“ Pour commencer, pourriez-vous définir ’données de la recherche’ ? ” Une tentative de réponse

Joachim Schöpfel, Eric Kergosien, Hélène Prost

« Pour commencer, pourriez-vous définir ‘données de la recherche’ ? »

Une tentative de réponse

Joachim Schöpfel[1], Eric Kergosien[1], Hélène Prost[2]

1. Laboratoire GERiiCO, Université de Lille
prenom.nom@univ-lille3.fr

2. CNRS, INIST, Nancy, membre associé du laboratoire GERiiCO

RESUME. *Le projet D4Humanities s’inscrit dans le champ des Humanités numériques – comment permettre l’exploration des données de la recherche en SHS (corpus textuels ou oraux, données brutes, images…) avec des techniques numériques (text and data mining, cartographie, visualisation…) afin de construire un sens nouveau ? Il s’inscrit dans la continuité des travaux du laboratoire GERiiCO et de ses partenaires à l’Université de Lille Sciences Humaines et Sociales (SCD, ED SHS, ANRT…) avec comme objectif d’accélérer la démarche des données de la recherche notamment par rapport aux doctorants et jeunes chercheurs, et de faciliter le montage d’un projet de recherche international. En particulier, le projet contient trois volets : (1) Pratiques et besoins dans le domaine des données de la recherche (enquête qualitative des comportements, attitudes, motivations et besoins par rapport à la gestion et au partage des données de la recherche) ; (2) workflow pour le dépôt des données des doctorants en SHS (dépôt, préservation et diffusion des données via le service NAKALA de la TGIR Huma-Num) ; (3) recherche sur les données et les thèses (concept et typologie des données en SHS ; évolution des contenus, formats, structures et prescriptions des thèses dans l’environnement de l’Open Science). Le projet sera mené avec l’ISN Oldenburg et d’autres partenaires étrangers ; il facilitera la création d’un consortium et le montage d’un projet de recherche dans les Humanités numériques sur les thèses de doctorat de l’avenir, avec un financement européen (H2020) ou franco-allemand (ANR/DFG). Cette communication présente les grandes lignes de l’étude sur les données de l’axe 3, c’est-à-dire l’analyse du concept de données de la recherche, pour mieux cerner l’identification (granularité), pour mieux comprendre la distinction et les relations entre données primaires et secondaires et pour affiner la catégorisation des données en SHS. L’accent est mis sur une triple approche, conceptuelle, typologique et fonctionnelle.*

MOTS-CLES : *Données de la recherche, humanités numériques, open science, sciences humaines et sociales.*

1. Introduction

Les données de la recherche intriguent. Tout le monde en parle. Tout le monde est d'accord sur l'importance du sujet et sur la nécessité « de faire quelque chose » et « de se positionner ». Open science oblige. Mais se positionner par rapport à quoi ? Faire quelque chose, oui, mais avec quoi ? Autant de certitudes sur l'action et la politique, autant d'approximations sur le concept même des données de la recherche. D'ailleurs, faut-il dire « les données de la recherche » ou « la donnée de la recherche » ? En d'autres termes, existe-t-il une définition unique et exhaustive permettant de faire le lien entre tous les objets désignés par les chercheurs, informaticiens et bibliothécaires comme « données de la recherche » ? Ou s'agit-il d'un concept multiple, aux contours flous, une sorte de plus petit dénominateur commun avec un grand volume de données en formats multiples et flux continu qui n'ont en commun, outre les « 3V » du Big Data (volume, variété et vélocité, cf. Cointot & Eychenne 2014 et Davenport 2014) que le fait de constituer un « vaste océan d'opportunités » ?

Dans le cadre du projet *D4Humanities* nous poursuivons l'analyse du concept de données de la recherche, pour mieux cerner l'identification (granularité), pour mieux comprendre la distinction et les relations entre données primaires et secondaires et pour affiner la catégorisation des données en SHS. Concrètement, cette étude est menée entre avril et décembre 2017, par une équipe scientifique de GERiiCO en partenariat avec plusieurs initiatives et organismes allemands et néerlandais (cf. plus loin). Il s'agit en premier lieu d'une analyse de dispositifs et d'outils existants, et également d'exploiter les résultats d'autres recherches, y compris par une nouvelle analyse de nos propres enquêtes de 2015 sur les données dans les thèses (Schöpfel et al. 2015) et sur les pratiques et attentes des chercheurs de l'Université de Lille SHS (Prost & Schöpfel 2015).

Dans nos séminaires, colloques et enquêtes, nous sommes régulièrement interpellés sur ce point : « Pour commencer, pourriez-vous définir ce que veut dire 'données de la recherche' ? » Une bonne question : pouvons-nous les définir ? Que savons-nous de ces objets à la fois scientifiques, politiques et économiques que sont les données ? Souvent, nous répondons avec Borgman et al. (2012) que les données, c'est un concept difficile à définir à cause de leur grande variété, qu'elles fussent physiques ou numériques. Mais est-ce vraiment satisfaisant ?

Selon Chignard (2012, p.10), « tout est donnée ». Dans l'introduction de son livre *Big data, little data, no data*, C. Borgman (2015) constate qu'il est impossible de s'accorder sur une seule définition, en particulier en sciences humaines et sociales. Certains guides, manuels et présentations omettent tout simplement de définir leur objet, se limitant à énumérer quelques exemples pour délimiter le périmètre. Or, ce genre d'absence de consensus sur la définition de concepts-clés est caractéristique pour des domaines émergents (De Mauro et al. 2016) mais ne suffit

pas ni pour la curation des données, ni pour la recherche, ni pour la mise en œuvre d'une politique raisonnée d'information scientifique et technique.

2. Quelques définitions

Par rapport aux publications sur le Big Data, De Mauro et al. (2016) évoquent un concept aussi populaire que nébuleux et un état de l'art chaotique, avec beaucoup de définitions « implicites » faites d'anecdotes, de success stories, de descriptions, d'aspects technologiques, de tendances et d'impact sur les organisations et la société. Sur la base d'une analyse de 1437 articles et communications, De Mauro et al. proposent une définition consensuelle : « Big Data is the Information asset characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value ».

L'idée du « information asset » se retrouve chez Borgman (2015) qui d'une manière générale définit les données de la recherche comme « inputs, outputs, and assets of scholarship » (p.4). Les données comme « capital » ou « biens » de la recherche, à exploiter, analyser, travailler pour en révéler ou extraire une valeur scientifique ; un « avoir » en amont ou à la base de l'information. C'est une conception rejoignant l'approche de Chignard (2012) qui voit la donnée comme « un fait brut, qui n'est pas – encore – interprété » (p.10) ; un matériel à l'état brut que l'on peut (doit) manipuler, traiter, analyser et interpréter.

L'approche du gouvernement américain va dans le même sens quand il définit les données de la recherche comme « matériel », c'est-à-dire comme « the recorded factual material commonly accepted in the scientific community as necessary to validate research findings » (OMB Circular 110[1]). En français, dans la traduction de Pain (2016, p.17) : « Les données de la recherche sont des enregistrements factuels (chiffres, textes, images et sons) utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires à la validation des résultats de recherche. » En fait, cette définition est particulièrement intéressante dans la mesure où elle enrichit le concept des données comme matériel brut dans quatre directions :

- l'enregistrement comme préalable – d'une certaine façon, il s'agit de l'équivalent de la fixation matérielle d'une idée sur un support comme préalable de la reconnaissance et protection d'une œuvre de l'esprit ;
- la nature factuelle, malgré la grande diversité des données – mais comme la Royal Society (2012) précise, il peut s'agir d'une nature factuelle « supposée » quand elle définit les données de la recherche comme « qualitative or quantitative statements or numbers that are (or assumed to be) factual » ;

[11] https://www.whitehouse.gov/omb/circulars_a110#36

- le lien avec la communauté scientifique – la définition de l'OMB conteste l'idée d'un concept absolu et établit le lien avec les chercheurs eux-mêmes, au sens d'un minimum de pratiques, valeurs, méthodes, outils et concepts partagés ; en d'autres termes, une donnée est ce qui est acceptée par un groupe de chercheur (« communauté ») comme telle (consensus) ;
- et la finalité – les données de la recherche ont une fonction, jouent un rôle dans le processus scientifique, en particulier pour la validation des hypothèses et résultats.

Le lien avec la communauté (ou plutôt les communautés) peut être fondé par un cadre conceptuel, une thématique ou discipline ; souvent il s'appuiera avant tout sur un instrument, une procédure ou une méthodologie. Nous y reviendrons dans la section sur les approches classificatoires.

Le quatrième critère, la finalité, introduit une dimension fonctionnelle, et nous y reviendrons également un peu plus loin. Ici, soulignons un autre aspect, l'intégration organique des données dans le processus de recherche. Cet aspect ressort très clairement dans l'adaptation de la définition américaine par Reymonet (2017) : « Les données de la recherche sont un ensemble d'informations factuelles enregistrées sur des supports, produites ou collectées, selon divers procédés au cours d'un processus de recherche » (p.1). Borgman (2015) ne dit pas autre chose quand elle décrit les données de la recherche comme « inputs (and) outputs ». Cette approche accentue le caractère dynamique du concept ; avec les mots d'André (2015) : « la donnée scientifique est un objet dont les caractéristiques évoluent selon l'étape du processus de recherche auquel on s'intéresse (…) un objet complexe, dynamique, vivant » (p.82-83).

Les deux aspects, le lien avec la communauté aussi bien que la finalité, mettent en question toute tentative de définir les données de la recherche dans l'absolu, comme objets avec des caractéristiques propres, sans relation avec le contexte du travail scientifique. Il s'agit d'un concept relatif, impossible à comprendre hors contexte. Cette relativité ou contextualité explique sans doute une partie des problèmes de l'évaluation des données – en fait, les systèmes de 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).

La diversité (« variété ») est un aspect essentiel du concept du Big Data. Cette diversité se traduit entre autres comme un mélange d'éléments plus ou moins structurés : « Le big data est la combinaison des informations structurées des bases de données avec des informations semi-structurées de logs et des informations non structurées » (Cointot & Eychenne 2014, p.221). Cet aspect se retrouve dans la conception des données de la recherche à deux niveaux – dans la description d'un ensemble de manifestations quantitatives et qualitatives (cf. Royal Society 2012) et dans les différentes ébauches d'une arborescence ou hiérarchie de données.

Cette granularité pose un problème particulier pour la définition mais aussi pour la curation. Certains catalogues de métadonnées fournissent un niveau et une

spécificité assez bas pour décrire les différents aspects des données et des ensembles de données (Elbaek et al. 2010). Le projet-pilote de Rutherford-Appleton (Matthews et al. 2002) propose un modèle de collection de données avec deux niveaux et trois éléments différents (figure 1).

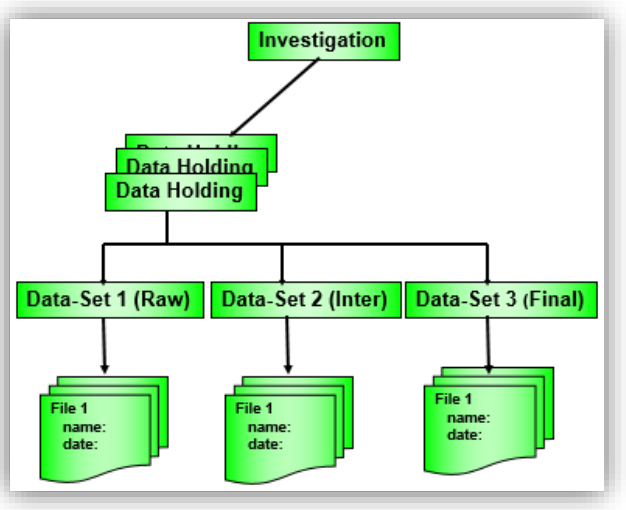

Figure 1. Arborescence d'une collection de données (Matthews et al. 2002)

Selon ce modèle, chaque collection de données prend la forme d'une hiérarchie ; une enquête génère une séquence d'ensembles de données logiques et chaque ensemble de données est instancié via un ensemble de fichiers numériques. Ces fichiers de données peuvent être des données brutes, des données intermédiaires ou finales. Ils sont liés par des métadonnées, mais ont des adresses différentes et peuvent être situés sur différents serveurs. Leur limite globale dépend de l'enquête initiale. Pourtant, les données ne sont pas des publications, leurs exigences ne sont pas identiques à celles des publications, elles n'ont pas toujours de limites nettes, et leurs métadonnées doivent être en mesure de supporter les changements.

Le *European Plate Observing System* (EPOS) définit un autre modèle de données avec quatre niveaux de données, compatible avec le projet *C4D* (Bailo & Jeffery 2014) :

- Niveau 0 : données brutes, ou données de base (ex. sismogramme) ;
- Niveau 1 : produits de données issus de procédures presque automatisées (ex. localisation des tremblements de terre) ;
- Niveau 2 : produits de données issus des recherches scientifiques (ex. modèles de la croûte terrestre) ;

- Niveau 3 : produits de données intégrés issus d'analyses complexes ou partagées (ex. cartes de risque).

Ces quatre niveaux dépendent explicitement du champ d'investigation (sciences de l'univers, sismologie) et le lien avec la communauté scientifique et ses instruments et méthodes est évident. Mais l'arborescence est exemplaire pour deux raisons – les quatre niveaux peuvent être transposés à d'autres domaines, comme une sorte de modèle hiérarchique indépendant des domaines ; et ils posent clairement la question de la définition du concept et de son identification, et ceci d'une manière très concrète, dans la mesure où le problème de l'attribution d'un identifiant unique et pérenne (DOI ou autre) ne trouve pas de solution – faut-il attribuer un DOI à une donnée de niveau 1 ? Ou plutôt de niveau 2 ? Quid des ensembles et produits de données plus complexes (cf. CODATA 2013) ?

Ces questions n'ont que l'apparence technique ; en réalité elles touchent au concept même des données de la recherche, en particulier en ce qui concerne leur relation avec la communauté et ses instruments, méthodes et conventions. Cette relation peut-être assez complexe dans un environnement multi-, inter- ou transdisciplinaire : « What constitutes data is determined by a given community of interest that produces the data. However, an investigator may be part of multiple, overlapping communities of interest, each of which may have different notions of what are data » (Koltay 2016, p.73).

Nous retrouverons ce lien dans la section suivante, cette fois-ci sous l'angle des approches typologiques ou classificatoires. Mais essayons une première schématisation d'une définition des données de la recherche qui devrait inclure quatre éléments (figure 2).

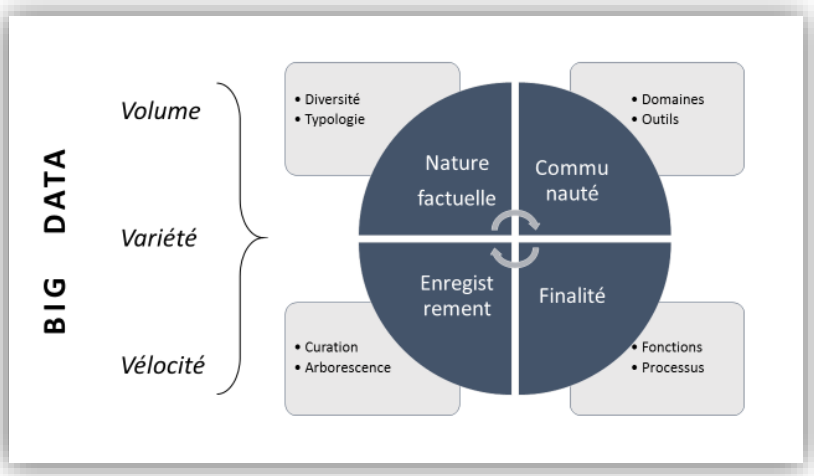

Figure 2. Éléments d'une définition des données de la recherche

3. L'approche typologique

Comme les données de la recherche sont difficiles à conceptualiser, leur définition passe souvent par la description d'exemples : « Data are most often defined by example, such as facts, numbers, letters and symbols » (Borgman 2015, p.19). Des faits et symboles, des chiffres et des lettres – ce que Borgman et ses collègues citent comme exemples de données, d'autres auteurs transforment en typologies, avec des approches classificatoires plus ou moins sophistiquées.

Ainsi, André a adapté une classification du *Research Information Network* en cinq larges catégories transversales (2015, pp.81-82) :

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

Ces différents types de données sont davantage liés aux méthodes et outils de la recherche scientifique qu'aux disciplines et thématiques. Cependant, plus les approches classificatoires sont détaillées et spécifiques, plus elles révèlent le caractère disciplinaire de certains types de données. Là où le *Research Information Network* a tenté une synthèse en fonction des finalités et procédures de leur génération (RIN 2008), le répertoire international *re3data* propose quatorze types de données, à partir de l'indexation des bientôt 2000 entrepôts de données référencés (Kindling et al. 2017, cf. figure 3).

L'indexation par *re3data* révèle deux caractéristiques des données de la recherche :

- Une répartition très inégale, avec plusieurs larges catégories, transversales aux disciplines, présentes dans une grande partie des domaines scientifiques, aux contours mal définis ; comme les *Scientific and statistical data formats*, les *Standard office documents*, le *Plain text*, les *Images* ou encore les *Raw data ;*
- Une longue traîne d'autres types de données présents dans ces entrepôts. D'après la catégorie *Other*, plus d'un tiers des sites indexés contient d'autres types de données, en dehors des quatorze catégories de la typologie ; ou bien sans correspondance avec la définition de ces catégories.
- Le rapprochement avec les disciplines montre que certains types de données sont surreprésentés, comme *Standard office documents*, *Plain text* et *Images* en SHS (Kindling et al. 2017). Mais avec une taille d'effet relativement faible, la réalité et la portée de cette observation restent à démontrer.

	Count (n)	Percentage (%)
Scientific and statistical data formats	1152	63%
Standard office documents	1088	59%
Plain text	903	49%
Images	895	49%
Raw data	809	44%
Structured graphics	697	38%
Structured text	585	32%
Archived data	425	23%
Audiovisual data	339	18%
Software applications	324	18%
Databases	313	17%
Networkbased data	112	6%
Source code	81	4%
Configuration data	43	2%
Other	668	36%
Total	1837	100%

Figure 3 : Types de données dans le répertoire re3data (N=1837 sites, 4 avril 2017)

Nos propres études sur le campus de l'Université de Lille 3 confirment une typologie propre aux sciences humaines et sociales mais illustrent surtout l'intérêt d'une distinction entre données primaires (sources) et secondaires (résultats), avec une classification différente (figure 4).

	re3data	Prost & Schöpfel 2015, sources	Prost & Schöpfel 2015, résultats
Scientific and statistical data formats	63%	26%	49%
Standard office documents	59%		
Plain text	49%	64%	76%
Images	49%	25%	21%
Raw data	44%		
Structured graphics	38%		32%
Structured text	32%		
Archived data	23%	34%	
Audiovisual data	18%	6%	44%
Software applications	18%		9%
Databases	17%		37%
Networkbased data	6%		
Source code	4%		
Configuration data	2%		
Enquêtes et entretiens		47%	
Observations		41%	
Expériences		36%	
Cartes et plans			10%
Other	36%	7%	3%
Total	100%	100%	100%

Figure 4 : Classification des données primaires et secondaires en SHS (en %)

La correspondance entre nos catégories et la typologie de re3data est imparfaite. Les différences sont certainement dues à la finalité (description du contenu des entrepôts de données vs. étude des pratiques des chercheurs) et l'approche méthodologique (indexation vs. enquête).

La correspondance entre ces catégories et les disciplines est forte, sans que l'on puisse parler de données spécifiques aux disciplines. Les analyses attestent davantage de profils disciplinaires pour les différents types de données, voire de profils de données pour certaines disciplines (cf. Schöpfel et al. 2015).

4. L'approche fonctionnelle

Nous avons évoqué plus haut un triple désir scientifique, politique et économique. Quelles sont les fonctions associées ? Et comment peuvent être définies les données de la recherche selon ces finalités ?

Concernant la production scientifique, les données ont un rôle central quelque soit les disciplines. De plus en plus, les chercheurs ont besoin de disposer de grandes masses de données, à côté de données de dimensions plus modestes, pour explorer, visualiser et comparer et/ou vérifier des résultats, valider des hypothèses : « the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. » (OMB Circular 110)[2], voir en formuler de nouvelles. Au regard du cycle de vie des données, les données de la recherche peuvent être catégorisées de la façon suivante : « les données préliminaires ou données préparatoires (exclues dans un contexte de diffusion) ; les données brutes (données acquises lors du processus de recherche et déjà potentiellement traitées) ; et enfin les données traitées et analysées : c'est-à-dire ayant subi une transformation telle qu'il n'est plus possible d'accéder aux données brutes, un graphique par exemple » (Pain 2016, p.18) (cf. figure 5).

Pour répondre à l'objectif d'accès aux données de la recherche, des accords successifs ont vu le jour. A la suite de plusieurs mouvements de chercheurs visant à l'ouverture de corpus de données scientifiques, la déclaration internationale sur le libre accès de Budapest (Budapest Open Access Initiative) propose le 14 février 2002 une première définition de l'Open data[3]. Aujourd'hui le terme d'Open data est réservé à l'ouverture des données obtenues sur fonds publics en général. De nombreuses autres initiatives suivent, comme la Déclaration de Berlin de 2003 sur le libre accès à la connaissance dans toutes les sciences. En février 2012, le texte de la Déclaration de Berlin avait été signé par plus de 360 universités ou assimilées. Pour les chercheurs, cette mouvance de l'open data a deux objectifs principaux, la réutilisation des données de la recherche et la possibilité de garantir la qualité scientifique d'une hypothèse en la justifiant par les données qui en sont à la source.

[2] https://www.whitehouse.gov/omb/circulars_a110#36
[3] http://www.budapestopenaccessinitiative.org/

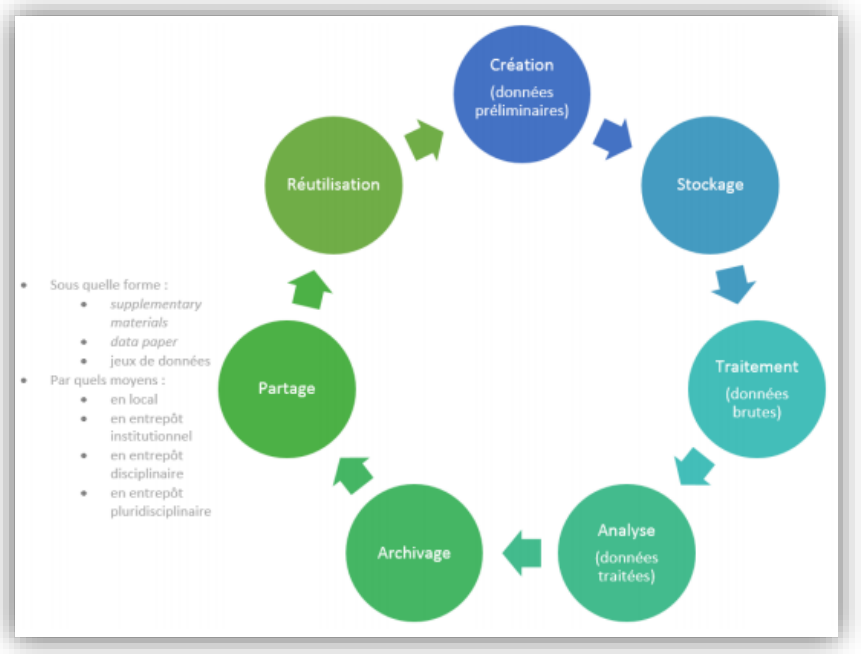

Figure 5 : Cycle de vie des données de la recherche (Pain 2016, p.18)

Au niveau politique, les fonctions de l'ouverture des données de la recherche diffèrent sensiblement. Davantage de données doivent être partagées gratuitement, en formats ouverts, avec liberté de réutilisation. En effet, les Etats sont d'importants pourvoyeurs de données produites, reproduites, collectées, diffusées ou rediffusées par les administrations publiques dans le cadre de leurs missions institutionnelles. Il s'agit notamment de données démographiques, géographiques, météorologiques, économiques, financières, culturelles, touristiques, etc., qui visent à assurer la qualité et la continuité du service public. Initié sur la base de la politique de l'Union Européenne en matière d'open science, l'effort d'ouverture des données vise trois grands objectifs :

- renforcer la démocratie (transparence, concertation, implication des citoyens) ;
- soutenir l'innovation économique et sociale, favoriser l'émergence d'un environnement propice à la croissance économique ;
- rendre l'action publique plus efficace (décloisonnement et adoption de stratégies fondées sur la donnée), notamment pour l'aide à la prise de

décision pour l'évaluation et le bon fonctionnement des services (Chignard 2012).

Au regard de ces finalités, les politiques valorisant l'ouverture des données publiques n'ont pas les mêmes objectifs que celles du partage des données scientifiques. Dans ce sens, il serait utile de différencier données scientifiques et données publiques, comme le précise le comité éthique du CNRS (COMETS)[4]. Les données scientifiques produites sur des fonds publics ont dans la majorité des cas vocation à devenir publiques. Les données publiques ont vocation à devenir scientifiques lorsqu'elles concernent l'environnement, le climat, la santé ou encore l'aménagement du territoire. Parmi les nombreux exemples existants, les données disponibles sur le portail de la métropole européenne de Lille (MEL)[5], notamment les données géolocalisées, sont utilisées par les chercheurs du projet BiblioMEL pour l'analyse des pratiques des citoyens dans et à l'extérieur des bibliothèques[6]. A noter qu'en France, c'est le service Etalab qui gère l'Open data public sous l'autorité du Premier ministre, avec pour mission de communiquer les données subventionnées sur fonds public avec mise à disposition libre et (quasi) gratuite.

Sur le point de vue économique, pour de nombreux services de l'Etat, ainsi que pour le secteur privé, les données créées dans le cadre d'activités ont une forte valeur lorsqu'elles sont ensuite régulièrement réutilisées, comme le précisent les auteurs de l'étude de l'US National Research Council, intitulée *Bits of Power*[7] : « La valeur des données réside dans leur exploitation. L'accès total et ouvert aux données scientifiques devrait devenir la norme internationale pour l'échange des données scientifiques issues de la recherche financée sur fonds publics. » Pour les organismes de recherche mais surtout pour les autorités politiques nationales et européennes, l'ouverture des données scientifiques a, du point de vue économique, deux finalités majeures :

- optimiser la recherche (réduction de la recherche redondante, nouvelles recherches sans collecte de données etc.) ;
- accélerer l'innovation industrielle notamment dans le domaine de la santé (traitement et prévention contre Zika ou Ebola etc.) et de l'environnement.

Une autre fonction est liée au concept du Big Data : « l'association dans une même analyse de données variées pour en déduire des informations que l'on n'était pas en mesure de trouver avec les analyses classiques de données structurées (…) souvent pour prendre une action en temps réel » (Cointot & Eychenne 2014, p.221). En d'autres mots, la gestion et mise à disposition des données scientifiques peuvent faciliter l'émergence de nouvelles formes d'analyses, avec des résultats novateurs,

[4] http://www.cnrs.fr/comets/IMG/pdf/2015-05_avis-comets-partage-donnees-scientifiques-2.pdf
[5] https://opendata.lillemetropole.fr/page/home/
[6] https://bibliomel.hypotheses.org
[7] https://www.nap.edu/read/5504/chapter/1

du simple fait de leur masse (volume) et leur diversité (variété), peut-être aussi (mais pas nécessairement) leur interopérabilité.

5. Conclusion

La suite de notre étude sera l'intégration de ces trois approches en un seul modèle, avec une définition qui non seulement fera le lien entre concepts, typologies et fonctions mais déterminera aussi les limites et questions ouvertes, en particulier dans le domaine des sciences humaines et sociales. Cette étude sera menée entre avril et décembre 2017 en partenariat avec l'Université Humboldt de Berlin (projet *eDissPlus*[8]), le Karlsruhe Institut of Technology (projet *bwDataDiss*[9]), la Bielefeld University (projet *CONQUAIRE*) et le service DANS aux Pays Bas. Il s'agit en premier lieu d'une analyse de dispositifs et d'outils existants et d'exploiter les résultats d'autres recherches, y compris par une nouvelle analyse de nos propres résultats de 2015 (données dans les thèses, enquête sur le campus de l'Université de Lille SHS). L'analyse s'appuiera notamment sur les 480 archives avec des données SHS répertoriées par *re3data*[10] et sur les travaux de CODATA[11], DataCite[12], DARIAH[13] et CESSDA[14]. L'objectif est de conceptualiser plus finement le terme de données de la recherche dans le champ des SHS, par une définition aussi bien que par une description des principaux types et niveaux de données et une problématisation de la distinction entre données primaires et secondaires. L'enjeu d'une telle définition est double : contribuer à une meilleure compréhension des données de la recherche, et contribuer à une meilleure prise en charge, aussi bien au niveau des chercheurs et leurs équipes qu'au niveau des laboratoires, établissements et organismes.

Bibliographie

André F. (2015). Déluge des données de la recherche ? In L. Calderan, P. Laurent, H. Lowinger, and J. Millet (Eds.), *Big data : nouvelles partitions de l'information. Actes du Séminaire IST Inria, octobre 2014*, pp. 77-95. Louvain-la-Neuve: De Boeck; ADBS.

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

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

[8] https://www2.hu-berlin.de/edissplus/
[9] https://bwdatadiss.kit.edu/
[10] http://www.re3data.org/
[11] http://www.codata.org/
[12] https://www.datacite.org/
[13] https://datacite.hypotheses.org/
[14] http://cessda.net/

Borgman C. L., Wallis J. C., and Mayernik M. S. (2012). Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work (CSCW)* 21 (6), 485-523. https://doi.org/10.1007/s10606-012-9169-z

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

CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal 12*, CIDCR1-CIDCR75.

Cointot J.-C. and Eychenne Y. (2014). *La révolution Big Data : les données au cœur de la transformation de l'entreprise*. Paris: Dunod.

Davenport T. H. (2014). *Stratégie Big Data*. Paris: Pearson.

De Mauro A., Greco M., and Grimaldi M. (2016). A formal definition of big data based on its essential features. *Library Review* 65 (3), 122-135. http://dx.doi.org/10.1108/LR-06-2015-0061

Elbæk M. K., Sandfær M., and Simons E. (2010). CRIS/OAR interoperability workshop. In *CRIS 2010: 10th International Conference on Current Research Information Systems, 2-5* June 2010, Aalborg.

Kindling, M., H. Pampel, S. van de Sandt, J. Rücknagel, P. Vierkant, G. Kloska, M. Witt, P. Schirmbacher, R. Bertelmann, and F. Scholze (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

Koltay T. (2016). Digital research data. In D. Baker and W. Evans (Eds.), *Digital Information Strategies*, pp. 71-84. Oxford: Chandos Publishing.

Matthews, B., M. D. Wilson, and 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

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

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. http://hal.univ-lille3.fr/hal-01198379v1

Reymonet N. (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

Royal Society (2012). *Science as an open enterprise. Summary report.* The Royal Society Science Policy Centre, London, https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe-summary.pdf

Schöpfel J., Prost H., and Malleret C. (2015). Making data in PhD dissertations reusable for research. In *8th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 21 October 2015, Prague. http://hal.univ-lille3.fr/hal-01248979

Schöpfel J., Prost H., and Rebouillat V., 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