Revue COSSI

N° 5 | 2018
Processus de normalisation et durabilité de l'information

Hors norme ? Une approche normative des données de la recherche

Joachim SCHÖPFEL

Édition électronique :
URL :
https://revue-cossi.numerev.com/articles/revue-5/77-hors-norme-une-approche-normative-des-donnees-de-la-recherche
DOI : numerev_1609
Date de publication : 30/11/2018

CertiScience© *Certifié évalué par les pairs*

Pour **citer cette publication** : SCHÖPFEL, J. (2018) Hors norme ? Une approche normative des données de la recherche. *Revue COSSI*, (5). https://doi.org/10.34745/numerev_1609

Nous proposons une réflexion sur le rôle des normes et standards dans la gestion des données de la recherche, dans l'environnement de la politique de la science ouverte. A partir d'une définition générale des données de la recherche, nous analysons la place et la fonction des normes et standards dans les différentes dimensions du concept des données. En particulier, nous nous intéressons à trois aspects faisant le lien entre le processus scientifique, l'environnement réglementaire et les données de la recherche : les protocoles éthiques, les systèmes d'information recherche et les plans de gestion des données. A l'échelle internationale, nous décrivons l'effet normatif des principes FAIR qui, par la mobilisation d'autres normes et standards, créent une sorte de « cascade de standards » autour des plateformes et entrepôts, avec un impact direct sur les pratiques scientifiques.

Mots-clés :
Science ouverte, Normes, Données de la recherche, Gestion des données, Standards, Principes FAIR

Abstract : We propose a reflection on the role of norms and standards in the management of research data in the open science policy environment. Starting from a general definition of research data, we analyze the place and function of norms and standards in the different dimensions of the concept of data. In particular, we focus on three aspects that link the scientific process, the regulatory environment, and research data: ethical protocols, research information systems, and data management plans. At the international level, we describe the normative effect of the FAIR principles, which, through the mobilization of other norms and standards, create a sort of "cascade of standards" around platforms and repositories, with a direct impact on scientific practices.

Keywords : research data, data management, norms, standards, FAIR principles, open science

INTRODUCTION

Le plan national pour la science ouverte, présenté début juillet 2018, a confirmé l'ambition de l'Etat français d'accélérer l'ouverture des données issues de la recherche

publique (MESRI 2018). Les mesures et actions annoncées se traduiront d'une manière ou d'une autre en normes, c'est-à-dire en référentiels incontestables et communs, tels que des lois, directives, normes industrielles, recommandations ou bonnes pratiques.

Ces nouvelles initiatives et prescriptions s'ajouteront, à différents niveaux de normalisation, aux normes qui sont déjà en place et qui organisent et contrôlent l'écosystème des données de la recherche.

L'article propose une vision globale de l'aspect normatif des données de la recherche, en mettant l'accent sur la France et sur les sciences humaines et sociales. Il s'appuie sur l'analyse de documents officiels, d'articles et de rapports, de communications, discours, sites d'initiatives et de projets, publiés plus particulièrement en France (Ministère, CNRS, etc.) et en Europe (Commission), sans pour autant négliger le paysage international (DataCite, RDA, etc.). Le discours sur la science ouverte donne parfois l'impression qu'il s'agit d'un paysage émergent, d'une sorte de no man's land ou plutôt d'un territoire vierge, à l'instar du *far west* de la conquête du continent américain. Or, il n'en est rien. La gestion des données de la recherche et leur partage ne constituent nullement un sujet « hors norme » mais sont strictement encadrés par des normes de toute sorte – trop pour les uns, pas assez pour les autres.

Notre compréhension du concept des normes est pragmatique, partant de l'importance centrale des normes dans la société contemporaine (Perriault & Vaguer 2011) et considérant les normes comme un « document établi par consensus et approuvé par un organisme reconnu, qui fournit, pour des usages communs et répétés, des règles, des lignes directrices ou des caractéristiques, pour des activités ou leurs résultats, garantissant un niveau d'ordre optimal dans un contexte donné » (Cacaly 1997).

L'omniprésence des normes

L'organisation de l'accès libre aux données scientifiques fait partie des objectifs de la recherche publique de la France (Code de la Recherche, article L112-1 alinéa e). La volonté d'ouvrir les données de la recherche a été confirmé par le plan d'action national 2018-2020 *Pour une action publique transparente et collaborative* dont l'engagement 18 vise à construire un écosystème de la science ouverte dans lequel « 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, p.57).

Le plan national pour la science ouverte a confirmé cette ambition (MESRI 2018). L'objectif est que les données produites par la recherche publique soient progressivement structurées en conformité avec les principes FAIR, préservées et, quand cela est possible, ouvertes. Les mesures annoncées incluent l'obligation d'une diffusion ouverte des données issues de programmes financés sur fonds publics, la généralisation des plans de gestion de données dans les appels à projets, et l'engagement d'un processus de certification des infrastructures de données.

D’un point de vue normatif, il s’agit d’un processus de plusieurs étapes, dont l’objectif est de créer un nouvel écosystème fonctionnel et cohérent, par l’interprétation des normes existantes, par leur ajustement et par leur remplacement, avec quelques acteurs-clés (administration centrale, organismes de recherche, établissements de l’enseignement supérieur, agences de financement, éditeurs, industrie de l’information…).

Qu’il s’agisse de la sécurité des systèmes et des données, de leur traitement, communication ou partage ou bien, de leur description et conservation pérenne, chaque aspect de la gestion et chaque phase du cycle de la vie des données de la recherche fait l’objet de multiples normes, à caractère légal, éthique ou industriel.

L’expérience utilisateur

Les chercheurs eux-mêmes connaissent généralement assez bien les normes, surtout en ce qui concerne les référentiels d’expertise de leur propre champ d’action, puisque l’efficacité et la qualité de leur travail dépend de la connaissance et de l’application de ces normes. Les enquêtes de terrain montrent des pratiques hétérogènes, plus ou moins formalisées, plus ou moins efficaces et adaptées, dépendant pour beaucoup des équipements et méthodes, des thématiques et disciplines, et des compétences scientifiques (Prost & Schöpfel 2015, Serres et al. 2017, Schöpfel et al. 2018). Les normes sont généralement ressenties comme contraintes, surtout pour des questions de sécurité et de protection des données à caractère personnel, où l’offre institutionnelle en matière de système d’information n’est pas toujours à la hauteur des besoins des chercheurs (Schöpfel 2018).

Aussi, malgré leur tradition séculaire de partage des résultats de leur recherche, une partie des chercheurs considère la politique d’ouverture des données comme une contrainte externe, où des finalités politiques et industrielles restent sans lien direct avec leurs objectifs et besoins au quotidien (Kaden 2018).

QUATRE *DIMENSIONS* NORMATIVES

« Les données de recherche en SHS ne se laissent pas aisément définir et saisir » (Serres et al. 2017). Souvent, les définitions des données en général et des données de la recherche en particulier sont implicites, faites d’anecdotes, de *success stories*, de descriptions, de listes, d’aspects technologiques, de tendances et d’impact sur les organismes et la société en général. Il y a dix ans, l’OCDE a tenté une définition qui est devenue *de facto* une référence pour les études dans ce domaine : selon l’OCDE, les données de la recherche sont « des enregistrements factuels (chiffres, textes, images et sons), utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche » (OCDE 2006). A regarder de près, on peut distinguer quatre dimensions constitutives au concept des données de la recherche (Schöpfel et al. 2017, cf. figure 1) :

- La nature factuelle des données et leur grande diversité.
- Le caractère communautaire des données.
- La finalité des données et leur lien avec le processus de la recherche.
- Le lien des données avec leur traitement documentaire (enregistrement).

Figure 1. Les quatre dimensions du concept des données de la recherche

Ainsi, nous nous proposons d'analyser la place et la fonction des normes et standards dans les différentes dimensions du concept des données, en particulier. Pour chacune des dimensions, nous dégagerons la tension entre plusieurs niveaux de standardisation, avant tout entre une approche générique au sein des organismes et grandes infrastructures[1], et l'insistance des communautés disciplinaires sur la spécificité des instruments et procédures, avec leurs propres normes et standards, mais également avec une revendication assumée d'être différent, « hors norme ».

La nature factuelle

Il n'y a pas de normes ou de standards proprement dits pour décrire la diversité des données de la recherche. En revanche, plusieurs typologies font office de référentiels *de facto*, plus ou moins reconnus, utilisés et acceptés. La définition de l'OCDE mentionne quatre types de données à titre d'exemples et comme modèles : chiffres, textes, images et sons, sans intention d'être exhaustif.

Christine Borgman (2012) insiste sur la distinction fondamentale entre les données collectées comme ressources primaires de la recherche (« inputs ») et celles produites comme résultats de la recherche (« outputs »). A partir d'une proposition du JISC, Francis André (2015) suggère une classification synthétique des données en fonction des finalités et procédures, de leur génération, davantage liée aux méthodes et outils de la recherche scientifique qu'aux disciplines et thématiques :

- Les données d'observation ;
- les données d'expérimentation ;
- les données de simulation ;
- les données dérivées ;
- les données de référence.

Si ces deux approches typologiques sont régulièrement citées pour structurer l'analyse et la compréhension du domaine des données de la recherche, deux autres systèmes de classification ont un certain impact, auprès notamment des professionnels et administratifs des services et infrastructures, dans la mesure où ils permettent de décrire et d'indexer d'une manière standardisée les dispositifs des données, l'un – re3data[2] – à l'échelle internationale, l'autre – Cat-OPIDoR[3] – en France.

Le répertoire re3data fait la différence entre treize catégories de données, suivant un méthodologie plutôt empirique que systématique. La répartition des dispositifs par rapport à cette typologie est très inégale. Plusieurs catégories sont très larges, transversales aux disciplines, aux contours mal définis, telles que « raw data » (données brutes) ou « archived data » (données archivées). Une catégorie « other » correspond à une longue traîne d'autres types de données dans un tiers des entrepôts de données. Tout cela ressemble davantage à un référentiel ad hoc, en cours de constitution et d'évolution, qu'un référentiel réfléchi, suivant une approche systématique et exhaustive.

Cat-OPIDoR de son côté propose une typologie de services et de fonctions, mais pas de classification des données, juste une indexation libre qui ne pourra pas servir de référentiel.

Les deux outils n'ont pas de caractère normatif au sens prescriptif, du « ce qui doit être ». En revanche, ils aident, sur un terrain nouveau et dynamique, à déterminer le périmètre et quelques éléments de l'état habituel, de la majorité des cas, ce qui est un autre aspect d'une norme.

Sans lien avec ces approches génériques, certains types de données ont fait l'objet d'une normalisation proprement dite. Dans le domaine des SHS, il s'agit avant tout des données à caractère personnel et des données de la santé. Des lois et réglementations déterminent leur nature et encadrent leur finalité, les obligations, droits et traitements autorisés, avec un fort degré de normalisation et de contrainte, aussi bien au plan national (Informatique et Liberté) qu'à l'échelle internationale (RGPD).

Le lien avec la communauté

Le lien étroit avec une communauté est une deuxième dimension importante du concept des données de la recherche. Il faut comprendre communauté au sens large, aussi bien autour des disciplines et thématiques qu'autour des équipements, procédures, méthodes et outils. Du point de vue normatif, cette approche ouvre la perspective sur un univers de recommandations, réglementations, bonnes pratiques et autres standards spécifiques, propres à tel ou tel groupe ou structure scientifique. Quelques exemples :

Équipement : des normes et protocoles pour la production de données recueillies grâce à des équipements importants, comme des IRM, accélérateurs de particules, observatoires etc.

Méthodologie : des recommandations, bonnes pratiques et règles pour la réalisation d'enquêtes, pour l'exploitation de statistiques, pour la collecte de données à l'aide d'observations etc.

Discipline : certaines disciplines se sont dotées de règles plus ou moins élaborées pour la collecte et le traitement de leurs données, comme l'archéologie ou les sciences

médicales.

Les normes communautaires peuvent aussi s'exprimer à travers les critères de publication des résultats. Par exemple, l'*American Psychological Association* (APA), la plus importante société savante en psychologie, demande désormais systématiquement l'accessibilité des données de la recherche pour les articles publiés dans les revues APA, créant ainsi une nouvelle situation de « ce qui doit être », ce qui impacte directement les pratiques des chercheurs.

A l'instar de l'APA, il est certain que la politique en faveur de la science ouverte incitera d'autres sociétés savantes, associations scientifiques, réseaux, laboratoires et instituts à formuler des recommandations (*guidelines*) pour une plus grande transparence, pour faciliter la réplicabilité des études et le partage des résultats. Ceci concernera donc la collecte, la production et la gestion des données de la recherche, par le biais de dispositifs (sites) labellisés, d'une description standard (cf. plus loin) et de règles partagées par tous.

Parmi ces règles, citons surtout la déontologie scientifique développée et maintenue par les chercheurs eux-mêmes, dont l'intégrité, le respect de la personne, les conflits d'intérêt… Nous y reviendrons.

La question d'une différence normative entre les SHS et les STM a été soulevée. Dans l'absolu, il n'y a pas de différence ; en réalité il existe bien une différence de taille, dans la mesure où les STM travaillent davantage avec des grands équipements et infrastructures, qui mobilisent un ensemble important de normes et standards industriels. Par exemple, il y a un réel débat autour du développement des standards concernant les sources, l'interopérabilité et l'échange de données dans la recherche clinique (Hammond 2012).

La finalité

Les données de la recherche n'ont pas d'existence en tant que telles. Elles remplissent des fonctions précises dans un écosystème constitué de politique, d'économie et de science. Ainsi, sous l'aspect fonctionnel, on peut distinguer trois grands domaines, chacun avec ses propres normes. Voici quelques finalités énoncées par l'Union Européenne, des organismes de recherche etc. :

- Finalités politiques : rendre le système politique plus transparent, rendre l'action publique plus efficace, et créer un environnement favorable pour la création de valeur par l'économie.
- Finalités économiques : optimiser la recherche, accélérer l'innovation.
- Finalités scientifiques : faciliter l'exploration et la fouille des résultats scientifiques, pouvoir comparer, répliquer et vérifier les résultats, valider des hypothèses etc.

Quels sont les éléments normatifs qui relèvent de ces trois domaines ? Citons les plus importants :

Politique : La loi numérique de 2016 (données publiques, ouverture des données, fouille de données) ; le plan national de 2018 (structuration des données, ouverture des données, développement des dispositifs) ; d'autres textes concernant les données publiques (car les données scientifiques produites sur des fonds publics ont dans la majorité des cas vocation aÌ devenir publiques et que les données publiques ont vocation aÌ devenir scientifiques lorsqu'elles concernent l'environnement, le climat, la santé, l'aménagement du territoire).

Économique : les textes, conventions, etc. préconisent l'interopérabilité des infrastructures et dispositifs des données de la recherche.

Scientifique : les documents relatifs à une politique institutionnelle dans le domaine des données de la recherche (déclarations, votes, décisions, statuts etc.), les programmes de financement incluant des conditions en lien avec les données de la recherche (H2020, Wellcome Trust, ANR etc.).

D'une manière plus fine, on peut aussi rapprocher les normes de la modélisation du processus de la recherche comme un « cycle de vie des données » (figure 2).

Figure 2. Cycle de vie des données (source : inist.fr)

Sous cet aspect, on découvrira un nombre important de normes, standards et recommandations de toutes sorte, tous en lien avec un ou plusieurs aspects de traitement des données de la recherche. Quelques exemples :

- Les normes de sécurité pour l'archivage numérique des données.
- Les normes de l'archivage.
- Le protocole d'échange de données NF ISO 20614 (cadre de transactions pour données et métadonnées).
- Le standard COUNTER avec la norme SUSHI (NISO) pour la production et la communication des données d'usage des bases de données et autres ressources numériques.
- Les référentiels pour la production des données liés à certains équipements.
- Les recommandations pour l'analyse des données (probabilités statistiques etc.).
- Les incitations au partage.

Il ne faut pas imaginer cet écosystème établi selon différents domaines et finalités comme un univers cohérent et fonctionnel. Du point de vue normatif, force est de constater des degrés de normalisation très différents, d'une directive européenne via une loi nationale vers des normes industrielles, des déclarations politiques et des simples incitations, et force est également de relever des incohérences et contradictions, avec des injonctions opposées : par exemple entre l'ouverture des

données et une valorisation économique des résultats de la recherche. Les incohérences rendent le système instable et dynamique.

L’enregistrement

Du point de vue documentaire, ce quatrième aspect fonctionnel est le mieux maîtrisé. Il s’agit de la fonction d’enregistrement, au sens d’Henry Oldenburg (*registration*) ou, en termes de la gestion des données, de la fonction de curation (Neuroth et al. 2013). A titre d’illustration, on peut évoquer trois niveaux de normalisation :

- La normalisation des métadonnées (cf. Qin 2013) : les standards de métadonnées génériques (Dublin Core), les standards de métadonnées pour les données de la recherche (DataCite Metadata Schema), les standards de métadonnées pour certains types de données (Data Documentation Initiative)…
- La normalisation des identifiants : les identifiants génériques (DOI), les identifiants pour les personnes (ORCID, ResearcherID…) et institutions (OrgID…), les identifiants à caractère national (IdHAL, les référentiels d’IdRef).
- La normalisation de la nomenclature de certains types d’objets scientifiques (Nomenclature of Celestial Objects du CDS…).

Si le DOI fait l’objet d’une norme industrielle internationale (ISO), d’autres formats ou identifiants sont compatibles avec une norme (comme ORCID) ou sont largement reconnus et acceptés comme un standard *de facto*, comme par exemple le format développé et maintenu par DataCite. Par ailleurs, DataCite annonce un *code of practice* pour l’attribution des DOI, à l’instar de la démarche du projet COUNTER, ce qui renforcerait encore le caractère standard et contournable de l’identifiant DOI pour les données de la recherche.

Sous l’aspect normatif, retenons trois autres éléments : certaines initiatives s’adressent directement aux chercheurs (ORCID, ResearcherID…), d’autres davantage aux professionnels de l’information et/ou des données (DOI, métadonnées…) ; les initiatives sont plus ou moins capables de faire le grand écart entre une approche générique, important pour l’interopérabilité des systèmes, et la spécificité disciplinaire ou communautaire ; la force normative de ces initiatives est liée au soutien industriel et/ou politique.

LA FORCE DES NORMES

Qu'est-ce qui fait la force des normes ? S’agit-il du risque d’une sanction, d’un caractère obligatoire, d’une autre forme de contrainte ? Il est certain qu’il faut discerner divers degrés, voire diverses natures de force, sans confondre force obligatoire et force contraignante, caractère impératif, incitatif ou inspiratoire (au sens de Thibierge 2009). Regardons à titre d’exemple trois domaines faisant le lien entre le processus scientifique, l’environnement réglementaire et les données de la recherche : le plan de gestion, le protocole éthique, et le système d’information recherche.

Le plan de gestion

En tant que tel, un plan de gestion des données de la recherche fait partie des bonnes pratiques d'un chercheur, de la même manière qu'un plan de gestion d'un projet avec ses *work packages*, ses échéances et livrables. Sous cet angle, son caractère normatif serait de l'ordre incitatif ou inspiratoire, basé sur des recommandations, modèles et formations, mis à disposition par les établissements et organismes, à l'instar des Universités Diderot ou Descartes, du JISC (avec DMPonline) et du CNRS (avec DMP-OPIDoR).

La réalité est plus compliquée dans la mesure où ces plans ont fait leur apparition avant tout avec les cahiers des charges du programme européen H2020, d'abord dans une version relativement simple (cinq questions) et comme une option volontaire, désormais dans une version conforme aux principes FAIR (cf. plus loin) et, du moins pour une partie du programme, avec un caractère obligatoire.

Rédiger un plan de gestion est donc nécessaire pour déposer une demande de financement et obtenir un budget de recherche pour certains domaines du programme H2020. C'est une contrainte forte qui, d'après les documents de travail pour le 9^{e} programme cadre de l'UE, se généralisera sans doute à l'avenir pour l'ensemble des domaines scientifiques et appels à projets de la Commission Européenne.

Néanmoins, à ce jour, aucune sanction ne semble prévue ou applicable en cas de non-respect du plan de gestion déposé. La CE insiste sur deux mises à jour du plan : une version intermédiaire à mi-parcours, l'autre à la fin, présentant le bilan du projet. Mais apparemment il n'y a pas de contrôle ou d'audit systématique ou ponctuel pour vérifier le traitement des données.

Quant aux programmes scientifiques en France, le plan d'action pour la science ouverte prévoit que les appels à projets de l'ANR contiennent à l'avenir l'obligation d'un plan de gestion. On connaîtra dans quelques mois, fin 2018 ou 2019, la nature exacte sera cette obligation, si elle sera accompagnée de contrôles etc.

Le protocole éthique

Protocole éthique et plan de gestion ont des liens organiques. Chaque protocole éthique contient un chapitre consacré à la définition, à la collecte, au traitement et à la conservation des données en question. De son côté, le plan de gestion pose systématiquement la question d'une éventuelle dimension éthique qui serait à développer, le cas échéant. Cependant, la force normative du protocole éthique est bien plus complexe que celle du plan de gestion, pour deux raisons.

D'une part, l'aspect éthique des données de la recherche touche des domaines très différents, comme les données personnelles, le respect des personnes et les conflits d'intérêt, mais également la sécurité, la propriété intellectuelle et la crédibilité des données (cf. Jacquemin et al. 2018).

D’autre part, ces différents domaines font appel à des normes de nature très diverse. Citons par exemple la politique scientifique des organismes de recherche (cf. les travaux et avis du COMETS du CNRS) ou les recommandations, consignes et bonnes pratiques des établissements et communautés scientifiques. Citons également les cahiers des charges des programmes de recherche (cf. la politique en matière d'éthique et d'intégrité scientifique adoptée par le conseil d’administration de l’ANR en 2014 ; cf. les *ethics reviews* du programme H2020 ; cf le Code de Nuremberg pour la recherche biomédicale). Les normes ou labels industriels (pour la sécurité des dispositifs), les lois et directives sur les données personnelles, la santé publique et la bioéthique sont d’autres exemples.

Il existe donc ici toute la gamme des contraintes, allant de l’obligation légale à l’incitation plus ou moins forte, avec le risque de sanction en cas de non-respect, par la jurisprudence, par les institutions et/ou par d’autres communautés scientifiques (sociétés savantes, associations, comités de rédaction etc.).

Le système d’information recherche

Un système d’information recherche sert entre autre à l’évaluation du fonctionnement et de la performance des équipes, structures, établissements et organismes scientifiques. Considérant les données de la recherche comme résultats du travail scientifique, il paraît normal qu’elles fassent partie des éléments évalués, tout comme les publications ou les brevets.

Les critères d’évaluation sont imposés par les agences de financement et d’évaluation; ils présentent le risque d’un impact négatif sur l’allocation des ressources humaines et financières en cas de non-conformité. Ces critères sont décidés et communiqués par les agences et trouvent leur reflet dans le modèle et le format des systèmes d’information recherche.

En France, pour l’instant, les données de la recherche ne font pas partie des critères évalués lors des campagnes du HCERES. D’autres pays les ont déjà ajoutées au catalogue des domaines à évaluer; les formats (dont CERIF, l’unique format standardisé dans ce domaine, maintenu par euroCRIS), les produits des éditeurs et d’autres sociétés (comme le système PURE d’Elsevier ou Converis de Clarivate) sont capables de les évaluer.

Or, quels sont ces critères ? L’analyse des dispositifs, formats et agences d’évaluation montre qu’il s’agit avant tout des critères de gestion, en conformité notamment avec les principes FAIR. En revanche, l’évaluation ne s’intéresse pas ou peu à la volumétrie ou à la qualité des données, contrairement par exemple à l’évaluation des publications (Schöpfel et al. 2016).

A ce jour, la force normative d’une telle évaluation est donc liée au caractère obligatoire de l’évaluation scientifique, imposée par la loi ou un arrêté, et par la légitimité de l’organisme désigné (en France, le HCERES) pour déterminer ses propres critères

d’évaluation.

Mais cette force normative s’appuie également sur la standardisation de facto des critères et formats d’évaluation des systèmes d’information recherche, et par ce biais, sur des normes de type industriel, même s’ils n’ont pas encore faire l’objet d’une norme ISO, AFNOR, DIN ou NISO.

LA GESTION *FAIR* DES DONNÉES DE LA RECHERCHE

Les principes FAIR ont été évoqués plusieurs fois déjà, notamment en lien avec le plan de gestion et avec l’évaluation. En fait, sur le terrain, la politique d’ouverture s’accompagne d’une forte incitation à mettre en œuvre des bonnes pratiques scientifiques compatibles avec certains principes définis au niveau européens comme “FAIR Guiding Principles” de la gestion et du pilotage des données de la recherche (Wilkinson et al. 2016, European Commission 2016, Mons et al. 2017) : les données doivent être faciles à (re)trouver, accessibles et si possible ouvertes, interopérables et réutilisables. Ces principes visent avant tout à faciliter le traitement automatique des données par des machines.

La normalisation joue un rôle important dans la démarche des principes FAIR. D’après la littérature et les documents officiels sur l’application de ces principes, on peut distinguer plusieurs axes prioritaires :

Findable

Pour faciliter la recherche des données, il faut privilégier des standards (normes) à trois niveaux :

- Identifiants standards (DOI, URI, handle…)
- Formats standards pour les métadonnées (DataCite Metadata Schema), enrichis de formats standards disciplinaires
- Mécanismes d’interrogation standards (API, SPARQL, SQL etc.)

Accessible

Pour garantir l’accessibilité des données de la recherche, les principes FAIR préconisent trois démarches normalisées :

- Utilisation de protocoles d’accès standardisés (http, API REST)
- Dépôt des données dans un entrepôt certifié (*Core Trust Seal*)
- Indexation avec des métadonnées disciplinaires standardisées

A regarder de près, l’exigence d’une certification implique l’application d’autres normes et standards, par rapport à la sécurité, à la pérennité, à la protection des données sensibles, etc.

Interoperable

Pour augmenter l'interopérabilité des dispositifs et infrastructures des données, la normalisation est requise à deux niveaux :

- Utilisation d'un langage formel et standard pour la représentation des connaissances, comme par exemple Web Ontology Language
- Utilisation de terminologies largement partagées et reconnues comme standards disciplinaires

Reusable

Finalement, pour faciliter la réutilisation des jeux de données, l'approche FAIR proposent trois démarches en lien avec la normalisation :

- Mise à disposition selon une licence ouverte explicite, accessible et reconnue (Creative Commons ou autres)
- Indication claire de la provenance des données, avec des identifiants et codes normalisés (auteurs, organismes, pays…)
- Utilisation d'un format de métadonnées standard, avec une indexation riche

Par le biais des programmes européens, des infrastructures et initiatives internationales et des plans ou projets nationaux, les principes FAIR sont en passe de devenir des standards *de facto*, avec une cascade de normes et de standards mobilisés. Soutenus non seulement par des organismes comme EUDAT, DataCite et RDA, portés par l'initiative GO FAIR à laquelle contribue activement la France, mais aussi mis en avant par le programme H2020, les principes FAIR s'imposent comme l'unique cadre de référence reconnu d'une stratégie de données de la recherche, que ce soit pour un pays, un organisme public ou un éditeur commercial.

Evaluer l'impact à terme de cette dynamique sur le fonctionnement et les pratiques au sein des équipes et laboratoires parait difficile, aussi bien que d'anticiper l'impact sur la diffusion des publications et sur l'évolution à venir de la science ouverte. Même si la politique en fait aujourd'hui une sorte de fer de lance pour accélérer le développement de la science ouverte, il faut garder en tête qu'il s'agit d'abord d'un ensemble de principes et de règles pour interconnecter des machines, une condition nécessaire pour l'exploitation massive du *big data* de la recherche. Autrement dit, le moteur et principal intéressé d'une telle démarche est l'industrie de l'information, avec les grandes infrastructures, qui a besoin de ces données comme « fuel of economy » (Neelie Kroes, EC) pour créer de la valeur. C'est peut-être la raison de leur succès.

CONCLUSION

Nous avons essayé de mettre en lumière les normes et standards mobilisés par l'écosystème émergent des données de la recherche en pleine effervescence. Nous avons opté pour une analyse à partir du concept des données de la recherche, avec sa diversité, ses fonctions, ses services et sa gestion, et avec ses liens communautaires. Le résultat de cette analyse est un ensemble normatif complexe, hétérogène et parfois,

sous tension, dans un cadre politique et réglementaire qui s'impose au sens de la « force normative des faits » (Georg Jellinek), même en dehors d'une construction normative cohérente. L'écosystème de la science ouverte peut paraître inédit, il n'est nullement « hors norme », ni dans son ensemble, ni au niveau des communautés, applications et équipements particuliers.

Pour aller plus loin, une analyse systématique des normes et standards pourrait aboutir à une matrice à trois dimensions :

1. La nature des normes
    1. Des normes légales
    2. Des normes techniques
    3. Des normes éthiques
2. Le degré de contrainte des normes
    1. Des lois
    2. Des normes industrielles
    3. Des standards reconnus et acceptés
    4. Des contrats
    5. Des recommandations
    6. Des bonnes pratiques
3. Les acteurs mobilisés
    1. Politique, scientifique
    2. Public, privé
    3. Infrastructures, industrie
    4. Métiers…

L'analyse devrait inclure une étude approfondie de la stratégie normative de la Research Data Alliance, qui regroupe les principaux acteurs concernés et qui travaille entre autre avec la NISO sur la protection des données personnelles (privacy), sur la normalisation des plans de gestion et sur l'interopérabilité légale des services de données (cf. aussi NISO 2018). Une telle approche permettra de comprendre comment ce nouvel écosystème de la science ouverte est en train de se structurer et de s'organiser, et quel en sont les véritables enjeux.

L'analyse des normes dans leur écosystème apporterait peut-être aussi des réponses aux questions relative à la durabilité des normes et standards dans un environnement en mutation constante et rapide. Quel sera l'impact des traditions et pratiques scientifiques ? Sont-elles de « simples verrous » sur la voie de l'industrialisation de la recherche (certains commencent à parler de l'Uberisation) ? Il est peut-être trop tôt pour donner une réponse mais il n'est certainement pas trop tard pour poser la question.

Remerciements

Une partie des travaux à l'origine de cette communication 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.

RÉFÉRENCES

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*. *Actes du Séminaire IST Inria, octobre 2014*, pages 77-95. De Boeck; ADBS, Louvain-la-Neuve.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059-1078. doi:10.1002/asi.22634

Cacaly, S. (dir.) (1997). Dictionnaire encyclopédique de l'information et de la documentation. Nathan, Paris.

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.

European Commission (2016). *H2020 programme. Guidelines on FAIR data management in Horizon 2020*.Version 3.0. European Commission Directorate-General for Research & Innovation, Brussels.

Hammond, W. E. (2012). Standards development and the future of research data sources, interoperability, and exchange. In Richesson, R. L. and Andrews, J. E. (dir.), *Clinical Research Informatics, Health Informatics*, pages 335-365. Springer, London. doi:10.1007/978-1-84882-448-5_18

Jacquemin, B., Schöpfel, J., Kergosien, E., & Chaudiron, S. (2018). L'éthique des données de la recherche en SHS. In *DocSoc2018, 6e conférence "Document numérique & Société", Information-Communication : le recours à l'éthique en contexte numérique*, Institut de la Communication et des Médias (Echirolles), 27 et 28 septembre 2018.

Kaden, B. (2018). Warum Forschungsdaten nicht publiziert werden. *LIBREAS*, 33. urn:nbn:de:kobv:11-110-18452/20046-8

Latif, A., Limani, F., & Tochtermann, K. (2018). Research data management for long tail research data - a generic research data infrastructure approach. In *CODATA 2018*, March 19, 2018, Göttingen.

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

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the

European open science cloud. *Information Services & Use*, 37(1):49-56. doi:10.3233/isu-170824

Neuroth, H., Strathmann, S., Oßwald, A., & Ludwig, J. (dir.) (2013). *Digital curation of research data. Experiences of a baseline study in Germany.* vwh, Glückstadt.

NISO (2018). *Strategic directions 2018.* National Information Standards Organization, Baltimore MD.

OCDE (2006). *Recommendation of the council concerning access to research data from public funding.* 14 December 2006 - C(2006)184, Organisation for Economic Co-Operation and Development, Paris.

Prost, H. & Schöpfel, J. (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.

Perriault, J. & Vaguer, C. (dir.) (2011). *La norme numérique. Savoir en ligne et Internet*. CNRS Editions, Paris.

Qin, J. (2013). Infrastructure, standards, and policies for research data management. In *2013 COINFO 8th International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, 25-27 October 2013, Nanjing, China.

Schöpfel, J. (2018). *Vers une culture de la donnée en SHS. Une étude à l'Université de Lille.* Université de Lille, Villeneuve d'Ascq.

Schöpfel, J., Ferrant, C., André, F., & Fabre, R. (2018). Research data management in the French national research center (CNRS). *Data Technologies and Applications*, 52(2):248-265. doi:10.1108/DTA-01-2017-0005

Schöpfel, J., Kergosien, E., & Prost, H. (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 (France).

Schöpfel, J., Prost, H., & Rebouillat, V. (2016). Research data in current research information systems. In *CRIS 2016,* St Andrews, 8-11 June 2016.

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.

Thibierge, C. (dir.) (2009). *La force normative. La naissance d'un concept*. Librairie générale de droit et de jurisprudence, Bruylant, Paris.

Timmermann, M. (2018). A framework for research data management. In *CODATA 2018*, March 19, 2018, Göttingen.

Wilkinson, M. D., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:sdata201618+.

---

[1] Cf. les travaux de CODATA, par exemple Latif et al. (2018) et Timmermann (2018)

[2] https://www.re3data.org/

[3] https://cat.opidor.fr