « RDA Metadata Standards Directory Working Group: Final Report ». 2016. Disponible sur : https://rd-alliance.org/system/files/MSDWG-Final-Report.pdf

Belter C. W. « Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets ». *PLoS One* [en ligne]. 26 mars 2014. Vol. 9, n°3. Disponible sur : < http://dx.doi.org/10.1371/journal.pone.0092590 > (consulté le 12 mai 2016)

Béranger J. *Les Big data et l'éthique: le cas de la datasphère médicale*. London (UK), Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : ISTE editions, 2016, 2016. 313 p.

Blanc I., Gaspin C., Hologne O., Partage des données de larechercheSéminaire de lancement de la mise en œuvre de la politique Inra, INRA [en ligne]. 2013. Disponible sur : https://www6.inra.fr/reseau-in-ovive/content/download/3209/32325/version/1/file/Accueil_atelier_IN-OVIVE.pdf (consulté le 18 avril 2016)

Borgman C. « The Conundrum of Sharing Research Data ». *Journal of the American Society for Information Science and Technology*. 2012. Vol. 63, n°6, p. 1059 1078.

Borgman C. L. *Big data, little data, no data*. Cambridge, Massachusetts : The MIT Press, 2015. p. 383

Boukacem-Zeghmouri C. « Nouveaux intermédiaires de l'information, nouvelles logiques de captation de la valeur ». *I2D–Information, données & documents* [en ligne]. 2015. Vol. 53, n°4, p. 34–35. Disponible sur : < http://www.cairn.info/resume.php?ID_ARTICLE=I2D_154_0034 > (consulté le 01 septembre 2016)

Boukacem-Zeghmouri C. *Mutations dans la sous-filière de la revue scientifique dans les domaines STM : une analyse par les industries culturelles* [en ligne]. Accreditation to supervise research : Université Claude Bernard Lyon 1, 2015. Disponible sur : < http://archivesic.ccsd.cnrs.fr/tel-01281524 > (consulté le 01 septembre 2016)

Brown S., Bruce R., Kernohan D., Directions for Research Data Management in UK Universities, JISC [en ligne] mars 2015. Disponible sur : <

http://repository.jisc.ac.uk/5951/4/JR0034_RDM_report_200315_v5.pdf > (consulté le 01 juin 2016)

Candela L., Castelli D., Manghi P., Tani A. « Data journals: A survey ». *Journal of the Association for Information Science and Technology* [en ligne]. septembre 2015. Vol. 66, n°9, p. 1747 1762. Disponible sur : < http://dx.doi.org/10.1002/asi.23358 > (consulté le 9 mai 2016)

Cabrera F. « Les données de la recherche en Sciences humaines et sociales: enjeux et pratiques ». Disponible sur : < http://hal-obspm.ccsd.cnrs.fr/mem_01128394/document > (consulté le 5 avril 2016)

Chao T. C. « Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences ». *Proc. Am. Soc. Info. Sci. Tech[en ligne]*. 1 janvier 2011. Vol. 48, n°1, p. 1 8. Disponible sur : < http://dx.doi.org/10.1002/meet.2011.14504801125 > (consulté le 17 mai 2016)

Chartron G. « Stratégie, politique et reformulation de l'open access », *Revue française des sciences de l'information et de la communication* [en ligne], 2016. Disponible sur : < http://rfsic.revues.org/1836 > (consulté le 01 septembre 2016)

Chavan V., Penev L. « The data paper: a mechanism to incentivize data publishing in biodiversity science ». *BMC Bioinformatics* [en ligne]. 2011, Vol. 12, n°15, p. 1 12. Disponible sur : < http://dx.doi.org/10.1186/1471-2105-12-S15-S2 > (consulté le 15 juillet 2016)

Cite C.-I. T. G. on D. C. S. and P. Of, Sices O. of M. T. C. « Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data ». *Data Science Journal* [en ligne]. 8 septembre 2013. Vol. 12, n°0. Disponible sur : < http://dx.doi.org/10.2481/dsj.OSOM13-043 > (consulté le 17 mai 2016)

CODATA-ICSTI Task Group on Data Citation Standards and Practices. « Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data ». Data Science Journal [en ligne]. 8 septembre 2013. Vol. 12, n°0. Disponible sur : < http://dx.doi.org/10.2481/dsj.OSOM13-043 > (consulté le 22 juin 2016)

Concordat on Open Research Data launched [en ligne]. Research Councils UK, 28 juillet 2016. Disponible sur : < http://www.rcuk.ac.uk/media/news/160728/ >

Convention de partenariat en faveur des archives ouvertes et de la plateforme mutualisée HAL[en ligne] 02 avril 2013. Disponible sur : < http://cache.media.enseignementsup-recherche.gouv.fr/file/HAL/93/3/01_Convention_HAL_246933.pdf >

Cordero-Llana L., Ramage K., Law K. S., Keckhut P. « LABEX L-IPSL Arctic Metadata Portal ». *Data Science Journal* [en ligne]. 12 janvier 2016. Vol. 15, n°0,. Disponible sur : < http://dx.doi.org/10.5334/dsj-2016-002 > (consulté le 30 août 2016)

Cox A., Verbaan E., Sen B. « Upskilling Liaison Librarians for Research Data Management ». Ariadne [en ligne]. 2012. n°70. Disponible sur : < http://www.ariadne.ac.uk/issue70/cox-et-al > (consulté le 20 avril 2016)

Defeo C., Introducing Elsevier DataSearch [en ligne]. Mendeley blog, 24 août 2016. Disponible sur : < https://blog.mendeley.com/2016/08/24/introducing-elsevier-datasearch/ > (consulté le 25 août 2016)

Délémontez R., Boukacem-Zeghmouri C. « Données de la recherche: entre discours, réalités et valeur ». *I2D–Information, données & documents* [en ligne]. 2015. Vol. 53, n°4, p. 56–57. Disponible sur : < http://www.cairn.info/resume.php?ID_ARTICLE=I2D_154_0056 > (consulté le 01 septembre 2016)

Direction de l'Information Scientifique et Technique - CNRS. Livre blanc — Une Science ouverte dans une République numérique [en ligne]. Marseille : OpenEdition Press, 2016. (Laboratoire d'idées). Disponible sur : < http://books.openedition.org/oep/1548 > (consulté le 15 juillet 2016)

Drachen T., Ellegaard O., Larsen A., Dorch S. « Sharing data increases citations ». *LIBER Quarterly* [en ligne]. 15 août 2016. Vol. 26, n°2. Disponible sur : < http://dx.doi.org/10.18352/lq.10149 > (consulté le 30 août 2016)

Edinburgh University Data Library Research Data Management Handbook, Edinburgh University Library[en ligne]. 2011. Disponible sur : http://www.docs.is.ed.ac.uk/docs/data-library/EUDL_RDM_Handbook.pdf

Fecher B., Friesike S. « Open Science: One Term, Five Schools of Thought ». In : Bartling S., Friesike S., *Opening Science*, [en ligne]. Springer International Publishing. 2014. Disponible sur : < http://book.openingscience.org/basics_background/open_science_one_term_five_schools_of_thought.html > (consulté le 23 juin 2016)

FORCE 11, « Joint Declaration of Data Citation Principles - FINAL ». FORCE11 [en ligne]. 2013. Disponible sur : < https://www.force11.org/group/joint-declaration-data-citation-principles-final > (consulté le 05 mai 2016)

Gaillard R. « De l'Open data à l'Open research data: quelle (s) politique (s) pour les données de recherche? ». 2014. p. 238. Disponible sur : < http://eprints.rclis.org/22746/ > (consulté le 5 avril 2016)

GilPress. *Data Scientists Spend Most of Their Time Cleaning Data* [en ligne]. *What's The Big Data?*. 1 mai 2016. Disponible sur : < https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/ > (consulté le 7 juin 2016)

Gitelman L. (éd.). *« Raw data » is an oxymoron*. Cambridge, Mass., Etats-Unis d'Amérique : MIT Press, 2013

Gómez N.-D., Méndez E., Hernández-Pérez T. « Data and metadata research in the social sciences and humanities: An approach from data repositories in these disciplines ». *El Profesional de la Información* [en ligne]. 29 juillet 2016. Vol. 25, n°4, p. 545. Disponible sur : < http://dx.doi.org/10.3145/epi.2016.jul.04 > (consulté le 16 août 2016)

Goodman S. N., Fanelli D., Ioannidis J. P. A., What does research reproducibility mean?, *Science Translational Medicine* [en ligne]. 01 juin 2016, Vol. 8. Disponible sur : < http://stm.sciencemag.org/content/8/341/341ps12.full > (consulté le 16 août 2016)

He L., Nahar V., Lewandowski D. « Reuse of scientific data in academic publications: an investigation of Dryad Digital Repository ». *Aslib Journal of Information Management* [en ligne]. 2016. Vol. 68, n°4. Disponible sur : < http://www.emeraldinsight.com/doi/abs/10.1108/AJIM-01-2016-0008 > (consulté le 7 juin 2016)

Hrynaszkiewicz I. « Promoting research data sharing at Springer Nature ». BioMed Central blog [en ligne]. 2016. Disponible sur : < http://blogs.biomedcentral.com/bmcblog/2016/07/05/promoting-research-data-sharing-springer-nature/ > (consulté le 12 juillet 2016)

Huwe T. K., « Your New Role as a Data Scientist », *Computers in Libraries*. Avril 2016, p. 23

Jost C., « Ghislaine Chartron : "Je ne transformerai pas mes étudiants en data scientists" ». *Archimag* [en ligne]. novembre 2015. N°289. Disponible sur : < http://www.archimag.com/veille-documentation/2015/11/26/ghislaine-chartron-transformer-etudiants-data-scientists > (consulté le 31 août 2016)

Kenall A., Harold S., Foote C. « An open future for ecological and evolutionary data? ». *BMC ecology* [en ligne] 2014. Vol. 14, n°1, p. 10. Disponible sur : < http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-14-66 > (consulté le 31 août 2016)

Kratz J., Strasser C. « Data publication consensus and controversies ». *F1000Research*[en ligne] 2014. Disponible sur : < http://f1000researchdata.s3.amazonaws.com/manuscripts/5878/397d927b-269a-4f45-826d-27fef7106eaf_3979_-_john_kratz_v3.pdf > (consulté le 7 juin 2016)

Lin J., « The article nexus: linking publications to associated research outputs », The art of persistence [en ligne]. Crossref. 25 août 2016. Disponible sur : < http://blog.crossref.org/2016/08/the-article-nexus.html > (consulté le 27 août 2016)

Monino J.-L., Sedkaoui S. *Big Data, Open Data et valorisation des données*. London, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : ISTE Éditions, 2016

NWO stimulates optimal access to research data, NWO[en ligne] 20 juillet 2016. Disponible sur : < http://www.nwo.nl/en/news-and-events/news/2016/nwo-introduces-data-management-protocol-in-all-calls.html > (consulté le 20 août 2016)

Piwowar H. A., Day R. S., Fridsma D. B. « Sharing Detailed Research Data Is Associated with Increased Citation Rate ». *PLOS ONE* [en ligne]. 21 mars 2007. Vol. 2, n°3, p. e308. Disponible sur : < http://dx.doi.org/10.1371/journal.pone.0000308 > (consulté le 28 juillet 2016)

Piwowar H. A., Vision T. J. « Data reuse and the open data citation advantage ». *PeerJ* [en ligne]. 1 octobre 2013. Vol. 1, p. e175. Disponible sur : < http://dx.doi.org/10.7717/peerj.175 > (consulté le 22 juin 2016)

Piron F., « Éthique, développement durable et libre accès ». *I2D–Information, données & documents* [en ligne]. 2016. Vol. 53, n°1, p. 42-43. Disponible sur : < http://www.cairn.info/resume.php?ID_ARTICLE=I2D_154_0034 > (consulté le 01 septembre 2016)

Rebouillat V. « Archives Ouvertes de la Connaissance. Valoriser et diffuser les données de recherche ». 2015. p. 84. Disponible sur : < www.enssib.fr/bibliotheque-numerique/notices/66039-archives-ouvertes-de-la-connaissance-valoriser-et-diffuser-les-donnees-de-recherche > (consulté le 5 avril 2016)

Prost H., Schöpfel J. *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3* [en ligne]. Université de Lille 3, 2015. Disponible sur : < http://hal.univ-lille3.fr/hal-01198379/document > (consulté le 3 juin 2016)

Schöpfel J., Prost H., Rebouillat V., « Research Data in Current Research Information Systems » [en ligne]. 2016. Disponible sur : < http://dspacecris.eurocris.org/handle/11366/501 > (consulté le 20 juin 2016)

Seadle M. S., « Managing and mining historical research data ». *Library Hi Tech* [en ligne]. 21 mars 2016. Vol. 34, n°1, p. 172 179. Disponible sur : < http://dx.doi.org/10.1108/LHT-09-2015-0086 > (consulté le 16 août 2016)

Tene O., Polonetsky J., « Big Data for All: Privacy and User Control in theAge of Analytics ». Northwestern Journal of Technology and Intellectual Property [en ligne]. Vol. 11, n°5. Disponible sur : < http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1191&context=njtip > (consulté le 16 août 2016)

Thessen A. E., Patterson D. J., « Data issues in the life sciences ». *Zookeys* [en ligne]. 28 novembre 2011. n°150, p. 15-51. Disponible sur : < http://dx.doi.org/10.3897/zookeys.150.1766 > (consulté le 06 juin 2016)

Wallis J. C., Rolando E., Borgman C. L., « If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology ». *PLOS ONE* [en ligne]. juil 2013. Vol. 8, n°7. Disponible sur : < http://dx.doi.org/10.1371/journal.pone.0067332 > (consulté le 22 juin 2016)

Zaken M. van B., « Amsterdam Call for Action on Open Science - Publicatie - EU2016.nl » [en ligne]. 2016. Disponible sur : < http://www.eu2016.nl/documenten/rapporten/2016/04/04/amsterdam-call-for-action-on-open-science > (consulté le 7 juin 2016)

ANNEXES

Table des annexes

LES CINQ ECOLES DE PENSEE DE L'*OPEN SCIENCE*

Table 1. *Five Open Science schools of thought.*

School of thought	Central assumption	Involved groups	Central Aim	Tools & Methods
Democratic	The access to knowledge is unequally distributed.	Scientists, polititians, citizens	Making knowledge freely available for everyone.	Open access, intellectual property rights, Open data, Open code
Pragmatic	Knowledge-creation could be more efficient if scientists worked together.	Scientists	Opening up the process of knowledge creation.	Wisdom of the crowds, network effects, Open Data, Open Code
Infrastructure	Efficient research depends on the available tools and applications.	Scientists & platform providers	Creating openly available platforms, tools and services for scientists.	Collaboration platforms and tools
Public	Science needs to be made accessible to the public.	Scientists & citizens	Making science accessible for citizens.	Citizen Science, Science PR, Science Blogging
Measurement	Scientific contributions today need alternative impact measurements.	Scientists & politicians	Developing an alternative metric system for scientific impact.	Altmetrics, peer review, citation, impact factors

LES ENTREPOTS IMPOSES PAR *NATURE*

Mandates for specific datasets

For the following types of data set, submission to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper. Examples of appropriate public repositories are listed below.

Mandatory deposition	**Suitable repositories**
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Microarray data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database

EXEMPLES DE REQUETES UTILISEES POUR L'ETUDE EXPLORATOIRE

Extraction des domaines des dépôts avec annexes :

http://api.archives-ouvertes.fr/search/?q=fileType_s:annex&rows=0&wt=xml&indent=true&facet=true&facet.field=level0_domain_s&facet.limit=250

Extraction des types document de tous les dépôts avec annexe :

```
SELECT C.TYPDOC, count(*) FROM (SELECT IDENTIFIANT, TYPDOC FROM DOC_FILE JOIN DOCUMENT ON DOCUMENT.DOCID = DOC_FILE.DOCID WHERE DOC_FILE.FILETYPE = 'annex' GROUP BY IDENTIFIANT ) AS C WHERE 1 group by C.TYPDOC
```

Extraction des différentes extensions des fichiers identifiés comme annexes :

```
SELECT `EXTENSION`,COUNT(*) FROM `DOC_FILE` WHERE 1 AND `FILETYPE` LIKE 'annex' GROUP BY `EXTENSION`
```

LES TYPES DE DOCUMENT PRINCIPAUX DE HAL

Catégorie	TYPDOC	Nom interface
Publications	ART	Article dans une revue
	COMM	Communication dans un congrès
	POSTER	Poster
	OUV	Ouvrage (y compris édition critique et traduction)
	COUV	Chapitre d'ouvrage
	DOUV	Direction d'ouvrage, Proceedings
	PATENT	Brevet
	OTHER	Autre publication
Documents non publiés	UNDEFINED	Pré-publication, Document de travail
	REPORT	Rapport
Travaux universitaires	THESE	Thèse
	HDR	HDR
	LECTURE	Cours
Données de recherche	IMG	Image
	VIDEO	Vidéo
	SON	Son
	MAP	Carte

LES FORMATS DE FICHIER ACCEPTES DANS HAL

tex, eps_tex, ps_tex, pstex, pdf_tex, pdf_t, pdftex, zip, odc, ods, pages, cls, clo, cnf, sty, bst, bib, bbl, toc, idx, aux, def, loc, table, pdf, doc, docx, txt, dot, dotx, rtf, odf, odt, ott, html, htm, ppt, pptx, pot, potx, pps, ppsx, pptm, ppsm, ps, eps, odp, ots, key, knt, xls, xlsx, xlsm, xltx, xlt, xml, xsl, jpg, jpeg, jpe, jps, png, gif, tif, tiff, ms3d, odg, otg, pct, svg, aac, ac3, aif, aifc, aiff, au, bwf, mp2, mp3, M4r, ogg, ogm, ra, ram, wma, wav, avi, flv, mov, movie, mp4, mpe, mpeg, mpg, qt, rm, rmvb, rv, vob, wmv, m4a

GLOSSAIRE

Archivage pérenne : conservation sur le long terme de données sélectionnées en s'assurant qu'un fichier est toujours présent sur le support de stockage et qu'il conserve son intégrité. L'indexation doit permettre de les retrouver facilement (ouverture et lecture du fichier) et les données doivent rester intelligibles (compréhensibles par les utilisateurs potentiels à travers le temps).

Archive ouverte : le terme archive ouverte désigne un réservoir où sont déposées des données issues de la recherche scientifique et de l'enseignement et dont l'accès se veut ouvert c'est-à-dire sans barrière. Cette ouverture est rendue possible par l'utilisation de protocoles communs qui facilitent l'accessibilité de contenus provenant de plusieurs entrepôts maintenus par différents fournisseurs de données.

Auto-archivage : l'auto-archivage est l'acte par lequel les chercheurs déposent eux-mêmes leurs articles (prépublications et postpublications) dans des archives ouvertes.

Big data : Ensemble des données produites en temps réel et en continu, structurées ou non, et dont la croissance est exponentielle.

Data paper : publication qui décrit un jeu de données scientifiques brutes, notamment à l'aide d'informations précises, appelées métadonnées

DOI : identifiant unique d'un objet numérique. Son lien URL pointe sur le site éditeur. On retrouve ce lien à l'aide d'un résolveur de DOI comme http://dx.doi.org/. Le but des DOI est de faciliter la gestion numérique sur le long terme de toute chose en associant des métadonnées à l'identifiant de l'objet : un article, une base de données…

Identifiant pérenne : identifiant qui est assigné à un objet de façon permanente. Il est disponible et gérable à long terme ; il ne changera pas si l'objet est renommé ou déplacé (changement de site, d'entrepôts de données...).

Intéropérabilité : l'interopérabilité est la capacité de différents systèmes informatiques à dialoguer entre eux, à communiquer sans ambiguïté et ainsi interpréter des informations correctement.

***Open access* / libre accès** : par « accès libre » à cette littérature, nous entendons sa mise à disposition gratuite sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et l'utilisation d'Internet. La seule contrainte sur la reproduction et la distribution, et le seul rôle du copyright dans ce domaine devrait être de garantir aux auteurs un contrôle sur l'intégrité de leurs travaux et le droit à être correctement reconnus et cités.

Open data : L'ouverture des données (*Open data*) a pour objectif la diffusion libre, gratuite et universelle, via internet, des données d'origine publique ou privée. Le terme ouvert est défini comme la liberté d'utiliser, de modifier et de redistribuer librement les données. L'Open data s'inscrit dans le mouvement mondial du libre accès à la connaissance (*Open knowledge*) et plus largement de la

Science ouverte (*Open science*), qui considère la science comme un bien commun dont la diffusion est d'intérêt public et général.

Protocole OAI-PMH : le protocole OAI-PMH est un protocole qui définit les conditions du transfert de métadonnées d'une archive ouverte, produite par un fournisseur de données, vers le serveur d'un fournisseur de services.

Revue par les pairs / *Peer-reviewing* : La révision, ou relecture, ou évaluation, des articles par les pairs (*peer review*) est l'étape incontournable avant la publication. Elle permet de vérifier le contenu scientifique de l'article et son apport original par rapport à ce qui a été publié dans le domaine concerné.

***Supplementary materials* / Données supplémentaires** : données additionnelles à une publication contenant des informations aidant à sa compréhension.

TDM : La fouille de textes et de données (*text and data mining* - TDM) consiste à explorer, via un algorithme de fouille, des corpus immenses - composés notamment d'articles scientifiques ou de données expérimentales - afin d'en déduire des nouvelles connaissances.

Voie dorée : la voie dorée s'applique à la publication d'articles dans des revues en libre accès, quel que soit leur mode de financement. Elle correspond à la deuxième stratégie recommandée dans l'Initiative de Budapest pour l'Accès Ouvert : « Revues alternatives : en second lieu, les savants ont besoin des moyens pour lancer une nouvelle génération de revues alternatives engagées dans le libre accès et pour aider les revues existantes qui choisissent d'opérer la transition vers l'accès libre. »

Voie verte : la voie verte qualifie l'auto-archivage par les chercheurs ou l'archivage par une tierce personne des articles dans des archives ouvertes. Elle correspond à la première stratégie préconisée dans l'Initiative de Budapest pour l'Accès Ouvert : « Auto-archivage : en premier lieu, les savants ont besoin d'outils et d'assistance pour déposer leurs articles de revues à comité de lecture dans des archives électroniques ouvertes, une pratique communément appelée auto-archivage. »

TABLE DES ILLUSTRATIONS

TABLE DES MATIERES