ÉTUDES DE COMMUNICATION
langages, information, médiations

Études de communication
langages, information, médiations

52 | 2019
Libre accès et données de la recherche. Quelle résonance dans les SIC ?

Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique

Data Repositories in Information and Communication Sciences. An Empirical Study

Hélène Prost et Joachim Schöpfel

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/edc/8604
DOI : 10.4000/edc.8604
ISSN : 2101-0366

Éditeur
Université de Lille

Édition imprimée
Date de publication : 1 juin 2019
Pagination : 71-98
ISBN : 978-2-917562-21-5
ISSN : 1270-6841

Référence électronique
Hélène Prost et Joachim Schöpfel, « Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique », *Études de communication* [En ligne], 52 | 2019, mis en ligne le 01 janvier 2021, consulté le 21 janvier 2021. URL : http://journals.openedition.org/edc/8604 ; DOI : https://doi.org/10.4000/edc.8604

© Tous droits réservés

View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by OpenEdition

Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique

Data Repositories in Information and Communication Sciences. An Empirical Study

Hélène Prost

CNRS, EA 4073 – GERiiCO, F-59000 Lille, France
helene.prost007@gmail.com

Joachim Schöpfel

Univ. Lille, EA 4073 – GERiiCO, F-59000 Lille, France
joachim.schopfel@univ-lille.fr

Pour alimenter le débat au sein de la communauté des sciences de l'information de la communication et accompagner l'émergence de la science ouverte, l'article présente les résultats d'une étude empirique sur les dispositifs numériques mis en place par et pour les chercheurs en SIC dans le domaine de la gestion des données de recherche. Quels sont les entrepôts thématiques et disciplinaires des SIC, et quels sont les services génériques d'accueil et de diffusion des données de recherche à disposition des SIC ? Après une analyse conceptuelle des données de recherche et des entrepôts, l'article présente les résultats d'une étude empirique à partir des répertoires Cat-OPIDoR et re3data, en particulier leurs contenus, métadonnées, dimensions disciplinaires et localisation géographique. La discussion porte sur trois aspects : la certification et la normalisation ; la question de la gestion, de l'archivage et/ou la diffusion des données ; et l'impact de la gestion des données sur la communauté des SIC. En guise de conclusion, l'article propose quelques recommandations pour le positionnement des SIC et quelques pistes pour des recherches futures.

Mots-clés : données de recherche, entrepôts de données, science ouverte, libre accès, sciences de l'information et de la communication.

This article offers a contribution to the ongoing debate within the Information and Communication Sciences community concerning the emergence of Open science. We present the results of an empirical study on digital devices implemented by and for researchers in Information and Communication Sciences in the field of research data management. What are the thematic and disciplinary data repositories in this discipline, and what are the generic repositories for the preservation and dissemination of research data? After a conceptual analysis of research data and warehouses, the article presents the results of an empirical study based on the Cat-OPIDoR and re3data directories, with information about their contents, metadata, disciplinary dimensions and geographic location. The discussion focuses on three aspects: certification and standardization; the issue of management, archiving and/or dissemination of data; and the impact of data management on the disciplinary community. By way of conclusion, the article offers some recommendations for the positioning of Information and Communication Sciences and for future research.

Keywords: research data, data repositories, open science, open access, information and communication sciences.

1.
Un triple enjeu pour les SIC

Dans la continuité des chantiers engagés pour la transformation numérique de l'État, le gouvernement français met en place une action publique plus transparente et plus collaborative. Le *Plan d'action 2018-2020* (Etalab, 2018) d'avril 2018 énumère différents engagements, parmi lesquels figure la construction d'un écosystème de la science ouverte, porté par le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI). Outre l'abandon du système des abonnements et le développement du *Text & Data Mining* (TDM), le *Plan d'action* préconise à partir de 2019 la « généralisation progressive, via un accompagnement, de la mise en place de plans de gestion des données dans les appels à projets de recherche, et l'incitation à une ouverture des données produites par les programmes financés » (Etalab, 2018).

La science ouverte figure désormais parmi les priorités de l'État français qui soutient « la mise en œuvre des principes du gouvernement ouvert pour renforcer [...] l'accès aux matériaux et résultats de la recherche ». Dans cet écosystème émergent de la science ouverte, la science sera « plus cumulative, plus fortement étayée par des données, plus transparente, plus intègre, plus rapide et d'accès plus universel (et qui) induit une démocratisation de l'accès aux savoirs, utile à la recherche, à la formation, à la société » (Etalab, 2018). Le *Plan d'action* ne définit pas clairement la science ouverte (*open science*) mais ce terme désigne généralement plusieurs dimensions du travail scientifique, dont l'accès libre aux publications, le partage des données de recherche, une évaluation plus transparente (y compris pour le *peer review*) et des ressources éducatives libres.

Le *Plan national pour la science ouverte* du MESRI présenté en juillet 2018 traduit cet engagement en 9 mesures et 27 actions, réparties sur trois axes prioritaires, dont la structuration et l'ouverture des données de recherche[1] (*cf.* figure 1). En tête des mesures figure l'obligation de la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics. Par rapport aux dispositifs et infrastructures, le *Plan* prévoit trois actions :

- développer des centres de données thématiques et disciplinaires ;
- développer un service générique d'accueil et de diffusion des données simples ;
- engager un processus de certification des infrastructures de données.

1 MESRI (2018). *Plan national pour la science ouverte*. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Paris. Disponible sur http://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html (page consultée le 27 février 2019).

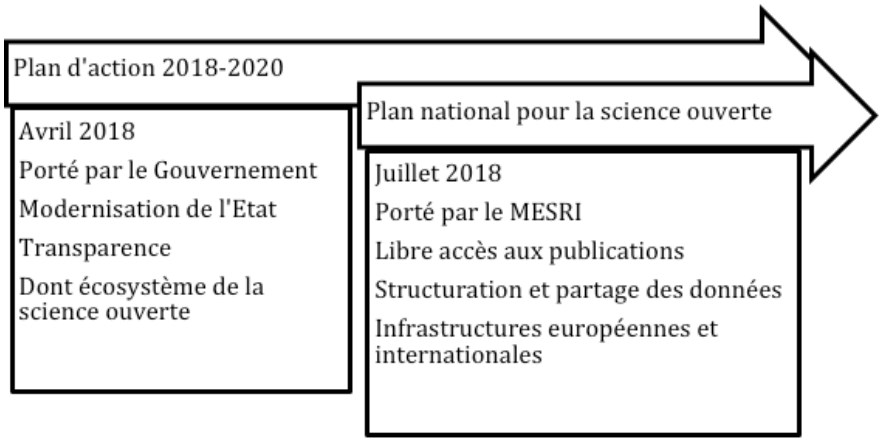

Figure 1 : Les deux plans français en faveur de la science ouverte

La situation évolue rapidement. Fin 2018, le Ministère a mis en place un Comité pour la Science Ouverte (CoSO)[2] pour coordonner la mise en œuvre de cette politique d'ouverture. Début décembre, le nouveau comité a organisé la première journée annuelle du groupe français de la *Research Data Alliance*[3], et son Collège Données a fait des recommandations concernant la gestion des données dans les projets de l'Agence Nationale de la Recherche ; le Collège travaille aussi sur l'accompagnement des politiques institutionnelles, sur les pratiques et la gouvernance des données et sur le recensement des services de données. Le groupe s'intéresse également à la question des entrepôts et pilote la fédération des usagers français de *Dataverse*[4], un logiciel *open source* développé à l'Université de Harvard pour le partage, la citation, l'analyse et la préservation de données de recherche.

Pour les Sciences de l'Information et de la Communication (SIC), cette nouvelle politique d'ouverture des données de recherche présente un triple enjeu :

1. un enjeu conceptuel : malgré un nombre toujours plus grands d'articles et de communications sur les données de recherche, leur définition précise comme ressources et résultats de la recherche reste incertaine. De même, les nouveaux services de données et notamment les entrepôts échappent à ce jour globalement à l'analyse des dispositifs d'information, aussi bien par rapport à leurs fonctions que par rapport aux besoins, usages et compétences en matière de données de recherche. Quel forme de « *data literacy* » pour ces dispositifs ? Quel impact sur les pratiques des professionnels et chercheurs ?

2 https://forum.ouvrirlascience.fr/.

3 https://www.rd-alliance.org/groups/rda-france.

4 Dont l'INRA, le CIRAD et Sciences Po Paris.

2. un enjeu méthodologique : pour l'accompagnement scientifique de la mise en œuvre de son plan d'ouverture, le Ministère soutient l'évaluation des différentes initiatives et l'élaboration d'indicateurs pour le suivi (*monitoring*) à plus long terme. Avec leurs compétences et approches qualitatives et quantitatives, notamment dans le domaine de la scientométrie, les SIC sont attendues dans ce domaine ;

3. un enjeu communautaire : une communauté scientifique se définit par ses valeurs, concepts, pratiques et outils. Comment les SIC se positionnent-elles par rapport aux données de recherche ? Quelles sont leurs pratiques en matière de gestion ? De quels outils disposent-elles ? Quelle est leur stratégie par rapport à la politique de l'ouverture et du partage des données de recherche ?

Afin d'alimenter le débat au sein de la communauté des SIC et d'accompagner l'émergence de la science ouverte, cet article présente les résultats d'une étude empirique sur les dispositifs numériques mis en place par et pour les chercheurs en SIC dans le domaine de la gestion des données de recherche. Pour utiliser les termes du Ministère, quels sont les « entrepôts thématiques et disciplinaires » des SIC, et quels sont les « services génériques d'accueil et de diffusion » des « données simples » à disposition des SIC ? Nous proposerons d'abord une approche scientifique des données de recherche et des entrepôts. Ensuite, après la présentation des résultats de notre étude, nous discuterons notamment trois aspects : la certification et la normalisation ; la question de la gestion, de l'archivage et/ou la diffusion des données ; et l'impact de la gestion des données sur la communauté des SIC, avec quelques pistes pour des recherches futures.

2.
Les données de recherche comme objet scientifique

Des résultats d'enquêtes, des transcriptions d'entretiens, des corpus de texte, des analyses statistiques, des images ou des photographies, des graphiques ou des tableaux : les données de recherche se caractérisent par leur diversité, leur volume et leur production et mise à jour continue. Que ce soient des données sources (primaires) ou des données élaborées (secondaires), les données de recherche « sont utilisées comme sources principales pour la recherche et sont reconnues par la communauté scientifique comme nécessaires à la validation des résultats de la recherche » (OCDE, 2007). Mais leur définition pose problème, et il paraît difficile de s'accorder sur une seule définition, sauf à constater à un niveau très général, à l'instar de Chignard (2012), que « tout est donnée », « un fait brut, qui n'est pas – encore – interprété », un matériel que l'on peut (doit) manipuler, traiter, analyser et interpréter. Le concept de la « valeur potentielle » se retrouve chez Borgman (2015) qui définit les don-

nées de recherche comme « *inputs, outputs, and assets of scholarship* ». Les données comme « capital » ou « biens » de la recherche, à exploiter, analyser, travailler pour en révéler ou extraire une valeur scientifique sous forme d'une information ou d'un savoir.

Nous avons identifié ailleurs (Schöpfel *et al*., 2017) quatre dimensions du concept de la donnée de recherche, à savoir :

- l'enregistrement (fixation sur support) ;
- la nature factuelle (réelle ou supposée) ;
- le lien avec la communauté scientifique ;
- la finalité (fonction).

Notre approche multidimensionnelle établit un lien entre données et chercheurs, au sens d'un minimum de pratiques, valeurs, méthodes, outils et concepts partagés ; une donnée est ce qui est acceptée par un groupe de chercheur (« communauté ») comme telle (consensus). Le lien avec la communauté peut être fondé par un cadre conceptuel, une thématique ou discipline ; souvent il s'appuiera sur un instrument, une procédure ou une méthodologie. Cette relation peut-être assez complexe dans un environnement multi-, inter- ou transdisciplinaire ; ce qui constitue des données est déterminé par une communauté d'intérêts qui les produit. Cependant, un enquêteur peut faire partie de multiples communautés d'intérêts qui se chevauchent, chacune pouvant avoir une notion différente de ce que sont des données (Koltay, 2016).

Un autre trait caractéristique est la dimension fonctionnelle, le rôle des données dans le processus de recherche, en particulier pour la validation des hypothèses et résultats. Ce lien entre « un ensemble d'informations factuelles enregistrées sur des supports » et leur production ou collecte « selon divers procédés au cours d'un processus de recherche » (Reymonet, 2017) a été conceptualisé sous forme d'un « cycle de vie » de la donnée de recherche (Pain, 2016), en soulignant la nature dynamique de cet objet « complexe, dynamique, vivant » dont les caractéristiques évoluent « selon l'étape du processus de recherche auquel on s'intéresse » (André, 2015). La relation avec le contexte du travail scientifique est essentielle pour définir les données de recherche. Cette relativité conceptuelle rend l'évaluation des données difficile ; en fait, les systèmes d'information recherche n'évaluent que la gestion des données (leur description, préservation, diffusion, etc.), mais pas les données elles-mêmes, ni leur volume, ni leur qualité, ni leur pertinence ou valeur (Schöpfel *et al*., 2016).

À ceci s'ajoute une très grande diversité qui a fait l'objet de nombreuses typologies, hiérarchies et modèles. À partir d'une analyse des finalités et procédures de la génération des données de recherche, le *Research Information Network* (RIN, 2008) a établi cinq larges catégories transversales :

- données d'observation ;
- données d'expérimentation ;

- données de simulation ;
- données dérivées ;
- données de référence.

Pour des besoins d'acquisition et de collection, Matthews *et al*. (2002) ont modélisé les données sous forme d'une arborescence dynamique organisée à partir de métadonnées, d'identifiants et d'adresses ; les différents fichiers d'un jeu de données sont liés par des métadonnées mais ont des adresses différentes et peuvent être localisés sur différents serveurs. Le mode de production et la complexité des données sont à l'origine du modèle de données du *European Plate Observing System*[5] qui distingue quatre niveaux (Bailo et Jeffery, 2014) :

- données brutes, ou données de base (ex. : sismogramme) ;
- données issues de procédures quasi-automatisées (ex. : localisation des tremblements de terre) ;
- données issues des recherches scientifiques (ex. : modèles de la croûte terrestre) ;
- données intégrées issues d'analyses complexes ou partagées (ex. : cartes de risque).

Ici, le lien du modèle avec les instruments et pratiques d'une communauté de recherche est évident, ce lien a un impact fort sur l'organisation et la gestion des données : comment décrire et identifier ces différents niveaux, comment stocker et diffuser les jeux de données ? D'une certaine manière, le plan du Ministère s'appuie sur ce genre de modèle, dans la mesure où il fait la distinction entre données « simples » dans des entrepôts génériques et données plus complexes dans des services disciplinaires ou thématiques.

À partir de l'indexation d'un grand nombre d'entrepôts de données, le répertoire international *re3data*[6] différencie quatorze types de données, selon une répartition inégale entre plusieurs larges catégories, transversales aux disciplines et aux contours mal définis ; le répertoire liste également une longue traîne d'autres types de données, dont certains ont un profil nettement plus disciplinaire (Kindling *et al*., 2017). Cette dernière approche est particulièrement intéressante pour notre propre étude ; là où les autres analyses essayent de conceptualiser le lien des données avec leur contexte et leur finalité, Kindling *et al*. étudient les données « froides » (définitives, finales, statiques) à partir de leur lieux de stockage et d'archivage, c'est-à-dire comme éléments des entrepôts. À première vue, cette approche parait décontextualiser les données de recherche, en coupant le lien avec leurs fonctions initiales et les pratiques des chercheurs. En réalité, il n'en est rien : l'analyse du concept des entrepôts comme dispositifs numériques va le démontrer.

5 EPOS https://www.epos-ip.org/.

6 https://www.re3data.org/.

3.
Les entrepôts comme dispositifs numériques

Les entrepôts de données occupent aujourd'hui une place centrale dans les politiques nationales et institutionnelles et dans le développement des infrastructures de données. Décrits comme des services en ligne « permettant la collecte, la description, la conservation, la recherche et la diffusion des jeux de données »[7], les entrepôts de données (*data repositories*) facilitent le stockage et le partage des données. Ils leur apportent également de la valeur ajoutée, par une description normalisée (métadonnées), par des identifiants (DOI), par des licences de diffusion et une authentification de l'accès, par une arborescence, des labels qualité (certification) ou des liens avec des publications associées.

Les annuaires et répertoires distinguent généralement plusieurs catégories d'entrepôts, notamment par rapport à leurs contenus mais aussi, en fonction de leur gouvernance et périmètre. Ainsi, il y a notamment des entrepôts :

- disciplinaires : il s'agit d'entrepôts réservés aux données issues d'une discipline ou domaine particulier (économie, chimie, histoire...), avec des thématiques plus ou moins larges ;
- génériques : il s'agit d'entrepôts pour accueillir et diffuser des données d'une large gamme de disciplines, avec des conditions techniques (taille et format des fichiers, etc.) mais sans restrictions quant au domaine ;
- institutionnels : il s'agit d'entrepôts destinés aux données produites par des chercheurs d'une institution (université, laboratoire, institut, organisme de recherche...).

Le *Plan* national introduit une autre distinction, entre « données simples » dans un service de données générique d'accueil et de diffusion, et données plus complexes (3D, etc.) dans des entrepôts plus spécifiques, limités à un champ de recherche, à un équipement ou une méthodologie.

Face à cette diversité, le répertoire international *re3data* a listé quatre éléments constitutifs d'un entrepôt de données (cf. figure 2), dont la fiabilité (*trustworthiness*), la qualité d'un tel dispositif pour garantir sa pérennité (*sustainability*) et son environnement scientifique, avec les chercheurs, agences de financement, éditeurs et établissements de recherche[8]. La garantie d'un

7 Cocaud S., Aventurier P. (2017). Les entrepôts de données de la recherche. In *Participer à l'organisation du management des données de la recherche, gestion de contenu et documentation des données*. Disponible sur https://anfdonnees2017.sciencesconf.org/data/pages/Entrepots_ANFRenatis_2017_Cocaud_Aventurier_1.pdf (page consultée le 27 février 2019).

8 *Re3data Metadata Schema for the Description of Research Data Repositories : version 3.0* http://gfzpublic.gfz-potsdam.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:1397899.

accès à long terme (pérenne) implique que les conditions d'accès soient clairement explicitées.

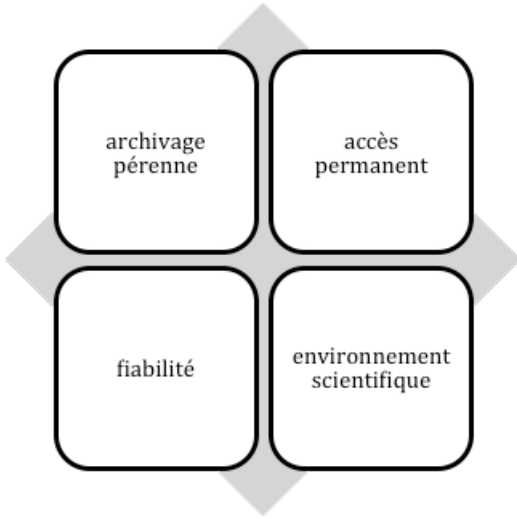

Figure 2 : Quatre éléments constitutifs d'un entrepôt de données (source : *re3data*)

Autour de ces quatre éléments constitutifs des entrepôts de données, d'autres fonctionnalités sont caractéristiques pour ces dispositifs, dont notamment le dépôt des données en ligne (y compris l'import en masse), l'identification pérenne des données avec des identifiants normalisés (DOI, handle…), une description des données suivant des formats normalisés de métadonnées (génériques ou spécifiques), des outils de découverte et de recherche, et le contrôle des droits d'accès et des conditions d'utilisation (licences). Moins spécifiques aux entrepôts, différentes options existent pour l'export des données et métadonnées et la gestion de différents groupes d'utilisateurs, suivant des profils et droits définis. Plus rarement, ces dispositifs proposent des services pour explorer, visualiser ou analyser les données, et on s'approche ici d'autres dispositifs de données, plates-formes de calcul, etc.

D'après l'analyse de Kindling *et al*. (2017), la plupart des institutions en charge d'un entrepôt se trouvent dans quatre pays, à savoir les États-Unis, l'Allemagne, le Royaume-Uni et le Canada, qui représentent 70 % des institutions du répertoire *re3data*. Leur étude révèle également que 86 % des entrepôts sont disciplinaires et, pour une majeure partie (74 %), spécialisés dans une seule discipline ou domaine. Quant aux disciplines, les sites couvrant les sciences naturelles (52 %) et les sciences de la vie (50 %) sont les plus nombreux, suivis par les sciences humaines et sociales (SHS, 27 %) et les sciences de l'ingénieur (12 %).

D'après ce répertoire *re3data*, il n'existe pas seulement des entrepôts au sens strict, c'est-à-dire des dispositifs qui contiennent des jeux de données avec leurs métadonnées. On trouve également des portails qui moissonnent les métadonnées d'autres sites sans conserver les données elles-mêmes. D'une manière plus générale, les entrepôts sont considérés comme une sous-catégorie des infrastructures informationnelles de la recherche, pour la conservation et la diffusion des données de recherche.

Si, à partir de certaines caractéristiques fondamentales et des fonctionnalités communes, on définit les entrepôts de données comme une catégorie particulière des services de données ou, plus largement, des systèmes d'information destinés à la recherche scientifique, on distingue alors une nouvelle famille de systèmes. Mais cette approche opérationnelle ne constitue pas une définition heuristique satisfaisante, car elle limite l'analyse à la dimension technique et exclut les autres dimensions du concept de dispositif, sa nature de « cadre dans lequel des techniques, des humains sont disposés pour permettre de réaliser des activités répétitives et distribuées » et la dimension « collective », riche des aptitudes, du savoir-faire et des pratiques (Larroche, 2018). C'est le point de départ et le challenge pour les SIC – comment appréhender ces entrepôts sous l'angle d'une théorie des « dispositifs sociotechniques d'information et de communication », c'est-à-dire des « lieux de médiation composés de multiples facteurs sémiotiques, esthétiques et techniques en interaction, qui relient des acteurs sociaux dans des agencements et processus médiatés » (Bonfils et Durampart, 2013) ? Comment analyser ces dispositifs non pas (uniquement) comme systèmes d'information mais comme des outils développés par et pour certaines pratiques scientifiques ?

Peut-on dire pour autant, à l'instar des données de recherche, qu'un entrepôt de données est ce qu'une communauté scientifique définit comme tel ? Ce serait sans doute exagéré et cela mettrait en question la politique de certification de ces dispositifs par le Ministère ; cette politique fait abstraction des pratiques disciplinaires et des liens avec une communauté, en appliquant (uniquement) des critères techniques et fonctionnels. Nous y reviendrons.

Dans ce contexte, notre étude tente de répondre à la question suivante : dans quels dispositifs numériques les chercheurs en SIC peuvent-ils déposer leurs jeux de données pour conserver, signaler et partager les résultats de leur recherche, et contribuer à l'ouverture de la science ? Si le focus est bien sur les SIC, l'analyse inclura les domaines limitrophes et les dispositifs multidisciplinaires, dans la mesure où certains jeux de données SIC s'y trouvent.

La discussion portera sur la certification et la normalisation des entrepôts, sur la réponse qu'ils apportent aux besoins des chercheurs, et sur leur caractère communautaire. À partir des résultats et discussions, nous évoquerons également des pistes de recherche sur les données et leurs entrepôts.

4.
Méthodologie

L'objectif de l'étude est l'identification et l'analyse des dispositifs numériques dont disposent les chercheurs en SIC, pour déposer, conserver, signaler et partager leurs jeux de données. Pour répondre à cette question et pour identifier des sites, nous avons sélectionné un échantillon pertinent d'entrepôts de données, à partir des deux sources d'information suivantes :

- le wiki *Cat OPIDoR*[9] (Optimisation du Partage et de l'Interopérabilité des Données de la Recherche). Ce catalogue vise à recenser les services français dans ce domaine et propose un référencement selon neuf catégories de service, dont des entrepôts de données et des plates-formes d'archivage (Rebouillat, 2017). Au moment de l'analyse (septembre 2018), *Cat OPIDoR* enregistre 166 services ;
- le répertoire international *re3data*[10] (Pampel *et al.*, 2013) regroupe 2 190 entrepôts de données dans plusieurs dizaines de pays, dont 98 en France (septembre 2018).

À partir de la grille du wiki *Cat-OPIDoR* et de la liste des sites de *re3data*, notre étude propose de faire un état des lieux des dispositifs spécifiques au stockage des données de recherche en SIC. Seront inclus les entrepôts dans lesquels des donnés SIC sont recensées ; puis nous croiserons les différentes caractéristiques de ces données, afin de préciser leurs formats et spécificités disciplinaires. Les résultats seront discutés sous trois angles : la certification et normalisation, les fonctionnalités et l'aspect communautaire. À partir de cette cartographie, nous proposerons quelques pistes pour la recherche sur les données et pour le développement d'une infrastructure de données de recherche en SIC, en France et dans le contexte du développement des infrastructures européennes (*European Open Science Cloud*[11]).

L'étude met l'accent sur l'offre de service en France mais s'ouvre sur les entrepôts d'autres pays. Cette approche permettra d'inclure davantage de dispositifs, plus diversifiés. Cependant, cette approche a un inconvénient : le terme des SIC n'existe qu'en France, ailleurs *library science* (ou *library and information sciences*, LIS) et *communication science* sont le plus souvent séparés, avec en plus l'existence d'un autre domaine, les *media studies*. La constitution de l'échantillon tient compte de cette terminologie hétérogène.

9 Financé par le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, *Cat OPIDoR* est hébergé par l'INIST (CNRS) : https://cat.opidor.fr.

10 Financé par l'agence allemande DFG et coordonné par l'initiative internationale DataCite, *re3data* est hébergé par le Karlsruhe Institute of Technology (KIT) : https://www.re3data.org.

11 https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud#.

5.
Résultats

5.1. Présence des SIC dans les répertoires d'entrepôts de données

Parmi les 166 services de données, *Cat OPIDoR* répertorie cinq services dans le domaine « Ingénierie des systèmes et de la communication : Ingénierie électrique, électronique, de la communication, optique et des systèmes », avec certaines thématiques proches des SIC : citons par exemple le site *Gestion et diffusion des données Irstea*, site conçu par une équipe transversale (professionnels de l'information scientifique et technique, désormais IST, informaticiens, juristes, chargé de valorisation), pour les chercheurs et toute personne intervenant sur les données, en appui autour du cycle de vie des données ; le site propose des actualités, des fiches pratiques et des outils[12]. Un autre site est dédié au droit et à l'éthique dans les SHS, avec des contenus relatifs aux pratiques informationnelles des chercheurs. Mais il s'agit d'un service d'information sur la plate-forme *Hypothèses.org*, pas d'un entrepôt de données[13]. Dans le domaine des SIC, on trouve les sites des URFIST mais là encore, il ne s'agit pas d'entrepôts ou de plates-formes d'archivage mais de services de formation.

En fait, parmi les 44 entrepôts de données listés par *Cat OPIDoR*, avec des fonctionnalités liées au stockage et à la conservation, mais également à l'exposition et à la réutilisation des données, aucun site n'est explicitement dédié aux SIC. En revanche, plusieurs entrepôts correspondent au type de services génériques d'accueil et de diffusion de « données simples » ; ces entrepôts permettent le dépôt de résultats issus de la recherche en SIC, tels que *MédiHAL* sur la plate-forme HAL (pour les images scientifiques), *NAKALA* de Huma-Num pour tout type de données, voire *BeQuali* de SciencesPo Paris pour les enquêtes qualitatives.

Au plan international, selon l'indexation du répertoire *re3data*, un seul entrepôt est issu d'un département de sciences de l'information (*Spec Patterns*, de la Kansas State University), mais il s'agit de données informatiques.

14 entrepôts relèvent du domaine « *communication science* ». Tous ces entrepôts sont hébergés et maintenus par des organismes publics ou à but non lucratif (cf. tableau 1).

12 https://donnees-recherche.irstea.fr/.

13 https://ethiquedroit.hypotheses.org/.

Site	Description
UNESCO Institute for Statistics, Data Centre	Données statistiques de plus de 200 pays sur les domaines Éducation et alphabétisme, Science technologie et innovation, Culture, Communication et information.
Center for International Earth Science Information Network	CIESIN travaille à la croisée des sciences sociales, naturelles et de l'information et se spécialise dans la gestion de données et d'informations en ligne, l'intégration et la formation de données spatiales et la recherche interdisciplinaire liée aux interactions humaines dans l'environnement.
Data.gov.au	Fournit un accès public, le téléchargement et la réutilisation des données brutes des gouvernements australien, étatique et territorial. *Open data*.
Stanford Network Analysis Project	Bibliothèque d'analyse de réseau et d'extraction de graphiques à usage général.
UCI Machine Learning Repository	Héberge des ensembles de données dans les domaines de l'apprentissage des sciences et des logiciels éducatifs. Le site fournit également des outils en ligne pour analyser et signaler les données. University of California, Irvine.
PSLC DataShop	Ensemble de bases de données, de théories de domaine et de générateurs de données utilisés par la communauté d'apprentissage automatique pour l'analyse empirique des algorithmes d'apprentissage automatique. Pittsburgh Science of Learning Center.
KONECT	Projet de collecte de grands ensembles de données de réseau de tous types afin de mener des recherches en science des réseaux et dans des domaines connexes. Koblenz Network Collection.

American National Election Studies	Des enquêtes nationales et études pilotes fournissent de vastes ensembles de données à multiples facettes. Universités du Michigan et de Stanford, National Science Foundation.
Reality Commons	Ensemble de données mobiles collectés avec des outils développés dans le MIT Human Dynamics Lab disponibles en tant que projets *open source* ou à prix coûtant.
Open Government Data Portal of Odisha	Plate-forme de soutien à l'initiative *Open Data* du gouvernement d'Odisha (Inde).
Informatics Research Data Repository	Entrepôt japonais où sont collectées des données liées à l'informatique. Pilotage par trois instituts nationaux.
Social Computing Data Repository	Données de plusieurs médias sociaux. Arizona State University avec deux établissements militaires.
Sound and Vision	La plus large archive audiovisuelle en Europe. Institution publique avec musée à Hilversum, Pays Bas.
TalkBank	Transcriptions, audio et vidéo des interactions communicatives pour la recherche en communication humaine et animale. Université Carnegie Mellon, infrastructure européenne CLARIN ERIC, National Science Foundation, National Institute on Deafness and Other Communication Disorders.

Tableau 1 : Entrepôts de données en sciences de la communication (source : *re3data*)

Trois autres sites – *Data.uel*, l'entrepôt institutionnel pour les chercheurs de l'University of East London, *Claremont Colleges Digital Library* (États-Unis) et *BABS*, l'archive numérique de la Bibliothèque de l'État de Bavière – évoquent la bibliothéconomie (*library science*) parmi les domaines couverts. Néanmoins il s'agit de services institutionnels et génériques (multidisciplinaires), pour accueillir et stocker les données produites par les équipes des institutions en question.

Finalement, 52 autres sites sont référencés dans les études des médias, dont *DARIAH-DE* pour les humanités numériques, *eLaborate* de CLARIN-ERIC avec des transcriptions et annotations de textes et la *Phonothèque* de la MMSH (Aix-en-Provence) sur Huma-Num. Mais la plupart de ces sites mélangent plusieurs domaines, dont les arts plastiques, la musique et le théâtre, sans qu'on puisse parler d'une quelconque spécificité de SIC. Une remarque concernant HAL : *MédiHAL*, créé en 2010, permet de déposer des données visuelles et sonores (images fixes, vidéos et sons) ; à ce jour, il n'y a pas de données issues des SIC. Quant à HAL, son positionnement comme entrepôt de codes sources dans le cadre du projet *Software Heritage*, en collaboration avec l'INRIA, n'était pas opérationnel au moment de l'étude ; il faudra dans les mois à venir sans doute être attentif et vérifier si HAL devient un entrepôt générique sur le modèle de *Zenodo* ou *figshare*, avec un intérêt pour les données des SIC.

5.2. Disciplinarité

L'exploration des répertoires comme *re3data* et *Cat OPIDoR* fournit finalement peu de sites relevant des SIC. Aucun entrepôt de données n'est explicitement référencé comme « entrepôt SIC ». Quelques-uns contiennent des données de recherche en sciences de la communication, d'autres sont marqués comme services pour la recherche dans le domaine des médias. Encore moins mentionnent les sciences de l'information, de la documentation et/ou de la bibliothéconomie.

Certains sites se situent à l'intersection de plusieurs domaines scientifiques ou, comme *SpecPatterns*, sont mis en place par d'autres disciplines (en l'occurrence, l'informatique), malgré leur référencement dans les SIC. D'autres sites contiennent quelques données en sciences de la communication, mais aussi celles d'autres domaines, comme le génie civil ou physique (télécoms), l'informatique, l'économie, l'éducation, les sciences de la vie, les sciences politiques ou la linguistique.

D'autres jeux de données sont déposés dans des entrepôts génériques, avec une multitude d'autres disciplines. Souvent il s'agit d'entrepôts institutionnels, destinés à la préservation et à la diffusion les résultats des chercheurs d'un établissement ; parfois il s'agit aussi de sites à caractère national (*EASY* aux Pays-Bas) ou international (*Zenodo, figshare, DRYAD*).

L'indexation des sites sur *re3data* et *Cat OPIDoR* n'est pas totalement fiable, cohérente ou exhaustive. Les disciplines sont indexées à des niveaux de granularité variables, et, pour connaître le contenu réel ou la présence éventuelle de données issues des SIC, il faut aller sur les sites et vérifier directement sur les serveurs.

5.3. Localisation géographique

Où sont hébergés les services pour accueillir, stocker et diffuser les données issues de la recherche en SIC ? Huit des 14 entrepôts en sciences de com-

munication (cf. tableau 1) se trouvent aux États-Unis, les autres se répartissent en Allemagne, Australie ou en Inde, au Canada, au Japon, ou aux Pays-Bas et dans l'Union Européenne. Aucun ne se trouve en France. Les trois entrepôts limitrophes mentionnés plus haut se trouvent aux États-Unis, au Royaume-Uni et en Allemagne.

En revanche, la France héberge plusieurs entrepôts génériques qui contiennent ou pourraient contenir des données SIC, comme *NAKALA*, *Médi-HAL* ou *beQuali*. À ceci, on pourrait ajouter la *Phonothèque* de la MMSH.

Nous n'avons pas compté les entrepôts institutionnels (11 en France, plus de 500 dans d'autres pays) qui, *a priori*, pourraient accueillir des données SIC produites par les chercheurs des universités et des organismes de recherche concernés, mais qui n'affichent pas explicitement les SIC parmi les domaines couverts.

Néanmoins, il existe un certain nombre d'entrepôts multidisciplinaires et généralistes au plan international, qui représentent un intérêt certain pour les chercheurs SIC en France, en tant que services génériques d'accueil et de diffusion des « données simples ». Parmi les sites les plus connus, citons *Zenodo* (CERN), *figshare* (*Digital Science*), *DRYAD* (consortium international) ou encore *EASY* (DANS). Soutenu par l'Union Européenne, *Zenodo* héberge notamment plusieurs communautés dans les domaines SIC, avec toutefois peu de jeux de données à ce jour.

5.4. Contenus

Quels types de données se trouvent dans ces entrepôts ? Trouve-t-on des données caractéristiques pour les SIC, peut-on parler d'un « profil de données » SIC ? Suite à une analyse des données de recherche dans les thèses de doctorat en SHS, nous avons constaté les faits suivants :

- il n'existe pas de « données spécifiques SIC » ;
- il se dégage un profil caractéristique entre d'une part les données primaires ou sources composées d'enquêtes, de corpus textuels et de données web et, d'autre part les données secondaires ou résultats, composés de textes, de tableaux, de graphiques et de statistiques ;
- on trouve une sorte de longue traîne d'autres types de données, y compris des cartes, bases de données, photographies et chronologies (Schöpfel *et al*., 2015).

En d'autres mots, la particularité des SIC semble être le fait que ses résultats et, à un moindre degré, ses sources couvrent l'ensemble de la typologie des données de recherche en SHS, mettant en exergue la transversalité des SIC, et non pas leur spécificité ; ou plutôt, la spécificité des SIC, sous l'aspect des données de recherche, réside dans la transversalité des sources et résultats.

D'après *re3data*, les entrepôts contenant des données en SIC couvrent l'ensemble des types de données, à l'exception des données de configuration.

Les catégories les plus représentatives sont les documents bureautiques (fichiers texte, tableaux, etc.), des données statistiques, des textes structurés ou non structurés (plein texte) et des données d'archives. Mais tout cela ne contribue pas à un profil de données spécifique aux SIC, dont une particularité semble être le fait d'utiliser et de produire une large panoplie de données, y compris des données issues des bases de données (scientométrie) ou du web (analyse des réseaux et des usages).

Faute d'outils appropriés pour réaliser des analyses sur l'ensemble des sites, nous n'avons pas les moyens pour une estimation de la répartition des jeux de données SIC dans les différents types d'entrepôts. De même, il est impossible de fournir une idée générale de l'accessibilité des données, avec différents degrés d'ouverture (en libre accès sans restriction, partagées avec une communauté prédéfinie, accessibles sur demande, confidentielles…). Peut-être à terme cela deviendra-t-il faisable, grâce aux outils, qui, aujourd'hui encore, sont en *bêta test*, comme *Google Dataset Search*[14].

5.5. Métadonnées

L'enjeu des métadonnées pour la conservation, la diffusion et la réutilisation est considérable, aussi bien par rapport à leur spécificité et granularité que par rapport à leur normalisation. Leur rôle est crucial notamment pour la conformité des dispositifs avec les principes FAIR de la gestion des données de recherche (Wilkinson *et al.*, 2016). Ainsi, la richesse des métadonnées contribue à la repérabilité (*findability*) des données de recherche ; leur normalisation augmente l'interopérabilité et la réutilisation des données ; et l'usage d'un identifiant pérenne et normalisé facilité leur accessibilité[15].

Dans l'échantillon des 14 entrepôts de *re3data* avec des données en sciences de la communication, six seulement indiquent l'application d'un standard ou d'une norme : deux sites utilisent des formats « standards » développés par eux-mêmes (*repository-developed metadata schema*) ; les autres entrepôts utilisent quatre formats standardisés pour leurs métadonnées, deux formats génériques et deux formats pour les données géographiques :

- *DCAT*, un format pour la description de jeux de données dans des catalogues, utilisé notamment par l'application CKAN (le système de la plate-forme *open data* de l'Union Européenne) ;

- *RDF Data Cube Vocabulary* pour les métadonnées des données multidimensionnelles sur le web ;

- *ISO 19115*, la seule norme au sens strict, pour la description des données géographiques ;

14 https://toolbox.google.com/datasetsearch.

15 *Cf.* la description sur le site de Force 11 https://www.force11.org/group/fairgroup/fairprinciples.

- le standard *FGDC-CSDGM*, un format préconisé par le gouvernement américain pour la description des données géospatiales, mais qui n'est plus maintenu.

Certes, l'analyse de si peu d'entrepôts n'est pas représentative mais permet deux observations en ce qui concerne les dispositifs avec des données SIC. D'une part, aucun site n'utilise un format de métadonnées standardisé et générique, tel que le *DataCite Metadata Schema* de l'initiative DataCite ou simplement le Dublin Core. D'autre part, il n'y a aucune trace d'un standard plus spécifique aux domaines des SIC, voire d'un format en adéquation avec les méthodes des SIC, à l'instar par exemple du format standard de la *Data Documentation Initiative* (DDI) pour les enquêtes. Ce dernier constat confirme ce que le répertoire de référence pour les métadonnées des entrepôts de données, maintenu par le Digital Curation Centre du JISC (UK), avait déjà révélé : apparemment, il n'y a tout simplement pas de format spécifique aux SIC ou aux domaines limitrophes[16].

Ce qui est peut-être encore plus surprenant dans notre échantillon, c'est l'absence quasi totale d'un recours à un ou plusieurs systèmes d'identifiants pérennes, comme par exemple le DOI, l'ORCID, le handle, etc. Seulement deux des 14 entrepôts de l'échantillon appliquent un tel système (*CIESIN* avec le DOI, *TalkBank* avec le handle). Un tel constat limite encore davantage la conformité de ces sites avec les principes FAIR, ici en particulier avec la repérabilité et l'accessibilité des données de recherche.

Deux autres observations : seulement une minorité des sites utilise des licences ouvertes pour la diffusion des données (*Creative Commons, Open Data Commons, Open Government License*), et aucun n'affiche une API conforme au protocole OAI-PMH.

Quant aux dispositifs génériques, tels que *Zenodo*, *figshare* ou *DRYAD*, ils appliquent des standards génériques des infrastructures internationales, dont notamment OAI-ORE pour les archives ouvertes, le Dublin Core et le format de DataCite, préconisé par l'Union Européenne; ils utilisent également les identifiants standards (DOI, handle, etc.).

6.
Discussion

L'objectif de notre analyse est de mieux comprendre la place et les opportunités des SIC dans le paysage émergent des entrepôts des données de recherche. Il ne s'agit pas d'une cartographie exhaustive, dans la mesure où le recensement des sites et l'étude de leurs caractéristiques s'appuie essentiel-

16 http://www.dcc.ac.uk/resources/metadata-standards.

lement sur les répertoires *re3data* et *Cat OPIDoR*, avec leurs richesses, mais aussi leurs limites. Aussi, nous cantonnerons la discussion de nos résultats à trois aspects, dont nous ferons une interprétation prudente.

6.1. Certification et normalisation

Le *Plan d'action* du Ministère préconise l'engagement d'un processus de certification des infrastructures de données. Cette certification a comme objectif de garantir des infrastructures de données avec une certaine fiabilité et qualité, afin de créer un « *trustworthy environment* » pour la conservation et le partage des données de recherche. Il existe plusieurs catégories de certificats ou de labels, certains pour les archives numériques en général (ISO 14641 pour les systèmes d'archivage électronique), d'autres pour des plates-formes d'information scientifique (*cf.* Schöpfel et Müller, 2014 pour les archives ouvertes) ; certains certificats sont spécifiques aux entrepôts de données : citons la norme ISO 16919 (*Systèmes de transfert des informations et données spatiales*), la certification du International Science Council World Data System (ICSU-WDS)[17], le Data Seal of Approval (DSA)[18], et plus récemment le CoreTrustSeal[19].

Trois des 14 entrepôts en sciences de la communication sont des plates-formes certifiées, c'est-à-dire des sites qui garantissent la conservation et l'intégrité de leurs données. Il s'agit des certificats du DSA (pour le site *Sound and Vision*), de CLARIN (certificat pour « centres B » dans le domaine des sciences du langage)[20] et du nouveau CoreTrustSeal (tous les deux pour *TalkBank*). *TalkBank* et *Sound and Vision* ont également engagé une démarche qualité, tout comme *UIS.stat* de l'UNESCO, *CIESIN* (Center for International Earth Science Information Network) et la plate-forme *data.gov.au* pour la diffusion des données ouvertes du gouvernement australien. *NAKALA, Zenodo* et *figshare* mènent également une politique de qualité, mais aucun n'affiche une certification à ce jour.

Si la certification contribue à la crédibilité d'un dispositif, la normalisation quant à elle augmente son interopérabilité et d'une manière générale, sa conformité avec les principes FAIR d'une bonne gestion des données de recherche. Or, en ce qui concerne les métadonnées, notre étude révèle un manque de normalisation pour la plupart des sites évalués. Seulement les grands entrepôts génériques appliquent les standards internationaux pour la gestion des données de recherche. Même constat pour les identifiants. Aussi, nous n'avons trouvé aucun format spécifique aux SIC, ni parmi les entrepôts, ni dans le répertoire des formats de métadonnées du JISC. S'agit-il d'un simple retard, ou est-ce l'expression de la diversité méthodologique et thématique

17 https://www.icsu-wds.org/services/certification.

18 https://www.datasealofapproval.org/en/.

19 https://www.coretrustseal.org/about/.

20 https://www.clarin.eu/content/assessment-procedure.

de la discipline ? La question reste sans réponse pour l'instant, en attendant l'évolution des sites thématiques et disciplinaires préconisée par le *Plan national* du Ministère, sachant que le Ministère s'engage clairement dans la voie d'une certification des entrepôts français avec le CoreTrustSeal.

L'objectif d'une telle démarche est de renforcer la sécurité de ces dispositifs numériques, avec une garantie de qualité de service basée sur les critères objectifs du certificat. Pour les SIC, la certification représente un triple enjeu scientifique.

Le lien avec la communauté : le concept des dispositifs numériques met en avant le lien organique entre les systèmes et les pratiques ; la certification met en œuvre une approche « agnostique », sans lien avec des pratiques et fonctions spécifiques ou disciplinaires. Mais les critères de « crédibilité », « fiabilité » ou « pérennité » ont-ils le même sens et le même intérêt pour tous les usages et pratiques de données ? Le caractère de « *trustworthiness* », la confiance dans les données est en passe de devenir un enjeu central de la gestion des données et du *Big Data* en général ; mais la notion de « *trustworthy* » a-t-elle la même signification pour les données médicales, pour les données produites par un accélérateur de particules et pour les données issues d'un sondage en sciences politiques ?

Le caractère statique : une certification attribue un label de qualité et de « *trustworthiness* » d'une manière statique, à un moment donné et pour un certain temps. Or, les dispositifs, leurs contenus et leurs caractéristiques évoluent vite, ce qui pose la question de la validité d'une telle démarche dans le temps. Une évaluation dynamique et continue de plusieurs indicateurs (standards, identifiants, ouverture, licences, volumétrie…) paraît davantage appropriée à ces dispositifs ; du moins, elle devrait accompagner et compléter la démarche de certification.

La qualité des données : la certification évalue certains aspects considérés comme cruciaux pour la qualité et la crédibilité des entrepôts. Or, ce genre de label ne garantit pas la qualité des contenus, c'est-à-dire des données déposées. C'est un peu comme le facteur d'impact et la sélectivité des revues scientifiques : l'impact ou la réputation du contenant ne disent *a priori* rien sur la qualité du contenu. D'autres variables entrent en jeu, y compris la « *trustworthiness* » du déposant (chercheur, établissement…). L'écosystème des entrepôts de données de recherche est loin, aujourd'hui, d'une solution à cette question pourtant essentielle pour la réutilisation des données et leur conservation à plus long terme.

6.2. Gestion, diffusion et/ou conservation

Notre étude s'est volontairement limitée aux seuls entrepôts de données, comme l'un des neuf types de services recensés par *Cat OPIDoR*. Cependant, un entrepôt de données peut remplir d'autres fonctions par rapport au cycle de

vie des données de recherche, dont notamment la conservation et la diffusion, mais aussi la gestion des résultats de la recherche.

Nos études antérieures ont montré comment la grande diversité de la nature des données et des conditions de leur production trouve son reflet dans les pratiques de gestion. L'enquête de 2015 (Prost et Schöpfel, 2015) en a fourni une image détaillée, composée d'une importante variation de pratiques mais aussi, de certaines caractéristiques prépondérantes. À titre d'illustration, le stockage en local est de loin le mode de sauvegarde privilégié, que ce soit sur l'ordinateur privé ou sur leur ordinateur professionnel. 19 % des chercheurs stockent « dans le *cloud* », alors que 8 % ont des données sur le serveur d'une autre institution. En réseau, 12 % des répondants se tournent vers le serveur de l'université. L'enjeu des entrepôts est lié au fait que certaines pratiques de stockage et de communication présentent des risques de sécurité (*cf.* aussi Simukovic *et al.*, 2014 ; Serres *et al.*, 2017).

Tous les sites étudiés garantissent un stockage des données à court ou moyen terme. En revanche, peu étendent cette garantie à une conservation de plus long terme, via une certification DSA ou autre. Concrètement, cela signifie que seulement trois sur les 14 sites pour les sciences de la communication affichent une telle garantie. Quant aux sites français, la conservation pérenne est prise en charge par le CINES, référencé par *Cat OPIDoR* comme l'unique dispositif d'archivage pérenne de données. Par le biais de partenariats et contrats, cette garantie de conservation à plus long terme concerne aussi d'autres services en France, en particulier ceux de la TGIR Huma-Num (*NAKALA, Phonothèque* MMSH).

De nouveau, les grands entrepôts génériques se démarquent par un engagement de conservation certifié. En revanche, la situation paraît bien plus incertaine et précaire pour beaucoup de sites institutionnels ou thématiques.

Outre le stockage, les entrepôts prennent généralement en charge la diffusion des données, par l'exposition des métadonnées et par la gestion des accès. Nous avons évoqué plus haut la question des métadonnées ; des métadonnées non normalisées ne facilitent pas la recherche et la découverte des jeux de données. Quant à la gestion des accès, tous les dispositifs analysés sont en libre accès, c'est-à-dire peuvent être interrogés librement et sans restriction. En revanche, certains sites contrôlent l'accès aux données elles-mêmes, par l'enregistrement obligatoire des utilisateurs. Les entrepôts permettent généralement aux propriétaires ou dépositaires des données de choisir le degré d'ouverture ou d'accessibilité des données (libre, avec des restrictions, sous embargo, etc.), certains entrepôts indiquent également les conditions légales de la diffusion – avec ou sans licence, protection par le copyright, etc. De tout cela se dégage de nouveau l'impression d'un paysage émergent, hétérogène, transitoire peut-être, incohérent, ce qui nuit à la compatibilité et conformité de ces sites avec les critères FAIR, y compris pour la réutilisation des données.

Reste la question de la gestion des données de recherche pendant la durée d'un projet (« données chaudes »), un point essentiel quand on parle avec les chercheurs de leurs préoccupation en matière de données (*cf.* Serres *et al.*, 2017). Les entrepôts de notre étude tiennent généralement compte de ces besoins et préoccupations, mais ils le font à des degrés variables.

Dépôt des données : à l'exception de deux sites, le dépôt des données est contrôlé et conditionné par l'affiliation institutionnelle ou par l'inscription au service. Trois sites déclarent le dépôt comme « fermé » ; en d'autres termes, l'auto-archivage de données par les chercheurs eux-mêmes n'est pas proposé, le dépôt est réservé au personnel de l'hébergeur.

Mises à jour : la majorité des sites étudiés permet la gestion de plusieurs versions du même jeu de données.

Sécurité : nous n'avons pas trouvé d'information précise sur la sécurisation des échanges et du stockage, notamment pour les données confidentielles ou à caractère personnel. Ceci veut probablement dire que ces sites ne sont pas spécialement conçus pour ce besoin : cependant, les entrepôts présentant une certification (voir plus haut) peuvent satisfaire une partie des contraintes.

Calcul, exploitation : les entrepôts de données ne sont généralement pas conçus pour le calcul ou l'analyse des dépôts de données. Néanmoins, certains sites, comme le *Stanford Network Analysis Project* ou le *Center for International Earth Science Information Network* (CIESIN) proposent des fonctionnalités et outils d'analyse. Ces entrepôts deviennent ainsi de véritables plates-formes pour la gestion, la conservation et la diffusion des données.

Toujours est-il que, comme mentionné plus haut, dans tous les cas de figure, il faudra faire le lien avec les pratiques et usages des chercheurs concernés, au lieu de comparer les fonctionnalités d'une manière abstraite. Ce qui peut paraître dysfonctionnel d'une manière générale et absolue peut faire sens sur un terrain particulier. De nouveau, il faudra mobiliser le concept des dispositifs numériques pour ne pas occulter ces liens.

6.3. Communauté et transversalité

L'objectif de notre étude était d'identifier et de décrire des entrepôts de données qui permettent de déposer, conserver et diffuser des résultats de la recherche en SIC. À l'issue de l'exploration des répertoires des entrepôts et services de données, nous avons trouvé des sites qui sont indexés SIC, mais qui contiennent surtout des résultats de la recherche en informatique (*computer science*) ou en génie civil (télécommunications), puis d'autres sites qui contiennent effectivement des résultats SIC, mais comme un domaine parmi d'autres, dont certains avec un caractère institutionnel, c'est-à-dire réservés aux seuls chercheurs de l'institution propriétaire du site. Nous avons également trouvé quelques entrepôts pour certains domaines ou outils des SIC (média, enquêtes…), eux aussi partagés avec d'autres disciplines. Autrement dit, on

trouve bien des entrepôts avec des données SIC. Mais aucun entrepôt SIC au sens strict, réservé aux chercheurs en SIC, avec une visibilité forte et un caractère communautaire explicite et confirmé. À ceci s'ajoutent deux autres constats.

- D'une part, les entrepôts avec des données SIC couvrent en général une large gamme de données. Ceci est certainement lié au caractère générique des sites institutionnels ou internationaux, comme *figshare* ou *Zenodo*. Mais il y a peut-être une autre raison, liée directement à la nature des SIC et évoquée plus haut. L'analyse des données dans les thèses de doctorat SHS de l'Université de Lille 3 avait abouti au même résultat (Schöpfel *et al*., 2015) : même si certaines catégories de données sont mieux représentées que d'autres (en particulier, les textes et tableaux), les résultats publiés avec les thèses couvrent l'ensemble de la typologie des données de recherche, y compris des bases de données, cartes et photos.

- D'autre part, apparemment il n'existe pas de métadonnées spécifiques aux SIC, ni dans les entrepôts, ni dans les sites de recensement des différents formats et standards.

Simple question de retard ou de transition ? Ou s'agit-il d'une transversalité revendiquée et assumée des SIC, où l'affinité avec les domaines limitrophes (informatique, linguistique, télécommunications, etc.) prend le dessus sur les pratiques et outils communautaires ? Les résultats de notre étude posent cette question, mais la taille réduite de l'échantillon et la problématique du référencement et de l'indexation par les répertoires incitent à une certaine prudence quant à donner une réponse définitive.

Les deux enquêtes récentes de Lille (Schöpfel, 2018) et de Rennes (Serres *et al*., 2017) ont révélé l'attente forte des chercheurs en SHS en matière de gestion des données, en particulier pour un stockage protégé et une communication sécurisée au sein des laboratoires et projets. La gestion des données est un défi quotidien qui mobilise les ressources et dispositifs locaux, mais également les services, outils et infrastructures d'envergure nationale ou internationale. Mais dans le domaine des données de recherche, il n'existe pas de solution unique. Pour répondre aux besoins des chercheurs, il faut prendre en compte leurs pratiques. Certaines communautés sont plus avancées que d'autres dans l'archivage et le partage des résultats scientifiques. Ce qui fonctionne chez les uns ne marche peut-être pas pour d'autres. Il faut abandonner l'idée d'une solution unique en faveur d'une approche modulaire, différenciée, par option.

Conclusion

Les entrepôts de données jouent un rôle central pour la gestion, le stockage, la conservation et la diffusion des données de recherche. Aussi, une part importante des efforts publics porte sur le développement et l'interconnexion des sites, plates-formes et infrastructures de données, à l'échelle nationale,

européenne et internationale. Le *Plan national pour une science ouverte* préconise la mise en place de dispositifs thématiques et disciplinaires et pose la question d'une plate-forme générique, tandis que l'Union Européenne soutient l'interopérabilité et l'interconnexion des infrastructures, au sein d'un *European Open Science Cloud* (EOSC).

Or, la plupart des entrepôts de données sont aujourd'hui répertoriés dans les sciences de la vie (médecine, biologie) et en sciences de l'univers, y compris la géographie. Quant aux sciences de l'information et de la communication, leurs données se trouvent éparpillées dans un certain nombre d'entrepôts, thématiques, institutionnels ou génériques, sans qu'aucun ne puisse revendiquer un label « entrepôt SIC ». Notre étude n'est qu'une photographie du paysage à un moment précis, fin 2018, et ce paysage évolue rapidement. Force est cependant de constater qu'il n'existe pas de site clairement labellisé SIC (ou au niveau international, « *Library and Information Sciences* »), et pour les autres dispositifs, il n'y a pas de distinction claire entre les SIC, l'informatique, les médias et les télécommunications. Ce constat concernant les sites est corroboré par l'absence d'une description disciplinaire puisqu'il n'y a pas de format de métadonnées propre aux SIC. Certes, l'application des formats génériques renforce l'interopérabilité des dispositifs, mais elle affaiblit aussi la cohérence communautaire au sein des infrastructures de la science ouverte.

On peut formuler le constat aussi d'une autre manière : les SIC, en tant que communauté disciplinaire, ne sont pas prêtes face aux enjeux de la science ouverte, en ce qui concerne les données de recherche. Dans cette situation, plusieurs options se présentent pour les SIC :

- contribuer à la mise en place d'un service de données SIC (simple agrégateur ou véritable entrepôt), compatible avec les principes FAIR des infrastructures européennes ;
- améliorer le référencement et l'indexation des données SIC dans les entrepôts institutionnels, thématiques ou génériques ;
- créer (ou fédérer) une communauté SIC sur les plates-formes génériques, notamment sur *Zenodo* ;
- participer à la démarche de normalisation des formats et métadonnées des données de recherche, en explorant l'intérêt potentiel de spécifications SIC ;
- promouvoir l'adoption d'une politique de données ouvertes associées aux articles publiés dans les revues SIC.

Un tout autre enjeu est l'appropriation scientifique du terrain des données de recherche et de leurs entrepôts par les SIC. À ce jour, la plupart des études abordent le sujet sous l'angle technique (comment faire ?), politique (quelle stratégie ?), opérationnel (quelle organisation ?) ou éthique (comment partager ?). Les SIC disposent de plusieurs cadres conceptuels pour apporter une contribution originale à la compréhension de l'évolution en cours, et notre étude

a évoqué plusieurs pistes : l'analyse des entrepôts comme nouveaux dispositifs numériques, l'analyse des pratiques sous l'aspect de la culture informationnelle, l'évaluation de l'information scientifique et l'analyse du concept des données dans l'environnement de l'organisation des connaissances.

La politique nationale et européenne pour la science ouverte offre aux SIC l'opportunité d'un positionnement communautaire autour d'un dispositif disciplinaire pour la conservation et la diffusion des données issues de la recherche en SIC. Un tel positionnement inclurait nécessairement une réflexion autour d'une description disciplinaire des données. Mais pour l'instant, on n'en est pas là, et peut-être suffit-il d'assurer la connexion entre publications et données dans les revues SIC et laisser aux chercheurs le choix de leur entrepôt, comme aujourd'hui, sans investir dans une infrastructure de données SIC. C'est l'option alternative, la plus simple, sans besoin d'investissement ou de coordination ; mais elle ne contribuera pas à l'affirmation et au développement de la discipline des SIC.

Remerciements

L'étude à l'origine de cette proposition a été réalisée dans le projet *D4Humanities* (*Deposit of Dissertation Data in Social Sciences and Humanities – A project in Digital Humanities*). Ce projet est financé dans le cadre des projets structurants de la MESHS 2017-2018 (Contrat de plan État-Région « ISI-MESHS »), par la MESHS et le Conseil Régional Hauts-de-France.

Bibliographie

André F. (2015). « Déluge des données de la recherche ? ». In Calderan L., Laurent P., Lowinger H., Millet J. (dir.). *Big data : nouvelles partitions de l'information*, Louvain-la-Neuve, De Boeck, p. 77-95.

Bailo D., Jeffery K. G. (2014). "EPOS: a novel use of CERIF for data intensive science". In *CRIS2014: 12th International Conference on Current Research Information Systems*, 13-15 May, Rome.

Bonfils P., Durampart M. (2013). « Environnements immersifs et dispositifs numériques. Études expérimentales et approches distanciées ». In *ESSACHESS – Journal for Communication Studies*, vol. 6, n° 1, p. 107-124. Disponible sur http://www.essachess.com/index.php/jcs/article/view/196 (page consultée le 27 février 2019).

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

Chignard S. (2012). *Open data : comprendre l'ouverture des données publiques*, Limoges, Fyp.

Etalab (2018). *Pour une action publique transparente et collaborative : plan d'action national pour la France 2018-2020*. Paris, Secrétariat d'État chargé de la Réforme de l'État et de la Simplification. Disponible sur https://www.etalab.gouv.fr/wp-content/uploads/2018/04/PlanOGP-FR-2018-2020-VF-FR.pdf (page consultée le 27 février 2019).

Kindling M., Pampel H., Van de Sandt S., Rücknagel J., Vierkant P., Kloska G., Witt M., Schirmbacher P. (2017). "The Landscape of Research Data Repositories in 2015: A re3data Analysis". In *D-Lib Magazine*, vol. 23, n° 3/4. Disponible sur http://www.dlib.org/dlib/march17/kindling/03kindling.html (page consultée le 27 février 2019). Doi : 10.1045/march2017-kindling

Koltay T. (2016). « Digital research data ». In Baker D., Evans W. (dir.), *Digital Information Strategies*, Oxford, Chandos Publishing, p. 71-84.

Larroche V. (2018). *Le dispositif : un concept pour les sciences de l'information et de la communication*, London, ISTE Editions.

Matthews B., Wilson M. D., Kleese Van Dam K. (2002). "Accessing the outputs of scientific projects". In *CRIS2002: 6th International Conference on Current Research Information Systems*, Kassel, August 29-31. Disponible sur http://dspace-cris.eurocris.org/handle/11366/149 (page consultée le 27 février 2019).

Organisation de Coopération et de Développement Économiques (OCDE) (2007). *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*. Editions OCDE, Paris. Disponible sur http://www.oecd.org/fr/science/

sci-tech/38500823.pdf (page consultée le 10 septembre 2018).

Pain M. (2016). *Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation : étude de cas pour l'archive HAL*. Mémoire de Master, Villeurbanne, Enssib. Disponible sur https://memsic.ccsd.cnrs.fr/mem_01374509 (page consultée le 27 février 2019).

Pampel H., Vierkant P., Scholze F., Bertelmann R., Kindling M., Klump J., Goebelbecker H.-J., Gundlach J., Schirmbacher P., Dierolf U. (2013). "Making Research Data Repositories Visible: The re3data.org Registry". In *PLoS ONE*, vol. 8, n° 11, e78080. Disponible sur http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078080 (page consultée le 27 février 2019).

Prost H., Schöpfel J. (2015). *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. Rapport final*, Villeneuve d'Ascq, Université de Lille 3. Disponible sur http://hal.univ-lille3.fr/hal-01198379v1 (page consultée le 27 février 2019).

Rebouillat V. (2017). "Inventory of Research Data Management Services in France". In Chan L., Loizides F. (dir.), *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices*, IOS Press, Amsterdam, p. 174-181. Disponible sur http://ebooks.iospress.nl/publication/46651 (page consultée le 27 février 2019).

Reymonet N. (2017). *Améliorer l'exposition des données de la recherche : la publication de data papers*. Paris, Université Paris Diderot. Disponible sur https://archivesic.ccsd.cnrs.fr/sic_01427978 (page consultée le 27 février 2019).

RIN (2008). *Stewardship of digital research data: a framework of principles and guidelines*. London, Research Information Network. Disponible sur http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines (page consultée le 27 février 2019).

Schöpfel J. (2018). *Vers une culture de la donnée en SHS. Une étude à l'Université de Lille*. Lille, Université de Lille. Disponible sur https://hal.archives-ouvertes.fr/GERIICO/hal-01846849v1 (page consultée le 27 février 2019).

Schöpfel J., Müller U. (2014). « Évaluer la qualité des archives ouvertes : Le certificat DINI ». In *Partnership*, vol. 9, n° 1, p. 1-21. Disponible sur https://journal.lib.uoguelph.ca/index.php/perj/article/view/2733/3226 (page consultée le 27 février 2019).

Schöpfel J., Primož J., Prost H., Malleret C., Češarek A.,Koler-Povh T. (2015). "Dissertations and Data" (Keynote Address). In *GL17 International Conference on Grey Literature*, 1-2 December, Amsterdam. Disponible sur https://hal.archives-ouvertes.fr/hal-01285304 (page consultée le 27 février 2019)

Schöpfel J., Prost H., Rebouillat V. (2016). "Research data in current

research information systems". In *CRIS 2016: 13th International Conference on Current Research Information Systems*, 8-11 June, St Andrews. Disponible sur http://dspacecris.eurocris.org/handle/11366/501 (page consultée le 27 février 2019).

Schöpfel J., Kergosien E., Prost H. (2017). « Pour commencer, pourriez-vous définir 'données de la recherche' ? Une tentative de réponse ». Toulouse, INFORSID 2017. Disponible sur https://hal.univ-lille3.fr/hal-01530937 (page consultée le 27 février 2019).

Serres A., Malingre M.-L., Mignon M., Pierre C., Collet D. (2017). *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2*. Université Rennes 2. Disponible sur https://hal.archives-ouvertes.fr/hal-01635186 (page consultée le 27 février 2019).

Simukovic E., Kindling M., Schirmbacher P. (2014). "Unveiling Research Data Stocks: A Case of Humboldt-Universität zu Berlin". In *iConference 2014 Proceedings*, 4-7 March 2014, Berlin, p. 742-748. Disponible sur http://hdl.handle.net/2142/47259 (page consultée le 27 février 2019).

Wilkinson M. D., Dumontier M., Aalbersberg I. J. *et al*. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In *Scientific Data*, vol. 3, 160018. Disponible sur https://www.nature.com/articles/sdata201618 (page consultée le 27 février 2019).