



HAL
open science

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

Hélène Prost, Joachim Schöpfel

► To cite this version:

Hélène Prost, Joachim Schöpfel. Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique. *Études de communication - Langages, information, médiations*, 2019, 52 (1), pp.71-98. hal-02878141

HAL Id: hal-02878141

<https://hal.science/hal-02878141v1>

Submitted on 7 Jul 2020

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

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

Les entrepôts de données en sciences de l'information et de la communication (SIC)

Une étude empirique

Hélène Prost

CNRS, membre associée du laboratoire GERiCO

helene.prost007@gmail.com

Joachim Schöpfel

Université de Lille, Laboratoire GERiCO

joachim.schopfel@univ-lille.fr

Un triple enjeu pour les SIC

Dans la continuité des chantiers engagés pour la transformation numérique de l'Etat, le gouvernement français met en place une action publique plus transparente et plus collaborative. Le Plan d'action 2018-2020¹ 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'Etat français qui soutient « la mise en œuvre des principes du gouvernement ouvert pour renforcer (...) l'accès aux matériaux et résultats de la recherche » (loc.cit.). 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 » (loc.cit.). Le Plan d'action ne définit pas clairement la science ouverte (*open science* en Anglais) 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 neuf mesures et 27 actions, réparties sur trois axes prioritaires, dont la structuration et l'ouverture des données de recherche² (cf. figure 1). En tête des mesures figure l'obligation de la diffusion ouverte des

¹ Etalab (2018). 'Pour une action publique transparente et collaborative : plan d'action national pour la France 2018-2020'. Secrétariat d'Etat chargé de la Réforme de l'Etat et de la Simplification, Paris.

<https://www.etalab.gouv.fr/wp-content/uploads/2018/04/PlanOGP-FR-2018-2020-VF-FR.pdf>

² MESRI (2018). 'Plan national pour la science ouverte'. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Paris. <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>

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.

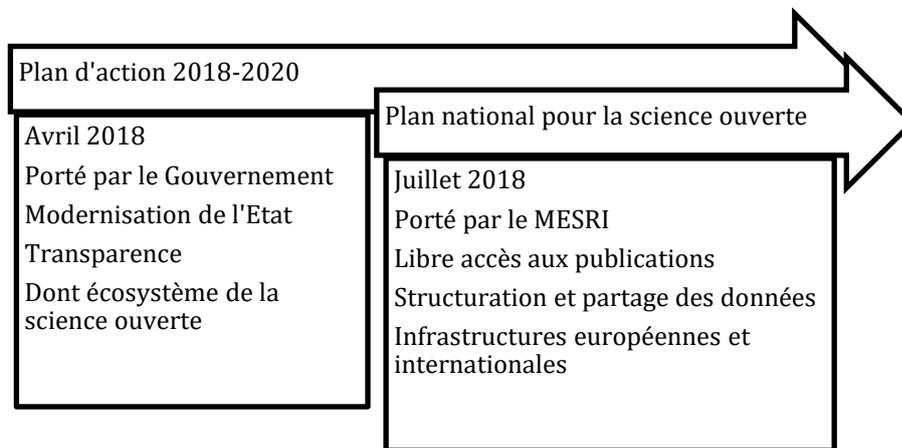


Figure 1. Les deux plans français en faveur de la science ouverte

La situation évolue rapidement. Fin 2018, le Ministère a mis en place un Comité pour la Science Ouverte (CoSO)³ 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 Alliance⁴, et son Collège Données a fait des recommandations concernant la gestion des données dans les projets de l'Agence Nationale de la Recherche, travaille 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 Dataverse⁵, 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. 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

³ <https://forum.ouvrirlascience.fr/>

⁴ <https://www.rd-alliance.org/groups/rda-france>

⁵ Dont l'INRA, le CIRAD et Sciences Po Paris

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 allons d'abord évoquer quelques problèmes de terminologie pour ensuite, après la présentation des résultats, aborder 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.

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 soit 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 données 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).

A ceci s'ajoute une très grande diversité qui a fait l'objet de nombreuses typologies, hiérarchies et modèles. A partir d'une analyse des finalités et procédures de la génération des données de recherche, le *Research Information Network* (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*⁶ qui distingue quatre niveaux (Bailo & Jeffery 2014) :

- Données brutes, ou données de base (ex. sismogramme)
- Données issus de procédures quasi-automatisées (ex. localisation des tremblements de terre)
- Données issus des recherches scientifiques (ex. modèles de la croûte terrestre)
- Données intégrés issus 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, et il 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.

A partir de l'indexation d'un grand nombre d'entrepôts de données, le répertoire international re3data⁷ différencie quatorze types de données, avec une répartition inégale entre plusieurs larges catégories, transversales aux disciplines et aux contours mal définis, et une longue traîne d'autres

⁶ EPOS <https://www.epos-ip.org/>

⁷ <https://www.re3data.org/>

types de données dont certains avec un profil nettement plus disciplinaire (Kindling et al. 2017). Cette dernière approche est particulièrement intéressante pour notre propre étude ; là où les autres analyses essaient 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. A première vue, cette approche paraît décontextualiser les données de recherche, en coupant le lien avec leur fonctions initiales et les pratiques des chercheurs. En réalité, il n'en est rien, comme l'analyse du concept des entrepôts comme dispositifs numériques va démontrer.

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 »⁸, des entrepôts de données (*data repository* en anglais) 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.
- Institutionnelles : 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.⁹ La garantie d'un accès à long terme (pérenne) implique que les conditions d'accès soient clairement explicitées.

⁸ Sylvie Cocard - Action Nationale de Formation RENATIS – MEDICI 2017 (3-6/07/2017) : Participer l'organisation du management des données de la recherche, gestion de contenu et documentation des données

⁹ 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>

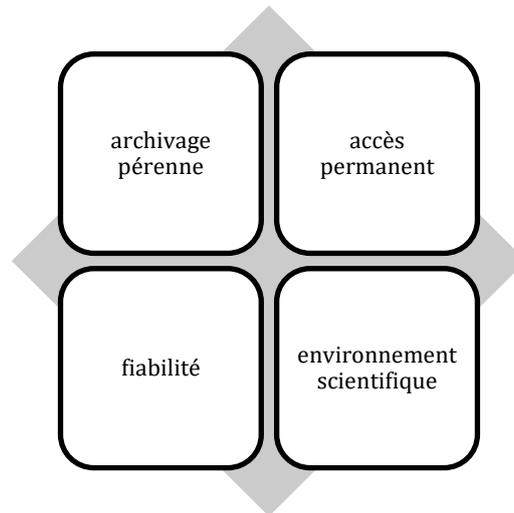


Figure 2. Quatre éléments constitutifs d'un entrepôt de données (source : re3data)

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 avec 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 sont différentes options pour l'export des données et métadonnées et la gestion de différents groupes d'utilisateurs, avec 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, plateformes 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 Etats 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 fréquents, suivis par les 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.

La définition d'entrepôts de données comme une catégorie particulière des services de données ou, plus largement, des systèmes d'information destinés à la recherche scientifique, à partir de certains caractéristiques fondamentaux et des fonctionnalités communes permet de distinguer et de décrire 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 », avec des aptitudes, du savoir-faire, 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 relie des acteurs sociaux dans des agencements et processus médiatisés » (Bonfils & 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 d'aller trop loin et mettrait en question la politique de certification de ces dispositifs par le Ministère qui 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 essaie 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 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 leur caractère communautaire. A partir des résultats et discussions, nous allons également évoquer des pistes de recherche sur les données et leurs entrepôts.

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 et sélectionner un échantillon pertinent d'entrepôts de données, l'étude exploite deux sources d'information :

- Le wiki *Cat OPIDoR*¹⁰ (Optimisation du Partage et de l'Interopérabilité des Données de la Recherche), un catalogue qui vise à recenser les services français dans ce domaine et propose un référencement de neuf catégories de service, dont des entrepôts de données et des plateformes d'archivage. Au moment de l'analyse (septembre 2018), *Cat OPIDoR* enregistre 166 services.
- Le répertoire international *re3data*¹¹ référence 2 190 entrepôts de données dans plusieurs dizaines de pays, dont 98 en France (septembre 2018).

A 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 données 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. A partir de cette cartographie, nous allons proposer 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 Cloud¹²).

¹⁰ Financé par le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, *Cat OPIDoR* est hébergé par l'INIST (CNRS), cf. <https://cat.opidor.fr>

¹¹ 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>

¹² <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud#>

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, avec une plus grande diversité. Cependant, elle 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.

Résultats

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 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 outils¹³. 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 plateforme *Hypothèses.org*, pas d'un entrepôt de données¹⁴. Dans le domaine des SIC, on trouve les sites des URFIST mais là encore, il ne s'agit pas d'entrepôts ou de plateformes 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", qui permettent le dépôt de résultats issus de la recherche en SIC, tels que *MédiHAL* sur la plateforme 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).

Site	Description
<i>UNESCO Institute for Statistics, Data Centre</i>	Données statistiques de plus de 200 pays sur les domaines Education et l'alphabétisme, Science technologie et innovation, Culture, Communication et information
<i>Center for International Earth Science Information Network</i>	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.

¹³ <https://donnees-recherche.irstea.fr/>

¹⁴ <https://ethiquedroit.hypotheses.org/>

<i>Data.gov.au</i>	Fournit un accès public, le téléchargement et la réutilisation des données brutes des gouvernements australien, étatique et territorial. Open data.
<i>Stanford Network Analysis Project</i>	Bibliothèque d'analyse de réseau et d'extraction de graphiques à usage général.
<i>UCI Machine Learning Repository</i>	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.
<i>PSLC DataShop</i>	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.
<i>KONECT</i>	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.
<i>American National Election Studies</i>	Des enquêtes nationales et études pilotes fournissent de vastes ensembles de données à multiples facettes. Universités du Michigan et de Stanford, National Science Foundation.
<i>Reality Commons</i>	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.
<i>Open Government Data Portal of Odisha</i>	Plate-forme de soutien à l'initiative Open Data du gouvernement d'Odisha (Inde)
<i>Informatics Research Data Repository</i>	Entrepôt japonais où sont collectées des données liées à l'informatique. Pilotage par trois instituts nationaux.
<i>Social Computing Data Repository</i>	Données de plusieurs médias sociaux. Arizona State University avec deux établissements militaires.
<i>Sound and Vision</i>	La plus large archive audiovisuelle en Europe. Institution publique avec musée à Hilversum, Pays Bas.
<i>TalkBank</i>	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* (Etats Unis) et *BABS*, l'archive numérique de la Bibliothèque de l'Etat de Bavière - évoquent la bibliothéconomie (library sciences) 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 *phonotheque* 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.

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 des résultats des chercheurs d'un établissement ou organisme ; 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 et la présence éventuelle de données issues des SIC, il faut aller sur les sites et vérifier directement sur les serveurs.

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 communication (cf. tableau 1) se trouvent aux Etats-Unis, les autres se répartissent sur l'Allemagne, l'Australie, le Canada, l'Inde, le Japon, les Pays Bas et l'Union Européenne. Aucun ne se trouve en France. Les trois entrepôts limitrophes mentionnés plus haut se trouvent aux Etats 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*, *MédiHAL* ou *beQuali*. A ceci on pourrait ajouter la phonotheque 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 niveau 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* notamment héberge plusieurs communautés dans les domaines SIC, avec toutefois peu de jeux de données à ce jour.

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 ? Une analyse des données de recherche dans les thèses de doctorat en SHS avait révélé qu'il n'y avait pas de « données spécifiques SIC » mais (1) qu'il y avait un profil caractéristique composé du côté des données primaires (sources) d'enquêtes, de corpus textuels et de données web et, du côté des données secondaires (résultats), de textes, de tableaux, de graphiques (schémas) et de statistiques, et (2) qu'il y avait 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 n'est pas possible 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 faisable, grâce aux outils qui aujourd'hui encore sont en bêta test, comme *Google Dataset Search*¹⁵.

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 facilite leur accessibilité.¹⁶

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 plateforme open data de l'Union Européenne).

¹⁵ <https://toolbox.google.com/datasetsearch>

¹⁶ Cf. la description sur le site de Force 11 <https://www.force11.org/group/fairgroup/fairprinciples>

- RDF Data Cube Vocabulary pour les métadonnées des données multidimensionnelles sur le Web.
- ISO 19115, la seule norme au sens stricte, pour la description des données géographiques.
- 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 limitrophes¹⁷.

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, et ils utilisent les identifiants standards (DOI etc.).

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

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 plateformes d'information scientifique (cf. Schöpfel & Müller 2014 pour les archives ouvertes), et puis certains pour les entrepôts de données, dont la norme ISO 16919 (Systèmes de transfert des informations et données spatiales), la certification du International Science

¹⁷ <http://www.dcc.ac.uk/resources/metadata-standards>

Council World Data System (ICSU-WDS)¹⁸, le Data Seal of Approval (DSA)¹⁹, ou plus récemment le CoreTrustSeal²⁰.

Trois des 14 entrepôts en sciences de la communication sont des plateformes 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)²¹ 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 plateforme *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 de l'expression de la diversité méthodologique et thématique 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.

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 & Schöpfel 2015) en a fourni une image détaillée, composée d'une importante variation de pratiques mais aussi, de certaines caractéristiques prépondérantes. A 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.

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

¹⁸ <https://www.icsu-wds.org/services/certification>

¹⁹ <https://www.datasealofapproval.org/en/>

²⁰ <https://www.coretrustseal.org/about/>

²¹ <https://www.clarin.eu/content/assessment-procedure>

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 préoccupations 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 plateformes pour la gestion, la conservation et la diffusion des données.

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. A 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 sciences) 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. A 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. 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 partie importante des efforts publics porte sur le développement et l'interconnexion des sites, plateformes et infrastructures de données, au niveau national, européen et international. Le Plan national pour une science ouverte préconise la mise en place de dispositifs thématiques et disciplinaires et pose la question d'une plateforme générique, tandis que l'Union Européenne soutient l'interopérabilité et l'interconnexion des infrastructures, au sein d'une 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.
- Création (ou fédération) d’une communauté SIC sur les plateformes 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. A 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, comme 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’est pas là, et peut-être il suffit 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.

Références

F. André (2015). ‘Déluge des données de la recherche ?’ In L. Calderan, P. Laurent, H. Lowinger, and J. Millet (coord.) *Big data : nouvelles partitions de l’information*, pp. 77-95. De Boeck, Louvain-la-Neuve.

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

- R. Aleixandre-Benavent, et al. (2016). 'Correlation between impact factor and public availability of published research data in Information Science and Library Science journals'. *Scientometrics* 107(1):1-13. <https://link.springer.com/article/10.1007%2Fs11192-016-1868-7>
- P. Bonfils & M. Durampart (2013). 'Environnements immersifs et dispositifs numériques. Etudes expérimentales et approches distanciées'. *ESSACHESS - Journal for Communication Studies* 6(1). <http://www.essachess.com/index.php/jcs/article/view/196>
- C. L. Borgman (2015). *Big data, little data, no data: scholarship in the networked world*. The MIT Press, Cambridge MA.
- A. Borrego & F. Garcia (2013). 'Provision of supplementary materials in librard information science scholarly journals'. *Aslib Proceedings* 65(5):503-514. <https://www.emeraldinsight.com/doi/abs/10.1108/AP-10-2012-0083>
- S. Chignard (2012). *Open data : comprendre l'ouverture des données publiques*. Fyp éd., Limoges.
- M. Kindling, et al. (2017). 'The Landscape of Research Data Repositories in 2015: A re3data Analysis'. *D-Lib Magazine* 23(3/4). <http://www.dlib.org/dlib/march17/kindling/03kindling.html>
- T. Koltay (2016). 'Digital research data'. In D. Baker and W. Evans (coord.). *Digital Information Strategies*, pp. 71-84. Chandos Publishing, Oxford.
- V. Larroche (2018). *Le dispositif : un concept pour les sciences de l'information et de la communication*. ISTE Editions, London.
- B. Matthews, M. D. Wilson, & K. Kleese Van Dam (2002). 'Accessing the outputs of scientific projects'. In *CRIS2002: 6th International Conference on Current Research Information Systems*, Kassel, August 29-31, 2002. <http://dspacecris.eurocris.org/handle/11366/149>
- M. Pain (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. https://memsic.ccsd.cnrs.fr/mem_01374509
- H. Pampel, et al. (2013). 'Making Research Data Repositories Visible: The re3data.org Registry'. *PLoS ONE* 8(11):e78080+. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078080>
- H. Prost & J. Schöpfel (2015). 'Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. Rapport final'. Université de Lille 3, Villeneuve d'Ascq. <http://hal.univ-lille3.fr/hal-01198379v1>
- V. Rebouillat (2017). 'Inventory of Research Data Management Services in France'. In L. Chan & F. Loizides (eds.), *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices*, pp. 174-181. IOS Press, Amsterdam. <http://ebooks.iospress.nl/publication/46651>
- A. Rege (2015). 'Retour sur l'enquête sur les pratiques de publication scientifique et de production de données de la recherche'. In: *Université de Strasbourg, Projet AOC, COPIL 27 mars 2015*.
- N. Reymonet (2017). *Améliorer l'exposition des données de la recherche : la publication de data papers*. Université Paris Diderot. https://archivesic.ccsd.cnrs.fr/sic_01427978
- RIN (2008). *Stewardship of digital research data: a framework of principles and guidelines*. Research Information Network, London. <http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines>

J. Schöpfel (2018). `Vers une culture de la donnée en SHS. Une étude à l'Université de Lille'. Université de Lille, Villeneuve d'Ascq. <https://hal.archives-ouvertes.fr/GERIICO/hal-01846849v1>

J. Schöpfel & U. Müller (2014). `Evaluer la qualité des archives ouvertes : Le certificat DINI'. *Partnership* 9(1):1-21. <https://journal.lib.uoguelph.ca/index.php/perj/article/view/2733/3226>

J. Schöpfel, et al. (2015). `Dissertations and Data (Keynote Address)'. In GL17 International Conference on Grey Literature, 1-2 December 2015, Amsterdam. <https://hal.archives-ouvertes.fr/hal-01285304>

J. Schöpfel, H. Prost & V. Rebouillat (2016). `Research data in current research information systems'. In: *CRIS 2016: 13th International Conference on Current Research Information Systems*, 8-11 June 2016, St Andrews. <http://dspacecris.eurocris.org/handle/11366/501>

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

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

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

M. D. Wilkinson, et al. (2016). `The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific Data* 3:160018+. <https://www.nature.com/articles/sdata201618>

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.