h e g

Haute école de gestion
Genève

Swiss Data Journal : Etude de conceptualisation d'un modèle de partage des données de recherche basé sur le concept de Data Paper dans le contexte helvétique

Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES

par :

Stéphane Zwahlen

Conseiller au travail de Bachelor :

Professeur Docteur René Schneider

Genève, le 30 juin 2016

Haute École de Gestion de Genève (HEG-GE)

Filière information documentaire

Hes·so GENÈVE
Haute Ecole Spécialisée
de Suisse occidentale

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre de spécialiste en information documentaire HES.

L'étudiant atteste que son travail a été vérifié par un logiciel de détection de plagiat.

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie ».

Genève, le 30 juin 2016, Stéphane Zwahlen

Remerciements

Je désire adresser ici mes remerciements pour les personnes ayant permis à ce travail de Bachelor d'être réalisé, notamment le corps enseignant de la Haute école de gestion.

Des pensées plus spécifiques sont adressées à :

- Mme Anne-Christine Maillard, pour ses encouragements et son soutien lors de mon questionnement sur la reprise d'une formation.
- Mme Valentine Costa, pour son soutien et sa confiance indéfectibles lors des épreuves traversées. Sans elle tout cela n'aurait pas été possible.
- Mme Lucie Sandoz pour sa relecture attentive.
- Mme Eliane Blumer, M. Jean-Blaise Claivaz et M. Patrick Ruch, pour leur disponibilité et pour les réponses pertinentes qu'ils ont apportées lors des entretiens réalisés.
- M. René Schneider, pour son soutien, sa compréhension et son encadrement.
- M. Michel Gorin et M. Alexandre Boder, pour leur compréhension et leur soutien.
- M. Mathieu Lourdin et Mme Laura George, pour avoir été les compagnons de ma route estudiantine.

Résumé

L'ouverture des données de la recherche financées par des fonds publics est au cœur de l'actualité récente, tant au niveau européen qu'à l'échelle helvétique, en raison du message fort porté par la présidence néerlandaise de l'Union européenne, véritable manifeste en faveur de l'accès libre aux résultats de recherche, s'appuyant sur des considérations politiques, économiques et éthiques. Cet engagement a entraîné dans son sillage une réaction similaire du Fond national suisse pour la recherche scientifique.

Les acteurs suisses de la recherche, qu'ils soient en charge de la promotion de la science et de l'innovation, institutions académiques, bibliothèques ou chercheurs tentent d'apporter des réponses aux enjeux soulevés par la production toujours plus importante de volume de données et l'évolution des exigences des agences de financement des partenaires européens et internationaux quant à la gestion des données.

Dans ce cadre fertile, le rédacteur de ce travail, a été amené à réaliser une étude de conceptualisation sur la mise en place, à l'échelle helvétique, d'une solution de diffusion basée sur le concept de data paper. Ce sont de véritables publications savantes, dont le sujet n'est plus les résultats de l'étude, mais les données produites dans le processus de recherche ayant mené à sa réalisation, afin de pouvoir déterminer si ce type de solution était envisageable et quels en seraient les éléments essentiels.

En résulte trois scénarios, qui ont été construits sur la base d'une étude contextuelle prenant en compte les besoins des principaux acteurs du domaine et les notions fondamentales de la gestion des données de recherche et de leur publication. Mais également plus spécifiquement sur le data paper et son influence sur les processus de communication scientifique.

Le rédacteur a considéré, une fois les éléments nécessaires acquis, qu'une solution de diffusion basée uniquement sur le type de publication sélectionnée au départ pouvait apporter une plus-value intéressante, mais ne constituait que l'une des facettes de la publication des données de la recherche.

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

> *« L'histoire de la science est une réplique de celle de la communication des savoirs et est basée sur la reproductibilité des données. Si une théorie ne peut être vérifiée, elle est alors remise en question. Les hypothèses qui ne sont pas testables avec des données vérifiables, ne sont pas scientifiques, mais se muent en un système de croyance sans fondement ».*[1] *(Heidorn 2008, p. 286)*

Comme l'illustre Heidorn, les données de recherche (ci-après DR) sont le moteur indispensable qui anime le fonctionnement académique et scientifique, en permettant de vérifier et valider les hypothèses qu'elles ont permis d'établir. Elles disposent d'un rôle primordial dans une société centrée sur l'information.

Si leur gestion est un sujet de réflexion pour les sciences depuis quelques années, son importance s'est accrue de manière très importante au cours de la période 2015 - 2016. En effet, les récentes prises de positions des acteurs européens et helvétiques de l'environnement scientifique, en faveur du libre accès des résultats de recherche (articles scientifique et DR), constituent les prémices d'une évolution structurelle des pratiques et usages des acteurs impliqués.

Le point d'entrée du rédacteur de ce travail, le poster : Data papers as a catalyst for Open Access and the long tail of research data, (Schneider & Prongué 2016), présente une infographie visant à illustrer le fonctionnement des data papers[2], et de leurs potentiels usages. Combiné avec un regard approfondi sur l'actualité et l'évolution du domaine scientifique, il ouvre la perspective de réaliser une analyse de faisabilité pour l'utilisation de ce nouveau type de publication, en qualité de modèle de diffusion pour les DR produites par les chercheurs affiliés aux institutions suisses.

---

[1] Traduction par le rédacteur de : « The history of science is a reflection of the history of communication of knowledge and is based on reproductible data. If evidence for a theory cannot take the form of replicable data, then the theory is in question. […] Theories that are not testable with verfiable data are not thories but rather unsubtantied belief systems ».

[2] Véritables productions scientifiques centrées sur la description des DR, ces articles feront l'objet d'une définition dans le cadre de ce travail en partie 6.1

1.1 Objectif

Cette étude de conceptualisation a pour objectif de fournir des scénarios de diffusion des DR pour le contexte helvétique, basées sur le concept de data paper. Il s'agit pour le rédacteur de proposer des solutions permettant la dynamisation et l'harmonisation du cycle de vie des données de la recherche par des propositions visant à développer une forme de data journal.[3]

1.2 Méthodologie

Ce travail est réalisé en analysant tout d'abord le contexte actuel, en tenant compte des influences politiques, économiques, technologiques, sociétales et légales qui constituent le terreau des développements récents dans la perception des DR.

Afin de disposer d'une meilleure compréhension du fonctionnement du domaine de la recherche helvétique, sont identifiés les principaux acteurs concernés pas la gestion des DR.

Puis sont définis, dans une première phase, l'ensemble des éléments essentiels à la compréhension de cette problématique :

- Les DR en elles-mêmes, leurs caractéristiques et les informations descriptives indissociables qui les accompagnent.
- Leur partage fait l'objet d'une analyse, afin de disposer des informations nécessaires à la compréhension des différents modèles de publications.
- Les data papers eux-mêmes, afin d'identifier les éléments nécessaires à la mise en place d'une solution de partage des DR centrée sur ceux-ci.

Au moyen des informations recueillies et du résultat des entretiens semi-directifs menés lors de la phase préparatoire de cette étude, sont alors construits les scénarios.

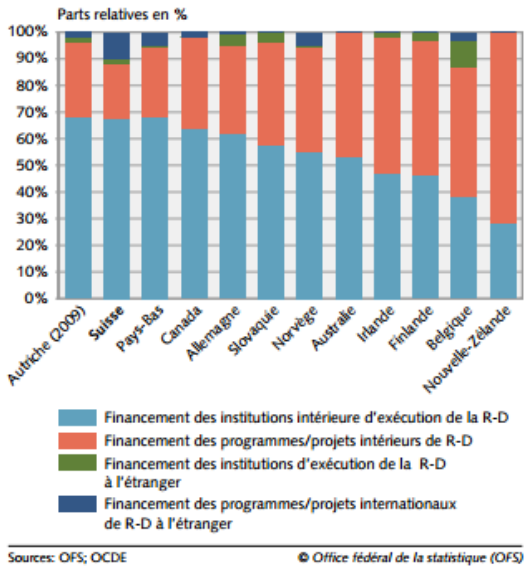
[3] Journal exclusivement composé de data papers.

2. Contextes

Afin de définir le contexte dans lequel s'inscrit cette étude, le rédacteur a opté pour une analyse de l'environnement politique, technologique, sociétal, légal et économique de l'écosystème de la recherche helvétique.[4]

Celle-ci repose essentiellement sur des interactions internes au territoire suisse, comme l'indique le rapport de l'office fédéral de la statistique : Financement public de la recherche en Suisse : Evolution 2000 – 2010. En effet, la lecture du document informe que, si « La Suisse a une longue tradition de participation à la recherche internationale [et …] est membre de la totalité des organisations intergouvernementales »[5] (OFS 2012, p. 17), les modes de financements privilégiés concernent les institutions et les projets intérieurs, avec une part relative de 68% (OFS 2012. P.18).

Figure 1 : Comparaison des types de financement de la recherche



Source : OFFICE FEDERALE DE LA STATISTIQUE, 2012. Financement public de la recherche en Suisse : Evolution 2000 – 2010. [Fichier PDF]. Septembre 2012.

[4] Critères sélectionnés selon les informations inclues dans les principes et lignes directrices de l'Organisation de coopération et développement économique (OCDE) pour l'accès aux données de la recherche, financées sur fonds publics.

[5] Selon le rapport de l'OFS, la part relative du financement des projets et organisations de recherches internationales représente 10% du budget alloué à la recherche. Ce qui en comparaison avec les statistiques internationales représente 8% de plus que la moyenne des pays membre de l'Organisation de coopération et développement économique (OCDE) sélectionnés (2%)

Selon la Confédération helvétique : l'investissement des autorités dans la recherche et l'innovation est l'un des plus élevé, avec presque 3% du produit intérieur brut (PIB) dont un tiers pour la recherche publique, soit plus de cinq milliards de francs. Elle confirme l'aspect essentiel des collaborations internationales, avec une place importante des organisations et programmes de recherches européens[6] en indiquant également l'existence de coopérations privilégiées avec une sélection de pays partenaires extra-européens[7]. Elle indique enfin que les collaborations internationales sont du ressort des Hautes Ecoles, au travers d'accords interinstitutionnels (Confédération helvétique 2013).

A la lumière des informations abordées ci-dessus, on peut constater que les acteurs disposant d'une forte influence sur la recherche scientifique helvétique sont essentiellement indigènes et européens. C'est sur ces derniers qu'il a paru pertinent au rédacteur d'appuyer le développement du contexte.

2.1 Politique

Le libre accès aux publications scientifiques et aux DR est au centre de l'actualité politique européenne avec la récente décision des ministres en charge de la recherche et de l'innovation, annoncée dans le communiqué de presse du 27.05.2016 :

> *« A partir de 2020, toutes les publications scientifiques sur les résultats de la recherche financée par des fonds publics seront en accès libre […], [ce qui] permettra une réutilisation optimale des données de la recherche. Celles-ci doivent être […] disponibles, sauf en cas de raisons fondées de déroger à ce principe, telles que le droit d'auteur, la sécurité ou le respect de la vie privée ».*
>
> (La présidence néerlandaise de l'UE 2016)

Ce communiqué annonce le plan d'action européen pour la science ouverte[8], qui indique que l'ensemble des parties prenantes : « scientifiques, [institutions académiques…] organismes de recherche, éditeurs, délégués des états membres […] » (La présidence néerlandaise de l'UE 2016) devront collaborer afin de répondre à deux exigences.

---

[6] A titre d'exemple, le Centre européen sur la recherche nucléaire (CERN) et le Programme-cadre de recherche de l'UE.

[7] Les états concernés sont : l'Afrique du sud, le Brésil, la Chine, la Corée du sud, l'Inde, le japon et la Russie et représentent de « forts potentiels de développement scientifique et technologique » (Confédération helvétique 2013b).

[8] La conférence « science ouverte - de la vision à l'action » est le point de départ du plan d'action d'Amsterdam sur l'innovation en matière de science ouverte, qui a donné lieu au communiqué de presse cité plus haut dans ce chapitre. Disponible à l'adresse : http://francais.eu2016.nl/a-la-une/actualites/2016/04/05/plan-d%E2%80%99action-europeen-pour-la-science-ouverte

1) Permettre une accessibilité complète et libre, dès 2020, à l'ensemble des produits de la publication scientifique financée par des deniers publics.

2) Que l'ensemble des DR produites dans le cadre de recherches financées par des fonds publics soient l'objet d'un partage systématique afin d'en garantir la réutilisabilité (La présidence néerlandaise de l'UE 2016).

Cette prise de position de l'UE, en faveur du libre accès aux résultats de recherche est une véritable évolution de l'écosystème scientifique. Elle invite l'ensemble des acteurs du domaine à l'échelle internationale à collaborer afin de contribuer à une harmonisation des processus de chacun, pour parvenir à accroître le partage, l'impact et la prise en compte des résultats de recherche.

En Suisse, le Fond national de la recherche scientifique (FNS)[9], l'un des acteurs majeurs du paysage académique, a rapidement réagit à cette annonce dans la « news room » de son site internet en publiant le communiqué de presse intitulé : Partager le savoir : Open Access[10] pour tous.

Véritable prise de position en faveur d'une recherche à l'accès publique et gratuit, ce texte confirme l'engagement de l'institution en faveur de l'Open Access (ci-après OA) et indique que celle-ci poursuit le même objectif temporel, conformément à l'appel d'Amsterdam pour une science ouverte : « la part des publications scientifiques en OA devrait atteindre 100% d'ici 2020 ». (FNS 2016a).

Il informe que les bases du développement d'une stratégie nationale pour la publication en OA, en collaboration avec Swissuniversities[11] sont en cours d'élaboration. Il s'appuie également sur les résultats intéressants de la politique d'encouragement[12] mise en place

[9] Description du FNS selon le site internet de l'organisation : « Sur mandat de la Confédération, […] il encourage la recherche fondamentale dans toutes les disciplines scientifiques […], ce qui fait de lui la principale institution suisse d'encouragement à la recherche scientifique ». (FNS 2016b)

[10] Selon l'Université de Genève : les documents OA « offrent […] un droit d'accès gratuit, irrévocable et universel, [la possibilité de] réutilisation […], en réservant l'obligation de citation F ». (UNIGE 2014)

[11] Description de l'Association selon son site internet : « fondée en 2012 par les hautes écoles universitaires, spécialisées et pédagogiques de Suisse qui vise à renforcer la collaboration [institutionnelle] et à favoriser l'expression commune de l'espace suisse de l'enseignement supérieur ». (Swissuniversities 2015a)

[12] Selon l'article 47 de son règlement des subsides (FNS 2016c) : Publication et mise à disposition des résultats de la recherche : « les bénéficiaires de subsides s'engagent à ce que les résultats de recherche […] soient rendus accessibles au public de manière appropriée ».

actuellement, en indiquant un taux de 40% des publications financées par l'institution disponible en libre accès. Il est intéressant de noter que les DR doivent également être mises à disposition sur des infrastructures digitales de données scientifiques reconnues (FNS 2016b).

Pour parvenir à mettre en œuvre cette évolution, l'institution appelle à un changement fonctionnel, une restructuration du système de publications qui permettra d'attribuer les fonds publics aux organismes de recherche et non aux éditeurs scientifiques (FNS 2016a).

2.2 Economique

Les acteurs évoluent dans un contexte économique fragilisé, qui souffre toujours des marasmes qui trouvent leur origine dans la crise financière de 2008[13], qui, selon le Fond monétaire international a provoqué un ralentissement de l'économie mondiale et une forte diminution de la croissance (FMI 2008). La Suisse, quant à elle, doit encore faire face aux conséquences de la crise du franc fort.[14] Comme le suggère l'article Making data sharing count : a publication based solution : « la détérioration de l'environnement économique et des conditions de financements, a rendu crucial le ratio coûts/bénéfices dans l'ensemble des discussions liées au partage des [DR] »[15] (Poline et al. 2012 cité dans Gorgolewski, Margulies, Milham 2013, p.1).

Selon l'OCDE : « Les systèmes scientifiques publics des pays membres […] sont fondés sur le principe de l'ouverture et de la libre circulation des données » (OCDE 2007). Cette situation implique de prendre en compte, dans les processus de financement de la recherche, les coûts de mise en place et de maintenance des conditions nécessaires à l'application de cette doctrine. Celle-ci s'appuie sur le fait « que l'agrégation des DR en collections augmente en raison de la croissance de leur potentiel de réutilisation pour résoudre de nouvelles questions de recherche ».[16] (Kim 2013, p. 494)

[13] Cette récession est la conséquence d'un krach boursier en 2008, suite à la problématique de l'explosion de la bulle immobilière aux Etats-Unis d'Amériques (Wikipedia, 2009).

[14] Cette crise s'est déclenché le 15 janvier 2015, suite à la décision de la BNS de supprimer le taux plancher de change pour l'Euro (JANSON 2015).

[15] Traduction par le rédacteur de : « a worsening economic and funding environnement, the cost-benefit ratio has besome central to any disscussion of sharing ».

[16] Traduction par le rédacteur de : « Data's value increase as they are aggregated into collections and they become more available for reuse to adress new research questions » (Kim 2013, p. 494).

Pour parvenir à optimiser la rentabilité du partage des DR et son retour sur investissement (ROI), des charges telles que : les infrastructures digitales, qui « nécessitent une planification budgétaire spécifique continue et un soutien financier adéquat ». (OCDE 2007, p. 14) et les solutions de diffusion, car « communiquer le[s] résultats [de recherche] pour les rendre exploitables […] est un moyen de les rentabiliser ». (Fayet 2013, p.3) Les sources de financements pour les coûts indiqués ci-dessus sont principalement de nature publique, qu'elles proviennent d'agences de financement, de sociétés savantes, ou d'associations institutionnelles (Tenopir et. al. 2011, p.2)[17].

2.3 Technologique

Les acteurs de l'écosystème de la recherche évoluent dans un contexte technologique à l'évolution rapide qui ouvre de nouvelles perspectives dans la conduite des processus de recherche. De plus, elle permet d'envisager l'utilisation de nouveaux types d'outils, afin d'en augmenter la performance.

L'usage de technologies de l'information et de la communication (TIC), est intégré au fonctionnement des processus de recherche de l'ensemble des disciplines scientifiques et elles ont largement contribué à l'essor de la libre circulation des données. (OCDE 2007)

> *« La réussite de leur implantation vient de leur possibilité de soutenir les pratiques de recherche et de suivre l'évolution culturelle des communautés qui composent le milieu de la recherche académique. Mais leur évolution est plus rapide que les pratiques scientifiques qui ont un rythme de développement plus lent ».*
>
> (Kling & McKim 2000, p.1307)

Les avancées des technologies informatiques, qu'elles soient matérielles ou logicielles permettent d'envisager le développement d'instruments d'accès plus optimaux aux DR.

L'amélioration des capacités de calcul et de stockage des ordinateurs et l'utilisation courante d'internet, comme canal de communication et de transfert de données, ont donné naissance à de nouveaux champs d'application pour les résultats de la recherche et les données qui sont produites lors du processus scientifique. (OCDE 2007)

[17] Ce qui explique en partie le positionnement des acteurs politiques identifiés dans la partie 2.1 de cette étude en faveur de la publication OA pour les DR. La production d'un volume toujours plus important de données impose en effet un modèle de distribution du savoir moins onéreux que le modèle de publication traditionnel.

Les moteurs de recherche deviennent plus performants et sont les instruments d'accès aux données textuelles, mais des développements doivent être envisagés, afin de permettre d'effectuer des recherches plus intelligentes portant sur des données structurées et d'en extraire l'essence au moyen de solution de fouillage de données. L'étape suivante dans le processus d'innovation étant le traitement sémantique des données pour leur apporter du sens.[18]

Le taux de pénétration des technologies du Web 2.0 est très important dans la société actuelle. Le développement d'applications de ce type, configurées pour soutenir les processus de recherche, peut répondre à l'interdisciplinarité et au développement des collaborations scientifiques. En effet, ce type de média :

> *« Partage des objectifs communs avec la communication scientifique et il permet de développer des échanges informels entre chercheurs non seulement sur les résultats d'une recherche, ses données ou ses […] rapports de conférence et peut leur donner un moyen d'atteindre une plus large diffusion ».*[19]
>
> (Collins 2013, p. 90)

2.4 Sociétal

L'idée que les résultats des recherches financées par des fonds publics doivent servir le *bien commun*[20], est en plein essor. Cette perception collective est traduite par une volonté citoyenne marquée par une exigence de transparence des autorités quant à l'utilisation des deniers étatiques. Elle soutient que l'ensemble des productions et innovations scientifiques devraient pouvoir être librement accessibles à la fois à l'ensemble de la communauté scientifique, mais dans un sens plus large, à la société elle-même.[21]

L'importance prise par l'éthique dans la société actuelle impacte également le domaine de la recherche. Elle est un facteur à prendre en considération. En effet, le droit à la protection de la vie privée entraîne une nécessité de mettre en place des processus

---

[18] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

[19] Traduction par le rédacteur de : « Social media share many aims in common with scholarly communications. […] informal ways in whichresaercher communicate with each other throughout tes reseach process, and how they engage with wider audiences ».

[20] « Le bien commun est ce qui est profitable à long terme pour l'ensemble des membres de la société » (Flauhault 2013).

[21] Cette volonté remet en cause le fonctionnement même de la diffusion des savoirs actuelle qui bénéficie principalement aux éditeurs commerciaux.

d'anonymisation des données, afin de protéger les participants humains de certaines expériences, pour ne pas limiter leur partage.

2.5 Légal

La loi fédérale sur le principe de la transparence dans l'administration (LTrans 2014), adoptée en 2006 et qui a pour objectif de « promouvoir la transparence quant à la mission, l'organisation et l'activité de l'administration […] en garantissant l'accès aux documents officiels ». (LTrans 2014), entraîne, selon le mouvement Opendata.ch[22], une modification importante de l'écosystème des documents publiques. Ceux-ci doivent, sauf exceptions, être accessibles pour l'ensemble des citoyens (Opendata.ch 2013), ce qui entraîne, de fait, la nécessité de partager les DR produites dans le cadre de la recherche financée par des fonds d'origine publique.

Les institutions académiques et les acteurs du domaine de la science sont tenus de respecter le cadre législatif dans lequel ils évoluent. En effet, le partage de données sensibles, tels que des dossiers médicaux, doit se référer à loi sur la protection des données personnelles (LPD 2014), alors que les problématiques d'attribution des droits de propriété sur les données sont concernées par la loi fédérale sur le droit d'auteur et les droits voisins (LDA 2011). Enfin, le libre partage des DR implique l'utilisation de solutions permettant d'identifier leurs créateurs et les droits de diffusion appliqués à celles-ci.[23]

[22] Voir le chapitre 3.1.5 dédiée aux mouvements.

[23] Ce qui implique l'utilisation de licence qui permettent de définir les modalités de partage sur les aspects de droit d'auteur et de propriété intellectuelle.

3. Acteurs

La réalité de l'augmentation des volumes de DR impliquera tôt ou tard, l'ensemble des acteurs du domaine scientifique (Quaglia 2015). Aussi, l'un des objectifs du présent travail est d'identifier les acteurs-clés impliqués dans le processus de partage des DR, ce qui constitue une étape essentielle pour la mise en place d'une solution de diffusion. En effet, une mise en lumière des besoins, attentes et du rôle des parties prenantes est une base non facultative, pour permettre le développement de propositions de solutions de diffusion cohérentes.

3.1.1 Chercheurs

Le partage des DR est « une responsabilité fondamentale de leur auteur » (Huang, Hawkins, Qiao, 2013, p. 5), en effet, « Les questions d'acquisition, d'accès et d'exploitation des DR sont indissociables des impératifs scientifiques et ne peuvent être envisagées qu'avec la communauté des chercheurs » (Fayet 2013, p. 8).

Dès lors, en qualité de producteurs de données, les scientifiques sont fortement impactés par les mutations des pratiques de partage des DR. Ils doivent répondre aux exigences contractuelles imposées par les agences de financement et les éditeurs scientifiques. Ils sont aussi déterminant dans le processus de tri des DR, car ils ont la responsabilité de définir quelles données disposent d'une valeur suffisante pour être conservées et partagées.

- On peut donc considérer qu'un modèle de diffusion des DR doit proposer un **cadre** permettant aux chercheurs de disposer d'informations sur le **type de données devant être stockées et diffusées**, en fonction notamment de leur valeur et de leur potentiel de réutilisabilité.

La préparation des données, pour en permettre le partage, leur incombe également. Mais, comme l'indique Gorgolewski, Margulies et Milham (2013 p.1), dans leur article Making data sharing count : a publication-based solution, Il est complexe de démontrer aux chercheurs que les bénéfices du partage des DR pour la communauté scientifique dépassent les investissements nécessaires à la préparation de celles-ci, ce qui peut être perçu comme une perte de productivité à l'échelle des chercheurs.

- Les articles consultés dans le cadre de ce travail indiquent l'importance de la mise en place d'un système qui permet de **récompenser les producteurs de données pour le travail effectué** lors de leur préparation. Les scénarios proposés doivent donc tenir compte de cet élément, en intégrant des fonctionnalités permettant d'attribuer la paternité scientifique des sets de données partagés.

Enfin, leur qualité d'utilisateurs potentiels des DR préexistantes implique que leurs besoins soient également pris en compte par les autres acteurs de cet écosystème.

- Dans ce cadre, on peut noter que les possibilités d'accès aux données, par des canaux hors du domaine institutionnel, entraîne des **modifications de leurs pratiques de recherches informationnelles**, illustrées par la baisse statistique des consultations des produits d'Elsevier par les chercheurs affiliés à l'UNIGE.[24]
- Les **pratiques et les besoins des chercheurs pour le partage et l'utilisation des DR sont hétérogènes** et sont liés à la fois à leur discipline scientifique et au fonctionnement des institutions auxquelles ils sont affiliés. Les résultats de l'enquête effectuée par le projet *Data Life Cycle Management* (ci-après DLCM) l'illustrent. Ils ont montré que la perception des chercheurs sur la gestion de DR était très hétéroclite au sein de l'UNIGE. En revanche et de manière générale, les scientifiques les plus favorables au partage, sont ceux qui dans leurs processus de recherche, sont confrontés à de nombreux formats et types de données.[25]
- Les outils mis en place doivent donc pouvoir répondre aux besoins spécifiques des scientifiques, afin de garantir que les **DR constituent une véritable plus-value** pour ceux-ci.

3.1.2 Agences de financement

Comme indiqué dans la description du contexte de ce travail, les agences de financement les plus importantes pour la recherche helvétique se positionnent de manière claire. Elles sont en faveur du partage des DR et de leur réutilisation pour l'ensemble des projets de recherche qu'elles financent (Claivaz, Dieudé, Krause 2015)

En Suisse, il existe deux partenaires qui sont chargés par la Confédération de soutenir la recherche scientifique en attribuant des fonds ; le FNS, qui est la principale institution d'encouragement à la recherche et la commission pour la technologie et l'innovation (CTI).[26]

Il est également important de comprendre que les institutions académiques elles-mêmes sont des sources potentielles de financement, comme l'indique Swissuniversities (2015a) : « les hautes écoles mettent leurs propres fonds à disposition pour la recherche ».

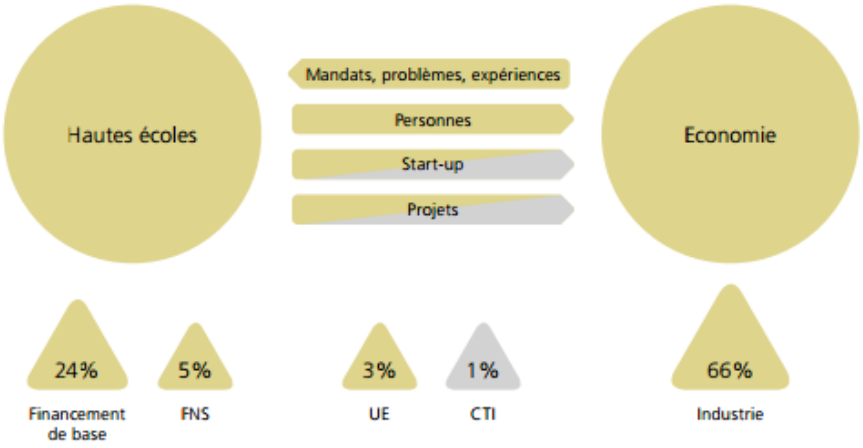
[24] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

[25] Entretien avec Eliane Blumer, coordinatrice du projet DLCM pour l'UNIGE, Genève, 24 mai 2016.

[26] « Agence de la Confédération, la CTI est l'organe chargé de l'encouragement de l'innovation ». (Confédération suisse 2016)

L'illustration ci-dessous permet de souligner que les instances académiques suisses sont avant tout soutenues par le FNS. Mais suggère que des financements attribués dans le cadre de projets impliquant des acteurs non académiques, une partie des ressources financières peut être mise à disposition par la CTI, dans le cadre de développement de start-up ou de projets, par exemple.

Figure 2 : CTI dans l'écosystème suisse de l'innovation



Source : CONFEDERATION SUISSE, 2015. *2017 à 2020.* [fichier PDF]. Commission pour la technologie et l'innovation CTI. Avril 2015.

Programme pluriannuel

S'il n'exige actuellement pas la présence d'un plan de gestion des données (ci-après DMP), le FNS considère que le partage des DR est quant à lui obligatoire. En effet, selon l'article 47 du règlement des subsides du FNS : Publication et mise à disposition des résultats de la recherche :

> *« [1] Pendant et après l'achèvement de leurs travaux de recherche, les bénéficiaires de subsides s'engagent à ce que les résultats de recherche soutenus par des ressources du FNS soient rendus accessibles au public de manière appropriée ; […] Cela implique notamment : […]*
>
> *b. qu'ils mettent à disposition d'autres chercheurs […] les données recueillies durant les travaux de recherche soutenus par le FNS et les déposent dans des bases de données scientifiques reconnues, conformément aux prescriptions du FNS […] »* *(FNS 2016b, p.15)*

De plus, la mise en place du partenariat avec Swissuniversities, dans la création d'une politique OA, inclut le partage des DR dans les réflexions. Ceci implique une volonté d'adaptation aux changements structurels induits par les prises de décisions politiques, dans le domaine de la recherche et de l'innovation.

Comme indiqué dans la description du contexte, le troisième acteur intégré aux réflexions, menées dans le cadre de cette étude est la recherche européenne. En effet, au travers du Conseil Européen de la recherche (ERC) et de son programme Horizon 2020, elle établit les objectifs pour la recherche et le développement en Europe, pour la période 2014-2020.

Comme indiqué dans la section consacrée aux avantages du partage des DR, celui-ci permet d'augmenter le ROI pour les organismes de financements. Toutefois, cela engendre la nécessité d'une gestion des données méthodiques par les scientifiques qui bénéficient de financements de ces partenaires publics.[27]

- La mise en place d'un modèle de diffusion pour les DR doit prendre en compte les acteurs du système de financement décrit ci-dessus, ce y compris leurs prises de position et **leurs exigences**.
- Il est notamment indispensable de tenir compte **des conditions particulières qu'impliquent la confidentialité obligatoire** de certaines DR sensibles.[28]
- Si les réflexions réalisées dans le cadre de ce travail concernent en particulier la diffusion de DR publiques, il paraît important de tenir compte de l'implication importante de l'économie privées dans les processus de recherches helvétiques. En effet, les recherches financées par des fonds privés peuvent engendrer **des collaborations avec des entreprises commerciales, qui ne partagent pas les missions des institutions académiques**.[29]

3.1.3 Institutions académiques

Les institutions peuvent jouer un rôle prépondérant dans l'amélioration de l'accès aux DR en prenant la main sur la question de l'archivage des DR et en mettant en place des politiques claires, compréhensibles et faciles d'accès. En effet, la création de dépôts de données institutionnel est une alternative raisonnable à la surcharge des serveurs des éditeurs, malgré le transfert de coûts impliqué par cette solution. (Sturges et al. 2015)

Les hautes écoles et les universités peuvent mettre en place des mesures d'encouragement au partage des DR. En attribuant des récompenses pour les

[27] Qui dans le cadre du pilote de libre accès aux DR Open Research Data, se traduit par l'exigence de la production d'un DMP dans les six premiers mois de la vie d'un projet, et du dépôt des DR sur un dépôt de données le plus rapidement possible, avec les métadonnées associées. (Ministère de l'éducation nationale, de l'enseignement et de la recherche, 2013)

[28] A titre d'exemple, les chercheurs bénéficiant de subventions du FNS doivent dûment justifier les intérêts de confidentialité et les soumettre pour approbation à l'agence de financement. (FNS 2016)

[29] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

chercheurs et en mettant à disposition des jeux de données de qualité, elles répondent aussi au besoin de reconnaissance du travail de préparation.

- La mise en place de métriques pour mesurer les citations des DR est l'une des pistes potentielles pour la valorisation des activités liées au partage de ces dernières. Cet élément doit être pris en compte dans les scénarios, car il peut constituer une récompense substantielle.
- Comme indiqué plus avant de cette étude, dans le cadre du fonctionnement de la recherche helvétique, **les collaborations interinstitutionnelles impliquant des échanges internationaux sont réalisées sous l'égide des institutions académiques**. De fait, elles doivent faire face aux exigences d'agences de financement dépassant le cadre développé ici.

Enfin, elles doivent établir des politiques de gestion des DR qui encouragent leur partage. Ces nouvelles manières de gérer doivent aussi permettre de faire face aux défis que représentent l'augmentation du volume de données, le maintien de leur intégrité et de garantir la reproductibilité de la science.

- La HES-SO a mis en place deux groupes de travail distincts. Le premier, de nature informelle, a pour mission de mettre en place une politique de gestion des DR et de collaborer aux projets nationaux dans le domaine.

  Le second est placé sous la direction des domaines HES-SO[30] et gère les DR au moyen d'une archive ouverte thématique : ArODES.[31] [32]
- La responsable de la *DIS*[33] de l'UNIGE, estime que les bibliothèques ont un rôle stratégique à jouer dans la gestion des DR. Une transformation d'une partie des missions traditionnelles doit permettre de justifier l'existence des services d'information documentaires.

  Les démarches en cours, au sein de l'institution, sont l'élaboration d'une politique institutionnelle et le développement de services dédiés à la gestion des DR.[34]

3.1.3.1 DLCM

Afin de répondre aux enjeux identifiés ci-dessus, les instances académiques suisses, ont mis en œuvre le projet DLCM, en 2015. Ceci s'est faut dans le cadre d'un programme

[30] Par analogie, peut être comparé aux facultés des universités.
[31] Archive institutionnelle du domaine économie et services de la HES-SO, qui valorise et conserve les publications et données des chercheurs affiliés et qui correspond probablement, selon Ruch, (2016) aux solutions actuelles les plus performantes.
[32] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.
[33] Division de l'information scientifique
[34] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

de soutien national visant à améliorer leurs performances dans la mise à disposition et le traitement des informations scientifiques : *CUS-P2*[35].

- Le projet n'a pas pour objectif de prendre en considération l'ensemble des DR. En effet, les *Big Data* et les disciplines disposant d'une forte culture de partage, possèdent déjà des modèles et des outils de diffusion. DLCM vise la *Long tail of science*.[36]

Selon sa coordinatrice, cette collaboration entre les partenaires des hautes écoles suisses[37], développe des solutions adaptées aux besoins des chercheurs dans chacune des étapes du cycle de vie des données. Ceci se fait dès le dépôt du projet de recherche, avec un accompagnement dans la rédaction d'un DMP, conformément aux exigences des agences de financement. A cela s'ajoute un soutien dans les processus de gestion : dépôt, partage et publication.

Le projet, commencé en septembre 2015 a obtenu une prolongation jusqu'à fin 2018 et son objectif est le développement de services par les institutions partenaires dans le cadre national.[38]

S'il n'existe, à l'heure actuelle, aucune politique institutionnelle définissant ce qui doit être réalisé dans le cadre de la gestion des DR, l'un des livrables prévus par le projet est un modèle de politique[39], que les partenaires DLCM tentent de valoriser sur le plan national. Le FNS sera d'ailleurs contacté en vue d'une collaboration. L'UNIGE et l'UNIBAS tentent de développer une politique sur la base dudit modèle. Il est à noter qu'une implémentation de solutions présente de nombreuses difficultés et que celle-ci ne sera probablement pas effective d'ici la fin 2016. [40]

La *Data Management Checklist* (DCLM 2016), produite également dans le cadre du projet, par une collaboration entre les équipes de spécialistes[41] des deux Ecoles

---

[35] Les objectifs de ce programme sont : l'harmonisation de l'offre numérique de contenus à caractère scientifique et des instruments de traitement y relatif. (Swissuniversities 2015)

[36] Voir ref. 34

[37] La collaboration comprend des professionnels des Ecoles polytechniques fédérales de Lausanne (EPFL) et Zurich (ETHZ), de la Haute école de gestion de suisse occidentale (HES-SO), des universités de Lausanne (UNIL), Bâle (UNIBAS), Zürich (UNIZH), et Genève (UNIGE), ainsi que des membres de SWITCH, partenaire privilégié des hautes écoles sur les questions de TIC (Switch 2016)

[38] Entretien avec Eliane Blumer, coordinatrice du projet DLCM pour l'UNIGE, Genève, 24 mai 2016.

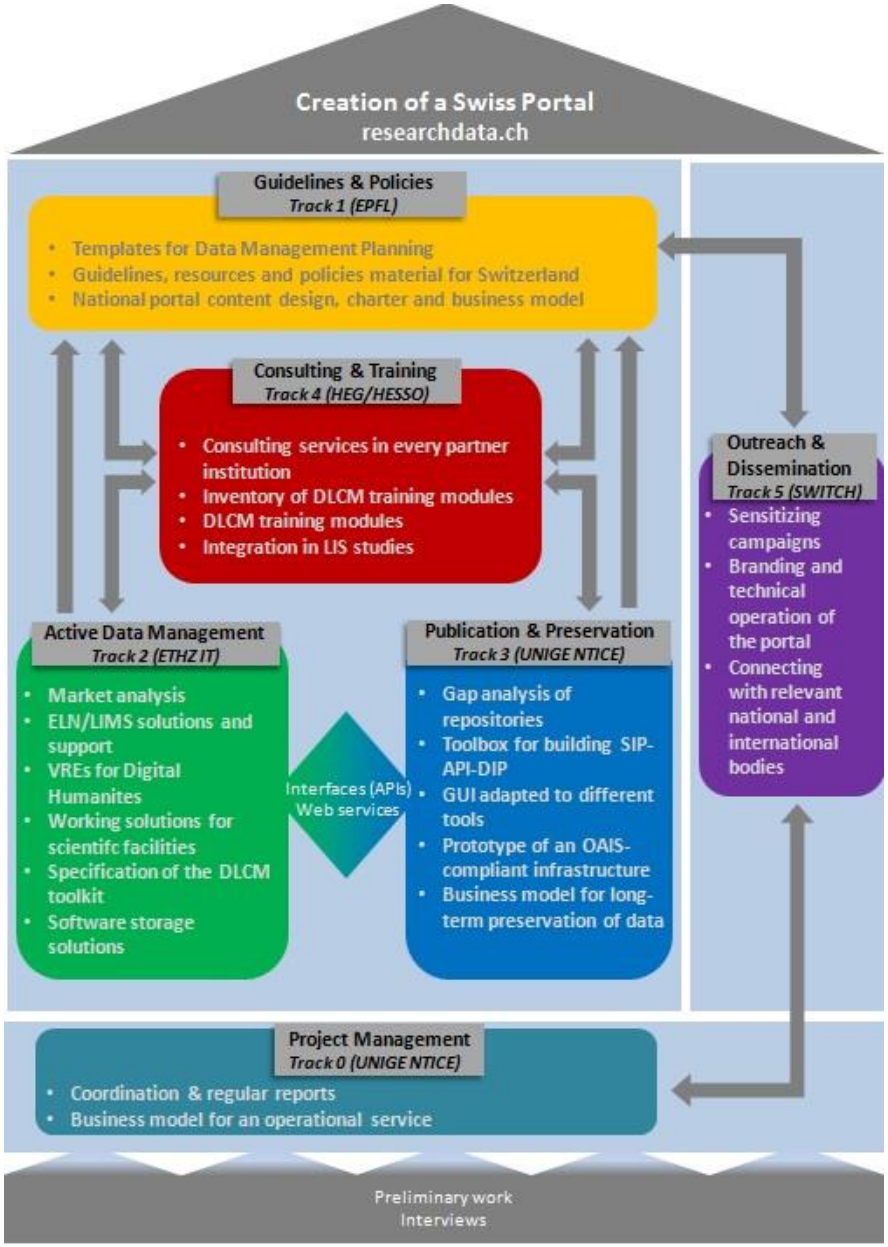
[39] Le modèle se présente sous la forme d'une structure modulaire qui permet de sélectionner le contenu suivant le degré de profondeur souhaité par l'organisme utilisateur.

[40] Voir réf. 38

[41] Le Digital Curation Team (ETHZ) et le Research Data Team (EPFL).

polytechniques fédérales, est une liste de points essentiels dont les chercheurs doivent tenir compte dans leurs pratiques de gestion des DR. Elle présente l'avantage d'être assez flexible pour permettre une adaptation aux exigences des agences de financement (ETH ZURICH, EPFL 2016).

Figure 3 : Représentation graphique du DLCM



Source : DLCM, 2016. First outcomes. *dlcm.ch* [en ligne]. 26 mai 2016. [Consulté le 30.05.2016]. Disponible à l'adresse :

http://dlcm.ch/2016/05/26/first-outcomes/

- Le projet DLCM doit être considéré comme un partenaire dans la recherche de solutions de diffusion, son objectif terminal consistant à la création d'un portail helvétique pour les DR : dlcm.ch. (DLCM 2016)

3.1.3.2 openresearchdata.ch

Figure 3 : Logo Open Research Data

Source : OPENRESEARCHDATA, 2016. Openresearchdata.ch [en ligne]. [Consulté le 30 mai 2016]. Disponible à l'adresse : https://openresearchdata.ch/

Le projet Open Research Data Pilot plateforme Suisse[42] a, dans le cadre du financement du programme CUS P-2[43], établi une plateforme d'accès aux métadonnées pour la recherche ouverte en Suisse (Openresearchdata 2015). Celui-ci prône une approche pluridisciplinaire, visant à mettre à disposition des chercheurs une vision d'ensemble des DR, provenant de la recherche financée par des fonds publics.

3.1.3.3 Bibliothèques académiques

Comme l'indique Heidorn (2009) en citant Treylor (2006, 2007) « Les bibliothèques jouent un rôle de plus en plus important dans la curation[44] des [DR] »[45]. Le processus est illustré ci-dessous, par la représentation graphique du modèle développé par le *Digital Curation Center* (DCC)[46]. En effet, c'est au travers des services que les institutions académiques ont la capacité de mettre en place une aide structurée, destinée aux chercheurs qui leur sont affiliés.

---

[42] Cette collaboration implique le FORS, un centre de compétences national pour les sciences sociales (forscenter 2016), le Digital Humanities Lab d'UNIBAS, le service IT de l'ETHZ et le Swiss Intitute of Bioinformatics (SIB)

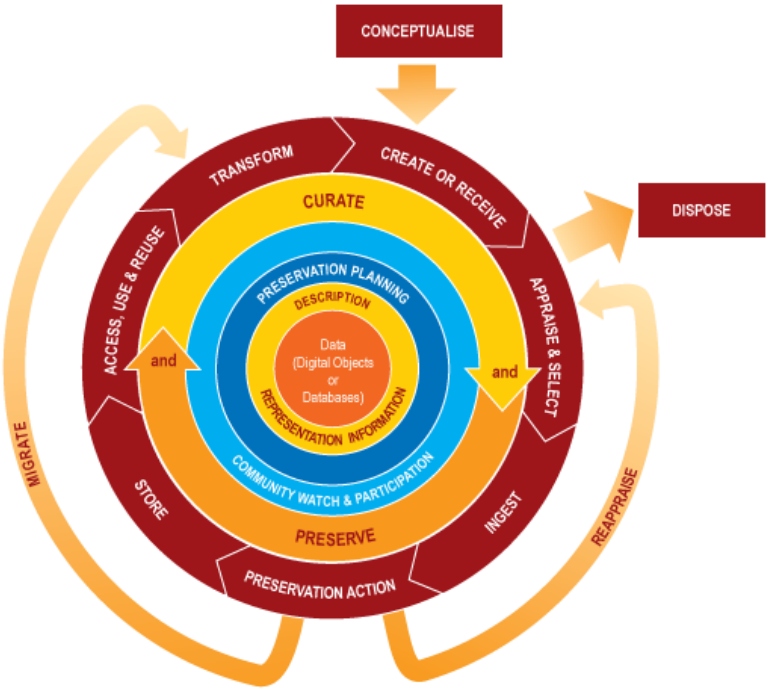
[43] Voir réf. 38

[44] « La curation désigne un ensemble de pratiques et une gamme d'outils destinés à des opérations de republication » (ADBS 2012) Développée à l'origine pour les contenus web, elle est appliquée aux DR, afin de garantir leur qualité et leur reproductibilité.

[45] Traduction par le rédacteur de : « Libraries are increasingly playing a role in science data curation ».

[46] Le DCC est un centre d'expertise dans la curation digitale avec un focus sur les compétences (DCC, 2016a)

Figure 4 : DCC Curation Lifecycle Model



Source : DCC, 2016b. DCC Curation Lifecycle Model. dcc.ac.uk [en ligne]. [Consulté le 1er juin 2016]. Disponible à l'adresse : http://www.dcc.ac.uk/resources/curation-lifecycle-model

Les professionnels de l'information peuvent faire valoir des compétences potentielles en s'appuyant sur leur expertise dans la gestion de métadonnées et sur la description, la curation et l'enrichissement des informations, ainsi que sur une certaine expérience de l'archivage à long terme.

Ceci permet aux bibliothèques de se positionner comme des actrices à part entière de la gestion des DR face aux chercheurs et de développer de nouvelles utilités pour eux. Car si les services d'information documentaire restent très pertinents pour répondre aux besoins de l'enseignement, ils sont de moins en moins considérés par les chercheurs, la totalité des informations nécessaires à leurs travaux de recherche étant disponibles en ligne.[47]

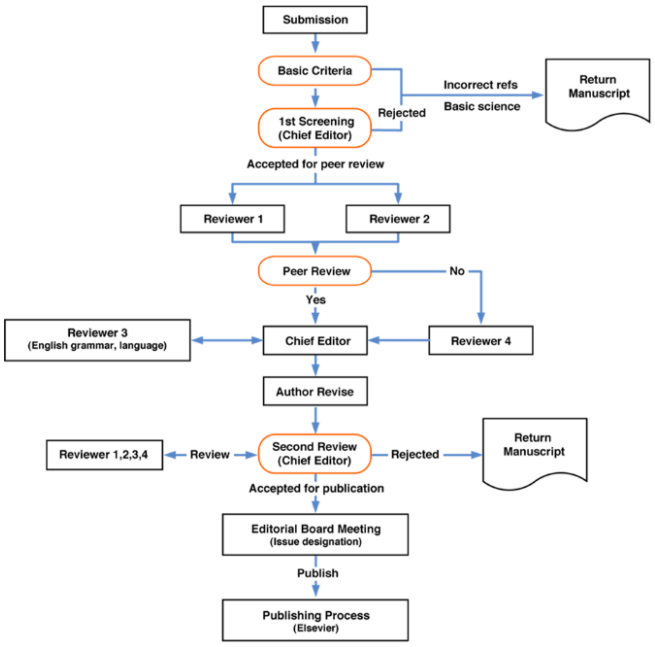
3.1.4 Editeurs

Selon McGrath (2013), dans The Future of Scholarly Communication, le rôle des éditeurs dans les processus de diffusion des résultats de la recherche consiste à gérer en premier

[47] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

lieu, la sélection des articles correspondant à leur public cible. Puis vient la validation de ceux-ci par les pairs, en organisant le *peer review*[48]. Il y a également la préservation de la communication scientifique et le soutien au développement de nouveaux formats.

Figure 5 : Représentation graphique du peer review



Source : ELSEVIER, 2016. What is peer review ? Elsevier.com [en ligne]. [Consulté le 15 mai 2016.] Disponible à l'adresse : https://www.elsevier.com/reviewers/what-is-peer-review

Il s'agit donc de valoriser les contenus proposés dans leurs journaux, revues scientifiques et solutions de bases de données. Cette valeur ajoutée justifie, de fait, les coûts de leurs services.

- Afin de contrôler la qualité des DR qu'ils diffusent, les éditeurs doivent disposer de métadonnées et de documentations les plus exhaustives possible. Ceci implique la mise en œuvre de politiques accessibles, qui informent des exigences que doivent respecter les DR pour pouvoir faire l'objet d'une publication. (Huang, Hawkins, Qiao, 2013)

Comme indiqué lors de la description du contexte technologique, en préambule de ce travail, l'augmentation des possibilités d'usage des moteurs de recherche et des logiciels

[48] La revue par les pairs est un processus de relecture par des pairs (comprendre ici d'autres scientifiques disposant d'une expertise dans le domaine concerné par l'article, ou le set de données), qui vise à valider les résultats de recherche afin de garantir la qualité des publications (Elsevier 2016)

permettant d'extraire le contenu d'un article, implique l'adaptation des outils proposés, poussant les éditeurs à :

> *« Développer plus avant le modèle de publication via les bases de données et continuer à chercher des moyens d'extraire plus de valeur de leurs contenus, au travers de développements des technologies de recherche, de lien et des outils de fouillage des données. »*[49]
>
> (McGrath 2013, p. 113).

Pour illustrer ceci, les parties prenantes du projet *DLCM* ont pu collaborer avec des professionnels de l'édition affiliés à Springer, lors de Workshop sur les DR. Ce dernier projet développe actuellement, une solution permettant de relier les données et les publications aux dépôts en collaboration avec un laboratoire de médecine. Il contient des fonctions de publication et d'analyse, ce qui entraîne une accélération de l'automatisation des processus de publication.[50]

- Comme le suggère Eliane Blumer, coordinatrice du projet DLCM, il est important de prendre en compte les avancées de tels parties prenantes et de les inclure dans les réflexions visant à développer un modèle de diffusion. Car, une solution intégrant la possibilité d'une publication rapide, en lien direct avec des publications scientifiques et disposant d'une excellente visibilité[51] est développée, cela peut représenter une concurrence trop importante pour les institutions académiques.[52]

Enfin, l'augmentation du volume de données, produites dans le cadre des processus de recherche, provoque chez certains éditeurs des questionnements quant à la résistance de leurs cyber infrastructures à cette croissance. Celle-ci entraîne la nécessité de mettre en place une plus importante capacité de stockage. (Sturges et al. 2015)[53]

---

[49] Traduction par le rédacteur de : « Publishers will move further into a database format of publishing and continue to look at ways to extract more value from the contnent, through developments in search technology, linking and text mining tools ».

[50] Entretien avec Eliane Blumer, coordinatrice du projet DLCM pour l'UNIGE, Genève, 24 mai 2016.

[51] Elle souligne ici le potentiel apport sur le H-Index des chercheurs qui disposeraient de ce type de solution.

[52] Et à terme exclure les institutions académiques de la gestion de leurs propres DR, à l'image de la situation actuelle quant aux publications scientifiques.

[53] Ce qui entraîne la possibilité, pour les institutions académiques de se positionner comme acteur dans les infrastructures de stockage, comme abordé en introduction du chapitre 3.1.3

3.1.5 Mouvements

Les mouvements et associations, qui travaillent à la promotion du libre accès des données publiques ou de recherches, sont des partenaires potentiels à prendre en compte dans les réflexions réalisée dans la thématique de ce travail.

3.1.5.1 Opendata.ch

Figure 6 : Logo Opendata.ch

Source : Opendata.ch 2016. Organisation. *Opendata.ch* [en ligne]. [Consulté le 4 mai 2016].] Disponible à l'adresse : http://fr.opendata.ch/organisation/

Opendata.ch est la section suisse de l'Open Knowledge Foundation, dont l'objectif est « d'ouvrir les données publiques, de façon libre et réutilisable pour plus de transparence, d'efficacité et d'innovation ».

3.1.5.2 Research data alliance

Figure 7 : Logo RDA Europe

Source : RDA 2016. The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data. *Europe.rd-alliance.org/* [en ligne] Disponible à l'adresse : http://europe.rd-alliance.org/

L'Alliance pour les DR (RDA) a pour objectif de constituer les bases sociales et technologique qui rendent possible le partage des DR. Ceci, tout en considérant que les chercheurs doivent partager librement leurs données au sein d'une collaboration à la fois interdisciplinaire et internationale, afin de répondre au défi sociétal que constitue une science ouverte. (RDA 2016)[54]

[54] En raison du contexte de cette étude, les représentations suisse et européenne de ces mouvements ont été sélectionnées.

4. Données de recherche

Afin de pouvoir proposer des critères et des scénarios pertinents pour la mise en place d'une solution de diffusion pour les DR, il est essentiel de pouvoir comprendre ce qu'elles sont, ce qu'implique leur gestion et quels sont les acteurs concernés par cette dernière dans le spectre du contexte défini.

4.1 Définitions

4.1.1 Données de recherche

Le concept de DR est complexe à définir, car celles-ci peuvent apparaître sous des formes hétérogènes. Afin de comprendre ce qu'elles sont, il faut tenir compte de la complexité qu'entraîne la nature même de ces informations factuelles qui ne peuvent exister en tant que telles que si leur enregistrement lui-même, constitue un acte scientifique (Borgman 2012).

Cet imbroglio intrinsèque débouche sur des définitions variables en fonction des approches des protagonistes du domaine de la recherche scientifique.

- Dans ses principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publiques, l'Organisation de coopération et de développement économiques (OCDE) considère les DR comme des :

> *« Enregistrements factuels (chiffres, textes, images et sons) qui sont utilisés comme sources principales pour la recherche scientifique et son généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de recherche. […] Ne concernent pas les données administratives, les supports de cours, les carnets de laboratoires, les objets matériels[55] et les communications avec les collègues ».* *(OCDE 2007, p. 18)*

- La Royal Society, institution britannique dédiée à la promotion des sciences introduit la notion de données brutes :

> *« Des informations qualitatives ou quantitatives […] qui sont factuelles. Ces données peuvent être brutes ou primaires (directement issues d'une mesure), ou dérivées de données primaires, mais ne sont pas encore le produit d'une analyse ou d'interprétation autre que des calculs »* [56] *(The Royal Society 2012 cité dans Gaillard 2014, p.17)*

[55] Par exemple : échantillons de laboratoire, souches bactériennes, etc.

[56] Traduction par R. Gaillard (2014) de : « Qualitative or quantitative statements or numbers that are (or assumed to be) factual. Data may be raw or primary data (eg direct from measurement), or derivative of primary data, but are not yet the product of analysis or interpretation other than calculation »

- Dans l'article intitulé : The Conundrum of Sharing Research Data, Borgman précise cette définition en indiquant les données qui nécessitent l'assistance de logiciels computationnels pour être utilisables, ainsi que celles générées et compilées par du travail humain ou des machines. (Borgman 2012)
- En raison du contexte dans lequel s'inscrit ce travail, intégrer la définition établie par la Commission Européenne pour Horizon 2020, qui inclut les métadonnées associées est primordial :

> *« 1) Les données, incluses les métadonnées[57] associées […], doivent valider les résultats présentés dans des publications scientifiques ; 2) les autres données[58], métadonnées associées incluses »[59]*
>
> *(European commission 2013, p. 14)*

- Dans le cadre des journées des archivistes des universités et hautes école suisses, l'université de Lausanne propose une définition construite en analysant plusieurs autres déterminations, caractères et types de données :

> *« Unité informationnelle, créée, collectée ou rassemblée, issue du travail savant qui représente un enregistrement objectif de faits objectifs […] organisée ou formatée pour être rendue communicable et interprétable[60], susceptible de subir un traitement informatisé, dont la volumétrie peut être importante et qui est accessible à terme, à la communauté ».* *(UNIL 2014, p. 7-8)*

C'est cette dernière option qui est retenue dans le cadre de cette étude, car en plus d'avoir été développée par des professionnels évoluant au cœur du contexte helvétique, elle intègre l'ensemble des éléments introduits dans les autres propositions. Elle permet de tenir compte de l'ensemble des composantes identifiées, en y ajoutant la notion de volumétrie qui, en raison de l'évolution de la recherche scientifique vers l'exploration des données produites par les instruments de mesure et des données produites par simulations informatiques, entraînent l'apparition du « 4e paradigme ».[61]

---

[57] Les métadonnées descriptives pour les DR déposées.

[58] Les données ayant bénéficié d'une curation qui ne sont pas directement attribuables à une publication, ou données primaires.

[59] Traduction par le rédacteur de : « 1) the data, including associated metadata (i.e. the metadata describing the research data deposited), needed to validate the results presented in scientific publications ; 2) other data (i.e. curated data not directly attribuable to a publication, or raw data), inculding associated metadata ».

[60] Ce qui englobe la présence de métadonnées descriptives et d'une documentation visant à permettre l'interprétation et la réutilisation des DR.

[61] Selon la théorie développée par Gray, les paradigmes scientifiques peuvent être définis en fonction de l'évolutions des pratiques au cours de l'histoire de la science. 1) La recherche empirique, qui vise à décrire les phénomènes scientifiques, 2) la recherche théorique, qui se base sur des modèles en faisant usage aux abstractions et généralisations, 3) la recherche computationnelle, intégration de la programmation informatique en qualité d'outil et d'expression qui permet des modelisations complexes, 4) la recherche exploratoire sur les données, qui permet d'unifier les paradymes précédent en appliquant la gestion des données et les statistiques. (Gray 2006, p.5)

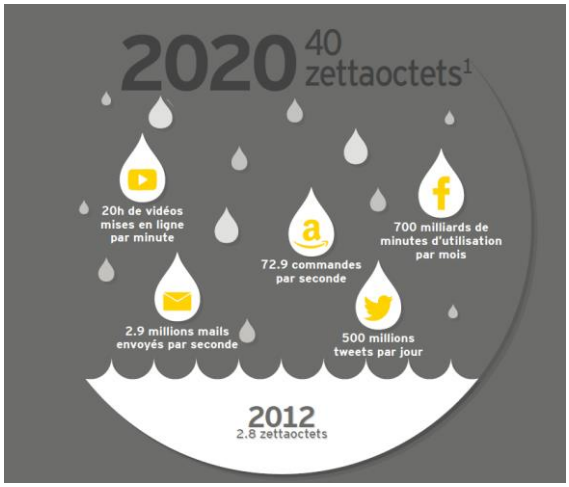
4.1.2 Sets de données

Les travaux scientifiques débouchent sur la création de DR qui sont regroupées en jeux de données ou *datasets,* afin d'en permettre le partage. Cette structure constitue une « agrégation de données brutes ou dérivées présentant une certaine « unité », rassemblées pour former un ensemble cohérent ». (Gaillard 2014) C'est sous cet aspect que sont stockées les données résultant de projet de recherche. Une structuration en sous sets est utilisée lorsque les données atteignent un volume important.

4.1.3 Volume & origine

La perception actuelle de la communauté scientifique sur les questions des DR, engendre une véritable *science des données*[62]*,* selon laquelle « toute information est appréhendée comme un élément d'un ensemble plus vaste et la gestion de ces ensembles de données est devenue un défi majeur ». (Claivaz, Aude, Krause 2015, p. 24-25)

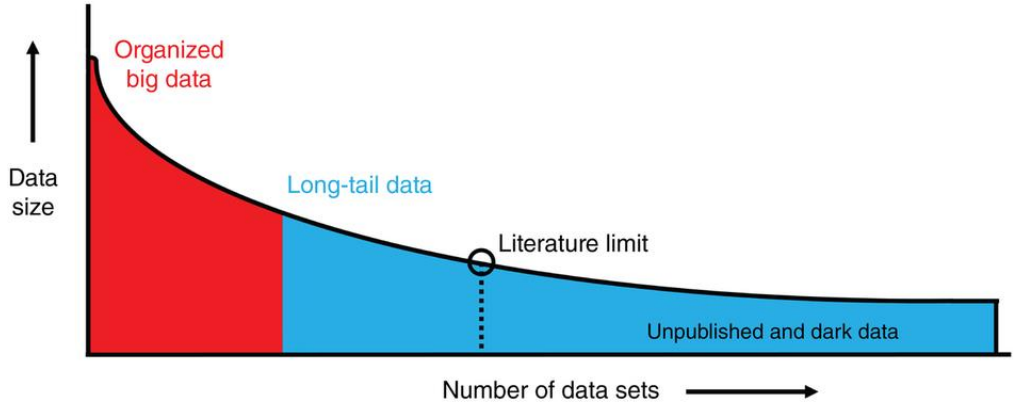
Figure 8 : Infographie sur l'origine des données



Source : Ey 2011. La donnée, un outil de transformation et d'innovation. ey.com [en ligne]. [Consulté le 12 mai 2016] Disponible à l'adresse : http://www.ey.com/FR/fr/Issues/ey-Comment-la-science-des-donnees-permet-elle-la-creation-de-valeur)

[62] Cette science, développée par Cleveland (2001), est une agrégation scientifique pluridisciplinaire autour de l'extraction de connaissance sur la base de données. On peut notamment citer : les statistiques, les mathématiques et les sciences de l'information.

Figure 9 : Schematic illustration of long-tail data



Source : FERGUSON, Adam R. et al. 2014. Big data from small data : data-sharing in the 'long tail' of neuroscience. *Nature.com*. [en ligne]. 28 octobre 2014. [Consulté le 13.04.2016]. Disponible à l'adresse :

http://www.nature.com/neuro/journal/v17/n11/full/nn.3838.html

4.1.3.1 Big data

Le *Big data* est « un concept permettant de stocker un nombre indicible d'informations sur une base numérique », (Lebigdata.fr 2015) dont le développement a été nécessaire en raison de l'explosion du volume des données produites et leur hétérogénéité, dans le cadre de la société de l'information. Il constitue également un important potentiel pour la recherche scientifique et requiert « une quantité massive de puissance informatique pour [être] traitée ». (Royal Society 2012, cité dans JACQUEMOT-PERBAL, COSSERAT, 2015.)

- A des fins d'illustration, dans le cadre d'un cours dispensé à la HEG, le rédacteur a pu être sensibilisé à l'une des applications potentielles de la gestion de ce type de données. En effet, Google peut, en analysant les critères de recherche des utilisateurs de son moteur, prévoir une pandémie de grippe de manière plus rapide que le système de détection mis en place par l'organe de surveillance des maladies infectieuses des Etats-Unis.

4.1.3.2 Small data

Le *small data,* quant à lui, désigne « la quantité de données que vous pouvez aisément stocker et utiliser sur une seule machine et plus précisément sur un ordinateur portable ou serveur de haute qualité » (Pollock 2013 cité dans Cassely 2013)[63]. Ces données constituent, selon l'auteur, la véritable richesse du partage des DR car, prisent dans leur

[63] Traduction de Jean-Laurent Cassely

ensemble, elles disposent d'un potentiel plus important que les *Big data.* Elles sont indiquées dans la figure ci-dessus avec l'appellation *Long-tail data*.

4.1.3.3 Dark data

Le terme *dark data* est associé par la littérature avec la *Long Tail of science*[64], les DR produite dans ce cadre pouvant faire l'objet d'une curation moins rigoureuse. Heidorn (2008, p.1), les définis dans son article, Shedding Light on the Dark Data in the Long Tail of Science. Selon l'auteur :

> *« Les dark data ne sont pas indexées et stockées avec soins, elles sont donc devenues presque invisibles pour les scientifiques et les autres utilisateurs potentiels et de fait restent sous utilisées et peuvent potentiellement être perdues ».* *(Heidorn 2008, p.1)*[65]

4.2 Description

Une donnée documentée au point de pouvoir la rendre indépendante de son contexte de création n'existe probablement pas. Elle est toujours liée à un contexte d'acquisition et intrinsèquement attachée à un domaine scientifique.[66] Il convient donc de la décrire de manière précise et compréhensible, afin de permettre de garantir sa réutilisabilité, car de la qualité et de la richesse des éléments descriptifs, dépendent son interopérabilité et sa bonne utilisation et réutilisation dans le cadre de collaborations.

4.2.1 Documentation

Des informations sur le projet dont sont tirées des DR doivent être formulées de manière adéquate. Elles sont généralement intégrées aux sets de données, au moyen d'un fichier au format .txt ou .pdf et détaillent les hypothèses explorées, les méthodes de collecte, les instruments ou logiciels utilisés (paramètres compris). (JACQUEMOT-PERBAL, COSSERAT, 2015.)

4.2.2 Métadonnées

Les métadonnées[67], véritables cartes d'identité des sets de données, (Enssib 2013) sont l'autre élément primordial de la description des DR. Comme indiqué lors du

---

[64] Fera l'objet d'un développement dans le chapitre consacré à la culture de partage.

[65] Traduction par le rédacteur de : « Dark data is not carefully indexed and stored so it become nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost ».

[66] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

[67] Elles peuvent être descriptives (sujet, etc.), de gestion (auteur, titre, etc.), de préservation (droits, format, résolution, etc.), ou sociales (Enssib 2013)

développement de la définition, elles en sont parties intégrantes. Elles consistent un « ensemble structuré de données qui servent à décrire une ressource, quel que soit son support (papier ou électronique) » (INIST 2014).

5. Partage des données

L'évolution des pratiques et des perceptions scientifiques plaçant la donnée au centre des processus de recherche, l'augmentation exponentielle des volumes de DR produites et la mutation des pratiques scientifiques vers une recherche collective et collaborative, provoquent une prise de conscience des communautés scientifiques sur l'importance du partage des DR.

Celle-ci entraîne une adaptation des exigences contractuelles des agences de financement et des éditeurs scientifiques qui publient les DR, ainsi que des pratiques des chercheurs qui peuvent déterminer la pertinence d'un partage d'un jeu de données, en fonction de son usage potentiel.[68]

Pensé autour de deux objectifs principaux, le partage des DR vise à améliorer l'accessibilité aux données et accroître les opportunités de réutilisation de celles-ci. La possibilité de reproduire des hypothèses de recherche et de les vérifier en réutilisant le matériel ayant permis de les construire étant « une part essentielle du processus scientifique ». (Tenopir et. al 2011, p.1)[69]

5.1 Avantages

Les sources consultées lors de la réalisation de cette étude s'accordent sur les avantages qu'un accès le plus complet et le plus libre possible aux DR apporte au développement de la science. En effet :

- La possibilité d'accéder aux résultats de recherches antérieurs faisant l'objet d'une documentation pertinente en vue de les reproduire, afin d'en vérifier la validité, est un gage de **qualité**.[70]

[68] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.
[69] Traduction par le rédacteur de : « key part of scientific process ».
[70] Ce qui permet d'éviter la production de données frauduleuses et manipulées en raison du contexte concurrentiel et des exigences de publication fréquentes. (Claivaz, Aude, Krause 2015)

- La large diffusion des DR exigeant une gestion attentive de celles-ci, tant au niveau de la description qu'à celui des activités de conservation, elle assure de fait, une meilleure **intégrité des données**[71].
- La réutilisation de données préexistantes pour vérifier de nouvelles hypothèses et la possibilité d'analyser plusieurs jeux de données en les comparant, entraîne à la fois une **accélération des processus d'innovation** et une **diminution des coûts** liés à la collecte des données. En effet, dans un contexte pluridisciplinaire, elle permet de multiplier les interprétations potentielles. (Tenopir 2011)
- L'augmentation de rythme de la recherche et développement provoque **une commercialisation plus rapide** des résultats de recherche, ce qui entraîne une croissance accélérée et **augmente le ROI** des agences de financement. (Conseil de l'Union européenne 2012)
- La mise à disposition systématique des DR contribue à l'amélioration de la **transparence** des processus scientifiques et permet d'intégrer les citoyens au fonctionnement même de ces derniers, comme l'illustre l'utilisation de *science citoyenne*[72] dans le cadre de projets scientifiques.
- La diffusion des DR intensifie une approche collaborative de la science en permettant de :

*« **Toucher une communauté internationale regroupée autour d'un domaine de recherche**, pour cette raison […], un dispositif de collecte et d'accès aux [DR doit] se mettre en place […] sous forme mutualisée, adossée à un réseau de chercheur »*

(Fayet 2013 p.5)

5.2 Obstacles

La littérature et les professionnels interviewés durant la réalisation de cette étude font état de plusieurs barrières se dressant sur l'ouverture des DR. Disposer d'une vue d'ensemble de ces derniers permettra d'en tenir compte dans le processus de réflexion menant à la proposition d'un modèle de diffusion. En effet :

- Un **usage frauduleux** de DR, qui peuvent être manipulées afin de modifier les résultats d'une recherche est envisageable, notamment dans un environnement très concurrentiel.

  Pour illustrer ce point, un chercheur a indiqué au DLCM, lors d'une enquête, qu'il ne souhaitait pas partager ses données, ni utiliser celles produites par

[71] « Etat de données qui […] ne subissent aucune altération […] et conservent un format permettant leur utilisation » (Wikipedia 2015)

[72] De l'anglais *Citizen Science*, le terme désigne les différentes méthodes qui permettent d'intégrer des amateurs aux processus de recherche. (Fondation Science et Cité 2015)

d'autres, car, n'ayant pas connaissance des conditions environnementales de l'expérience[73], il n'accorde pas sa confiance à ces résultats.[74]

- Des DR ne **disposant pas d'une documentation suffisante[75]** peuvent entraîner des erreurs et fausser les résultats d'une recherche visant à reproduire l'expérience de base. En tenant compte du fait qu'il est extrêmement complexe de réaliser une description exhaustive de l'ensemble d'un protocole de recherche, cela peut remettre en cause la pertinence du partage des DR.

  A titre d'exemple, une scientifique a indiqué, dans le cadre de l'enquête du point précédent, que lorsqu'elle réalise une expérience,[76] l'une des variables déterminantes identifiées par celle-ci ne figurait pas sur le protocole de recherche, ce qui compromettait la reproductibilité de l'expérience.[77]

- Des **limites technologiques** peuvent également remettre en cause l'utilité du stockage des DR en causant la **corruption des DR archivées**, la masse importante de données nécessitant des migrations fréquentes afin d'en garantir la pérennité.

  Ainsi, le *CERN* stocke une quantité énorme de données et effectue une migration chaque cinq ans, afin de maintenir les performances des technologies utilisées et de garantir leur pérennité. Lors de cette opération, un contrôle de l'intégrité des données[78] est réalisé. En raison de la quantité énorme de données dont dispose l'organisme, les logiciels réalisant ce check souffrent de dysfonctionnement.[79]

- Des **problématiques liées au droit d'auteur** peuvent également constituer un frein au partage des DR. Un conflit sur le Copyright d'un set de données peut en effet empêcher sa réutilisation, rendant de fait caduque la probité de son partage. La mise en place d'un système permettant de définir une période d'embargo devrait toutefois limiter la portée de cet obstacle.[80]

- L'utilisation potentielle des DR à des fins non souhaitées par les chercheurs, en raison d'un **manque de solidité des accords formels entre les partenaires académiques et privés** peut également constituer une barrière pour la diffusion de celles-ci. (Cragin et. al. 2010)

---

[73] Comprendre ici les conditions de température, le personnel effectuant le captage des données et son degré de conscience professionnelle, ou encore si la chaine de conservation des produits a été respectée

[74] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

[75] Comme l'indique Sturges et al (2015), dans l'article Reserach Sharing : Developping a Stakeholder-Driven Model for Journal Policies : le partage des DR implique des opérations de préparation qui visent à les rendre accessibles, le partage de données brutes pouvant se révéler complexe, car uniquement utilisable par leur producteur.

[76] Celle-ci consistait à faire effectuer des tâches à un sujet au moyen d'un ordinateur et enregistrer des facteurs tels que le mouvement des yeux ou ses temps de réaction.

[77] Voir réf. 67

[78] Au niveau du bit

[79] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

[80] Comme le développement des exigences politiques quant au libre partage des productions scientifiques abordé dans le cadre de l'introduction de cette étude.

5.3 Potentiel de réutilisation

Si le crédit accordé aux DR est en progression depuis qu'elles sont assemblées en collection et que leur qualité les rend plus à même de permettre une réutilisation dans le cadre de nouvelles recherches. (Kim 2013), leur nature hétérogène rend complexe la détermination de leur valeur pour un partage en vue d'une réutilisation future. La responsabilité des chercheurs dans la sélection de ce qui peut faire l'objet d'une diffusion entraîne une nécessité d'attribuer une forme de typologie des DR. Ceci vise à spécifier leurs caractéristiques, pour leur permettre d'opter pour les éléments qui représentent un potentiel de réutilisation optimal.

En effet, la valeur des données fluctue selon les disciplines scientifiques dans le cadre desquelles elles sont produites : « Dans certains cas, les données primaires auront de la valeur, dans d'autres seules les données intermédiaires devront être conservées de façon pérenne » (Bellamy 2014, p. 3). Il convient donc de distinguer les données primaires des intermédiaires.

5.3.1 Primaires

Les données primaires ou brutes sont des informations objectives enregistrées dans le cadre d'une recherche. Elles « ne sont pas encore le produit d'une analyse ou d'interprétation autre que le calcul ». (The Royal Society 2012, cité dans Gaillard 2014, p.17)

5.3.2 Intermédiaires

Les DR résultant d'une opération scientifique ne sont plus considérés comme primaires. Leur statut évolue suite à l'application d'une méthode de traitement[81] propre à la discipline dans le cadre de laquelle elles ont été produites. (Fayet 2013, p. 2)

Selon le CNRS (2015), ces données intermédiaires peuvent être déclinées en deux catégories distinctes :

- Traitées : « données : […] produites après calibration/étalonnage ou correction de données brutes ».
- Dérivées : « […] résumé ou […] représentation/vue spécifique des données [telles qu'une] agrégation, compilation […] »

[81] Il peut s'agir d'un processus de sélection visant à déterminer leur utilité dans le cadre du projet de recherche.

5.3.3 Finale

Selon le continuum des données proposé ci-dessus, le dernier stade atteint par les DR est celui de données publiées, celles-ci ont en effet été analysées et interprétées en fonction d'une hypothèse de recherche.

5.3.4 Typologie

En vue d'illustrer l'impact des conditions de récoltes et du type de DR sur leur potentiel de reproduction, un tableau inspiré du module d'introduction à la gestion et au partage de DR de l'*Inist* est présenté ci-dessous.

Tableau 1 : Typologie des DR

Type de données	Mode de récolte	Reproduction	Exemple
Observationnelle	Temps réel	Unique, pas de reproduction	Photographie astronomique
Expérimentale	Equipement de laboratoire	Reproductible, coûteuse	Chromatogrammes[82]
Computationnelle	Simulation / Modèle informatique	Reproductible à condition que le mode de récolte soit documenté	Modèle météorologique[83]
Dérivée / Compilée	Combinaison de données brutes	Reproductible, coûteuse	Fouille de texte
Référentielles	Collection de jeux de données	-	GenBank[84]

Source : INIST, 2014. Les standards de métadonnées. Inist.fr [en ligne]. [Consulté le 14.03.2016]. Disponible à l'adresse :

http://www.inist.fr/donnees/co/module_Donnees_recherche_32.html

[82] « La chromatographie est une méthode de séparation et de quantification de composé présents dans une phase homogène liquide ou gazeuse. […] [son résultat] le chromatogramme permet d'obtenir des informations qualitatives […] et quantitatives […] » (CHUV sd) sur l'analyse de la phase. (La chromatographie, aspects généraux), [S.d.])

[83] « Les modèles numériques de prévisions météorologiques permettent de prévoir à un instant T comment se présentera l'atmosphère selon les observations connues. Les prévisions issues de ces modèles sont établies grâce à de puissants ordinateurs pouvant traiter des centaines de millier d'informations » (Prévisonmeteo.ch 2016).

[84] GenBank est une banque de données des séquences génétiques du NIH, elle consiste en une collection de l'ensemble des séquences ADN disponibles publiquement (GenBank 2016)

- On peut, à la lecture de cet aperçu non exhaustif, constater les nombreuses variations relatives au type de DR. Celles-ci devraient être prises en compte lors de la mise en place de la recherche afin de déterminer si les données résultant d'un processus de recherche doivent être conservées et faire l'objet de traitements en vue de leur partage potentiel.
- Il est intéressant de noter ici que les évolutions technologiques des méthodes de récolte peuvent avoir un impact sur la décision de conservation. En effet, en génomique, certains chercheurs interrogés dans le cadre d'une enquête du projet *DLCM*, ont indiqué que la conservation d'une séquence de la mouche Tsé-Tsé n'avait pas d'intérêt car, si dix ans après, une telle donnée est nécessaire, la reproduction serait plus performante en raison de l'évolution des technologies.[85]

5.4 Culture du partage

Comme le laisse entendre la prise de position du *projet DLCM*, le partage des DR, varie fortement selon les disciplines scientifiques, qui sont confrontées à des obstacles, motivations et défis intrinsèques, que cela soit en termes de quantité ou de rythme de production des données ou en termes de capacité de financement et infrastructurelle. (Kim 2013 p.497-498)

En effet, « des sciences telles que l'astronomie et la physique tendent à disposer de pratiques standards de gestion et de partage des DR, » (Cragin et al. 2010, p. 4023) en raison d'une inclinaison culturelle de longue date pour la science collaborative, qui s'est d'abord cristallisée sous la forme d'une diffusion systématique des *preprints*[86] sur arXiv.org[87] (Brown 2001 cité dans Kim 2013, p. 500) qui s'est ensuite développé jusqu'à atteindre la création en 2014, pour le domaine HEP[88], d'un consortium visant à rendre l'ensemble des publications essentielles du domaine accessible en OA, SCOAP3[89].

De plus, ces disciplines peuvent s'appuyer sur des infrastructures dédiées, telles que des bases de données disciplinaires, ou des portails d'accès aux DR disponibles dans

[85] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.
[86] Etat d'un article scientifique avant que celui-ci ait été soumis aux processus de publication, tels que la revue par les pairs ou la mise en forme éditoriale.
[87] arXiv est une archive d'e-document, dans les disciplines physique, mathématique, informatique, etc.. (Cornell University library 2016).
[88] Acronyme désignant la physique des particules et des hautes énergies.
[89] Disponible à l'adresse : https://scoap3.org/

leur champ d'expertise, tel que le CERN Open data portal,[90] qui permet une visualisation des DR et l'utilisation d'outils sur les sets de données. (CERN 2014, 2015).

Les processus de recherche et leur conduite, dans le cadre des disciplines associées à la production de *small data*, telles que les sciences humaines, ne se reposent pas sur de telles infrastructures. Ces recherches sont « traditionnellement conduites par des hypothèses, menées par des chercheurs individuels »[91] (Cragin et al. 2010, p. 4024). Même si l'évolution des structures organisationnelles tend à entraîner une mutation des cellules de recherche vers une nature plus collaborative, que les auteurs considèrent comme un réseau communautaire. (Peterson 1993, cité dans Cragin et al. 2010, p. 4025)

Cette modification implique donc la mise en place d'infrastructures à l'échelle de ces entités de collaboration, exigeant la standardisation des méthodes de collectes, des processus de description. Ceci dans le but que la qualité des DR soit en adéquation avec la nécessité de partage ainsi engendrée et de fait, restreindre les différences de pratiques entre les champs de recherche.

Distinguer les DR en fonction du type de science dans lequel elles sont produites est une vision simplificatrice de la réalité scientifique. Les scinder dans une analyse n'est pas souhaitable, en raison de l'extrême hétérogénéité des pratiques et cultures de gestion et de partage des DR. En effet, on trouve certaines spécialités des sciences dites dures dans lesquelles les données peuvent être dissemblables et peu standardisées alors qu'on peut consulter des données normalisées dans certains champs de recherche en sciences humaines. [92]

La visibilité et l'utilité des résultats de certaines disciplines scientifiques pour la collectivité, comme par exemple la médecine, entraîne un déséquilibre de l'attribution des ressources financières. Celle-ci influence la production de DR elle-même et par là, le développement des modèles de partage : les champs de recherche disposant d'une large visibilité sont alimentés par des financements plus importants, permettant une croissance plus rapide des solutions de gestion et de partage des DR. (McGrath, 2013, p. 124)

---

[90] Disponible à l'adresse : http://opendata.cern.ch/

[91] Traduction par le rédacteur de : « hypothesis-driven research led by single principal investigator »

[92] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

Les disciplines scientifiques peuvent donc influencer les pratiques de recherche, en fonction notamment de la standardisation et de l'automatisation des processus de récolte, ou des moyens financiers liés aux investissements dans les spécialités intégrées en leur sein. Mais il existe peu de différences, en ce qui concerne les DR entre les sciences qualifiées de dures[93] ou de douces.[94]

> Au vu de l'hétérogénéité des DR et des pratiques disciplinaires, l'élément à retenir pour la mise en place d'une solution de diffusion est qu'il faut prendre en compte **les spécificités disciplinaires** comme point de départ de la réflexion. On peut considérer que le rapprochement des pratiques de gestion et de partage des DR entre les champs de recherche implique, qu'en dehors de l'**intensité de production des volumes**, celles-ci tiennent essentiellement aux **infrastructures**, aux **moyens financiers disponibles** et aux **convictions personnelles** des chercheurs (influencées par la perception du partage des DR de leur discipline d'expertise).

5.5 Gestion des données

Les acteurs du domaine de la recherche « ont réalisé [que les DR] doivent être traitées de manière responsable […], justifiable et économiquement rentable ». (Wood 2013, p. 79)

La mise en place de solutions de gestion consacrées aux DR, centrées sur leur cycle de vie de DR, répond à cette prise de conscience en identifiant l'ensemble des opérations nécessaires à la mise à disposition de DR de qualité pour en permettre le partage, la diffusion et la reproduction.

Motivée principalement par les exigences contractuelles des agences de financement et des journaux scientifiques publiant les résultats de recherche, comme indiqué en introduction du chapitre lié au partage des données, cette gestion est centrée par la préparation d'un plan de gestion des données (*DMP*). Un tel document, évolutif :

---

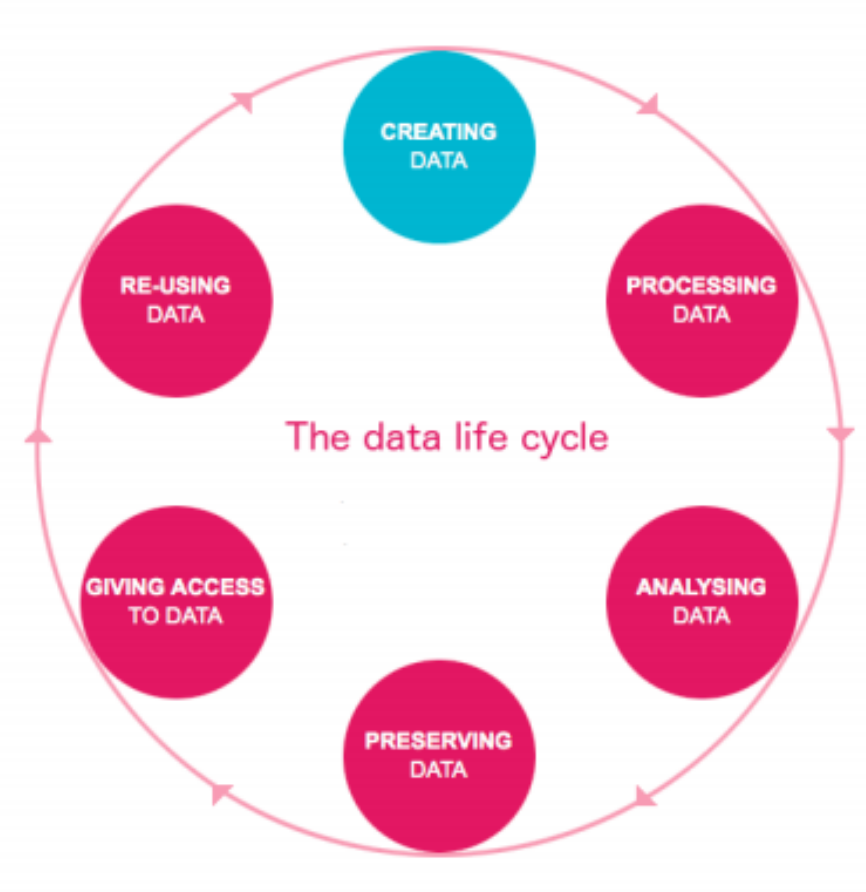
[93] Expression désignant dans un même ensemble les sciences de la nature : sciences de la vie et de l'environnement (agronomie, anatomie, botanique, etc.), les sciences de la Terre et de l'Univers (astronomie, météorologie, sismologie, etc.) et les sciences de la matière (chimie, physique) et sciences de formelles : algèbre, géométrie, informatique, logique, mathématiques et topologie. (Wikipedia 2016)

[94] Expression désignant dans un même ensemble, les sciences humaines et sociales, qui constituent un ensemble de disciplines étudiant des aspects de la réalité humaine (Wikipedia 2016)

« Ne devrait pas être une simple tâche administrative pour laquelle il suffit de copier des textes standardisés de modèles […] sans considérer ce qui est réellement nécessaire pour rendre le partage de données réalisable ».[95]

(Van den Eynden 2011, p. 6)

Figure 10 : Cycle de vie des DR



UK Data Archives 2015.Create and manage data. data-archive.ac.uk/ [en ligne]. [Consulté le 23 mai 2016]. Disponible à l'adresse : www.data-archive.ac.uk/create-manage/life-cycle)

Les différentes informations qui devront être intégrées dans le *DMP*, selon les informations recueillies dans le livrable du *projet DLCM*,[96] [97]concernent à la fois :

[95] Traduction par le rédacteur de : « A data management plan should not be thought of as a simple administrative task for which standardised text can be pasted in from model templates, [...], or without considering what is really needed to enable data sharing.

[96] Voir chapitre 3.1.3, p. 16

[97] La liste ci-dessous est inspirée de la data management check-list créée par : (EPFL, ETHZ 2016)

- La planification à réaliser lors de la préparation de la soumission du projet pour l'obtention d'un financement, qui vise à déterminer les parties prenantes impliquées dans le projet et leurs responsabilités.
- L'identification des DR, qui vise à identifier leur type et leur reproductibilité[98], leur origine si elles sont réutilisées, ainsi que leurs caractéristiques techniques (volume, gestion des versions, standards utilisés, etc.)
- La préparation et la sélection, qui permet d'indiquer selon quels critères et avec quels outils les DR seront évaluées et sélectionnées.
- La description, élément essentiel pour la qualité des données, qui vise à constituer les informations contextuelles qui permettront la réutilisation des DR partagées.
- La définition des formats, qui vise à contrôler la compatibilité de ceux-ci avec les standards de la discipline et leur interopérabilité.
- Les informations sur la conservation, telle que la durée, le volume nécessaire, les conditions de destruction éventuelle, la gestion des sauvegardes, etc.
- Les considérations éthiques à prendre en compte, en cas de diffusion de données sensibles, par exemple.
- Les informations sur les éléments de propriété intellectuelle et sur les éventuelles restrictions juridiques d'usage des DR.
- Les éléments concernant le partage des DR, tels que la présence d'une éventuelle période d'embargo, le niveau du partage (public, privé).
- Les informations spécifiques à la gestion à long terme des DR.

5.6 Publication

Pour illustrer l'importance de la publication des DR, on peut comparer dès 2001 « le taux de citation des articles publiés en ligne (7.03), par rapport à ceux qui ne le sont pas (2.74) » (Lawrence 2001 cité dans Heidorn 2008, p. 290), ce qui souligne l'aspect fondamental de l'accessibilité des DR en ligne.

Si la publication des DR fait « déjà partie, dans certaines disciplines, des canons de rédaction d'un article scientifique […] dans la partie Materials & methods) » (Fayet 2013, p. 2), deux ans après sa publication, les chances d'accéder aux données qu'il contient diminue de 17% par an, en raison de l'obsolescence des outils de stockage ou des liens (Vines et. al. 2014).

Comme l'illustre la pyramide des données ci-dessous, il existe plusieurs types de publications des DR, qui font toutes l'objet de solutions existantes. Cette illustration

[98] Selon la typologie définie au chapitre 5.3.4, p.30

permet également de définir quels prestataires correspondent aux types de publications qu'elle contient.

5.6.1 Formes

Figure 11 : Formes de publication des DR

(Inist 2014. Adapté de Reilly et. al 2011. Disponible à l'adresse : http://www.inist.fr/donnees/co/module_Donnees_recherche_20.html)

- Les journaux académiques de qualité intègrent actuellement les DR dans leur contenu, ce qui inclut leur description et les méthodes de récoltes (par le biais de la partie méthodologique).
- Les DR peuvent faire l'objet d'une description via les fichiers supplémentaires en lien avec les articles publiés.
- Les DR peuvent être conservées sur des centres de données ou des entrepôts de données. Ceux-ci peuvent être disciplinaires[99] ou généralistes.[100]
- Les DR peuvent faire l'objet d'une description par le biais d'un data paper.[101]

[99] A titre d'exemple GenBank: http://www.ncbi.nlm.nih.gov/genbank/, Protein Data Bank: http://www.rcsb.org/pdb/home/home.do

[100] A titre d'exemple Dryad : datadryad.org/, Zenodo : https://zenodo.org/, Figshare : https://figshare.com/, dataverse : http://dataverse.org/, archives institutionnelles.

[101] Cette solution fait l'objet d'un développement plus précis au chapitre suivant.

5.6.2 Déficits[102]

Comme indiqué lors de l'université d'automne 2015, organisée par la HEG Genève, sous la direction du professeur Schneider (Nosek & Bar-Anan cités dans Schneider 2015), la publication scientifique souffre de plusieurs déficits.

- Seuls les résultats positifs sont publiés, ce qui entraîne une perte de DR qui peuvent potentiellement être réutilisées dans le cadre de nouvelles recherches.
- Le modèle actuel souffre d'une certaine lenteur, provoquée notamment en raison du processus de revue par les pairs et les différentes corrections qu'il engendre.
- Le modèle repose principalement sur les chercheurs, ce qui peut générer des décisions erronées sur la conservation des DR.
- Le modèle actuel est figé. Une fois que les articles sont publiés, il n'y a pas moyen d'effectuer des modifications sans engendrer la publication d'un nouvel article.

5.6.3 Evolution

On constate également que le terme *partage* appliqué aux DR, laisse la place à *publication* dans les échanges entre les spécialistes du domaine. Cette modification de terminologie s'appuie, selon les auteurs « sur la conviction que les [DR] doivent disposer d'un statut de publications académiques ». (Kratz, Strasser 2014 p. 3)

Comme le présente le modèle graphique ci-dessous, afin d'être considérée comme publiée, une DR doit répondre à certains critères. Elle doit être disponible de manière pérenne, ce qui passe par l'attribution d'un espace de stockage. Cela peut être soit dans un dépôt de données ou sur un site web dédié à un projet de recherche. De plus, elle doit être accompagnée des éléments descriptifs qui permettent d'en assurer la reproductibilité, des métadonnées et une documentation. A cela doit être ajouté la notion de complétude des sets de DR. En effet :

> *« Seule une publication complète permet une compréhension claire des intentions de l'auteur, du projet de recherche et des options choisies pour la forme de l'expérimentation, ainsi que sur les décisions prises durant son déroulement »*[103]
>
> *(Gorgolewski et al. 2013, p. 5)*

---

[102] Les déficits de la publication scientifique déclinés dans cette partie doivent faire l'objet d'une réflexion dans la sélection d'une solution de diffusion. En effet, celle-ci devra pouvoir proposer des réponses visant à libérer la publication des DR du carcan ainsi défini.

[103] Traduction par le rédacteur de : « as only complete data publication provides a clear understanding of the intended experimental design and decisions made ».

- Le développement des entrepôts de données voit deux logiques s'affronter : institutionnelle et disciplinaire. Si leurs desseins sont similaires (dépôt de l'ensemble des productions académiques sur une solution de stockage), elles diffèrent sur l'organisation du découpage. La première vise à regrouper l'ensemble des productions académiques pouvant être rattaché à l'institution sur le même espace de stockage. Elle présente l'avantage de simplifier la gestion institutionnelle. La seconde quant à elle, est organisée autour des **solutions de partage disciplinaires** et elle est **plus en adéquation avec les besoins des chercheurs**.[104]

La nécessité d'attribuer une valorisation professionnelle aux producteurs de DR implique que celles-ci puissent faire l'objet d'une citation. Pour rendre possible cette pratique, l'attribution d'un identifiant pérenne est une obligation. Ce dernier permettra de relier les DR aux publications dont elles font l'objet ou pour lesquelles elles ont été utilisées.

- Le digital Object Identifier (*DOI)*, désigne un identifiant[105] numérique pérenne, attribué à un objet (Doi.org 2014, p. 3), qui constitue l'une des réponses[106] possibles à cette nécessité d'identification. Mais celle-ci reste toutefois partielle, car d'autres éléments sont à prendre en compte pour garantir l'accès, comme le maintien des solutions d'hébergement des DR.[107]

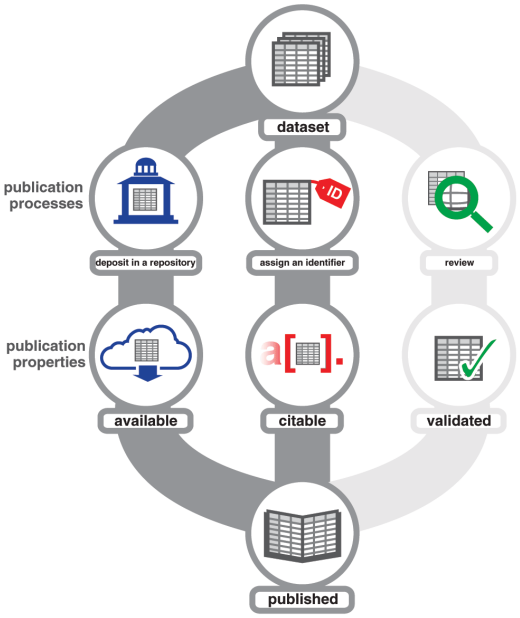
[104] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

[105] « Les identifiants sont des éléments indispensables pour faciliter l'accès aux [DR] et leur intégration dans les services d'information ». (Inist 2014)

[106] Les Archival Ressource Key (*ARK)*, constituent une alternative intéressante car ils sont gratuits et maintenus par des bibliothèques, des archives, des musées (confluence.ucop.edu 2016) qui en termes de pérennité des infrastructures sont un gage de sécurité. Voir réf. 104

[107] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

Figure 12 : Publication des données



Source : KRATZ, John et STRASSER, Carly, 2014. Data publication consensus and controversies. [version 3 ; referrees : 3 approuved]. *S1000 research*. 2014 3 : 94, pp.1-21

5.6.4 Validation

Le dernier axe du schéma de publication présenté ouvre la question de la validation scientifique des DR publiées. Ce processus consiste à « évaluer les méthodes de collectes des données, leur plausibilité et leur valeur de réutilisation » (Kratz, Strasser 2014, p.7)[108]

En effet, on observe une différence. Les données publiées au sein même des articles scientifiques font l'objet d'un processus de revue par les pairs, lors de l'examen de la plausibilité des résultats présentés par la publication. En revanche, les DR faisant l'objet d'une publication via un entrepôt de données, ne font pas toujours l'objet d'un contrôle en lien avec l'objet de l'étude qu'elles concernent.

Si la diffusion des DR, par le biais d'articles scientifiques ayant pour objet les sets de données permet d'appliquer le principe de revue par les pairs, en raison des similitudes éditoriales que présente ce format de publication avec le modèle traditionnel, des solutions sont développées également pour permettre de garantir la qualité des DR

[108] Ce qui implique que la personne en charge de la vérification scientifique du set de données dispose d'une expertise dans le champ de recherche à l'origine de leur création.

publiées seules. Ces processus de validations peuvent prendre la forme de guides pratiques ressemblant aux directives des journaux scientifiques, certains dépôts de données ayant jusqu'à exiger le contrôle des DR déposées par un organisme tiers. (Kratz, Strasser 2014, p. 7-8)

La publication séparée des DR implique enfin l'apparition d'une seconde forme de validation, en lien avec les caractéristiques techniques des sets de données. Ces vérifications consistent à contrôler la complétude du set et la cohérence entre les éléments descriptifs associés et son contenu. (Kratz, Strasser 2014, p. 7)

- Cette scission du processus de validation permet de **répartir la responsabilité des vérifications** entre les différents acteurs. Les compétences nécessaires à leur réalisation n'étant pas identiques.

5.6.5 Peer review

La question de la revue par le pairs et les débats qu'elle engendre actuellement doit être relevée dans le cadre de cette étude. En effet, le *post publication review* se développe dans de nombreuses disciplines (McGrath 2013, p.120), comme l'illustre le fonctionnement du journal F1000 Research, dont la validation est centrée autour des commentaires post publication.

Les partisans du processus de revue par les pairs avant publication, doivent faire face à ceux favorables aux vérifications post publications. Ce dernier courant, provoqué par le partage des DR, permet « à des chercheurs externes d'examiner les données d'une manière que leurs producteurs n'avaient pas envisagée, ce qui peut soulever des erreurs éventuelles » (Gorgolewski et. al. 2013 p. 4-5)

- Ce qui implique de prendre en considération l'ensemble des éléments intervenant après la publication des DR, entraînant la mise en place de solutions de gestion des versions des DR et le dynamisme des productions académiques post publication. Cela doit être un élément central de la solution de diffusion des DR résultant de ce travail.

5.6.6 Open Access

Les prises de position fortes, à la fois des instances européennes et suisses en faveur de la publication OA pour l'ensemble des projets qu'ils financent, exigent une prise en compte de cette forme de publication dans les processus de réflexion engagé pour la réalisation de ce travail.

Si l'OA entraîne les conséquences indiquées à la réf. 10, il s'agit également d'un modèle d'affaires pour de nombreux journaux[109], qui vise à rendre libre l'accès aux publications scientifiques en finançant leur parution au moyen des articles processing charges (*APC)*.[110] Il consiste également un modèle de diffusion[111], qui vise à rendre accessibles l'ensemble des publications scientifiques, via la mise à disposition de leur *preprints* sur les archives ouvertes institutionnelles.[112]

- Cet essor constitue un facteur essentiel qu'il convient d'intégrer dans les propositions faisant suite à cette étude[113]. Il est toutefois utile de ne pas perdre de vue l'importance de distinguer les DR publiques de celles produites dans le cadre de recherches financées par des fonds privés.

[109] Sous son aspect *Gold Road.*
[110] Les charges éditoriales sont facturées directement à l'auteur de la publication.
[111] Sous son aspect *Green Road.*
[112] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.
[113] L'unique restriction d'accès pour de telles publications étant de disposer d'une connexion web, richesse qui ne saurait être atteinte par le modèle traditionnel. (Bennett 2013, p. 109)

6. Data paper

Comme indiqué dans la référence 105, ce chapitre est consacré au développement d'une solution de diffusion des DR, se basant sur le concept de data paper. Le principe appliqué dans ce genre de production scientifique est que, contrairement aux articles classiques, tout n'est pas contenu dans le rapport scientifique. En effet, celui-ci se limitant à une description du set de données et à l'attribution de métadonnées. Le matériau faisant l'objet de la publication est disséminé sur un espace de stockage externe, par exemple un data repository.[114]

C'est cette forme de publication qui fait l'objet d'un développement particulier dans le cadre de cette étude. Il est important de prendre en considération que, malgré son potentiel, elle ne permet pas de répondre à l'ensemble des problématiques soulevées par la publication des DR, n'étant :

> *« Ni une solution ultime et complète pour l'ensemble des problèmes pour le partage et la réutilisation des [DR] et dans certains cas […] considérée comme à l'origine de certaines fausses attentes dans la communauté de recherche ».*
>
> *(Parsons, Fox 2013, cité dans Candela et al. 2015)*

De plus, certains scientifiques considèrent que favoriser ce type de modèle n'est pas une solution rentable et préconisent une amélioration du système existant. (Huang, Hawkins, Qiao 2013, p. 6)

6.1 Définition

Forme récente de publication scientifique centrée sur les DR, ce type d'article se rencontre sous différentes appellations : data article, dataset paper, data notes et sous son intitulé le plus courant, le data paper.

- Selon Chavan et Penev, dans leur article the Data paper : a mechanism to incentivize data publishing data publishing in biodiversity science, les data paper peuvent être définis comme :

  « Des publications scientifiques d'un document interrogeable par ses métadonnées, décrivant un set de données accessible en ligne, […] publié

[114] Voir réf. 116.

conformément aux standards des pratiques académiques ».[115]

(Chavan & Penev, 2011 cité dans Schneider Prongué 2015)

- Les auteurs de l'enquête Data Journal a Survey, précisent cette définition comme suit :

« Article qui décrit un set de données et donne des informations sur le quoi, où, pourquoi, comment et qui de la production des données. Celui-ci contient un lien (DOI) vers le repository sur lequel est déposé le set de données et le journal qui publie l'article ne les héberge pas, ce qui garantit que même en cas d'accès restreint de l'article, le set de données reste accessible librement ».

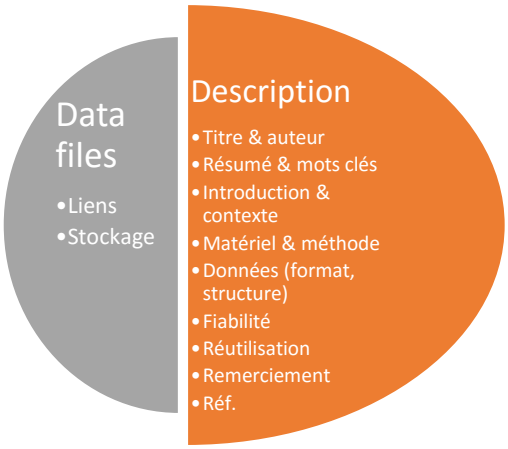
(Candela et al. 2015, p. 2)

- Cette description peut être complétée par la spécificité suivante : ce type d'article ne présente jamais d'analyses ni de résultats en lien avec le contexte de création des DR. (Gorgolewski, Margulies, Milham 2013 p. 2)
- On peut noter que ce type de publication se retrouve à la fois dans certains journaux scientifiques classiques[116] et sont la source d'apparition des *data journals*[117], des publications spécialisées qui ne contiennent que ce type de productions scientifiques.

6.2 Structure

La structure d'un article de ce type est présentée ci-dessous, selon les informations recueillies sur le document de Dedieu (2014) : Rédiger et publier un data paper dans une revue scientifique en 5 points.

Figure 13 : Représentation graphique du Data paper



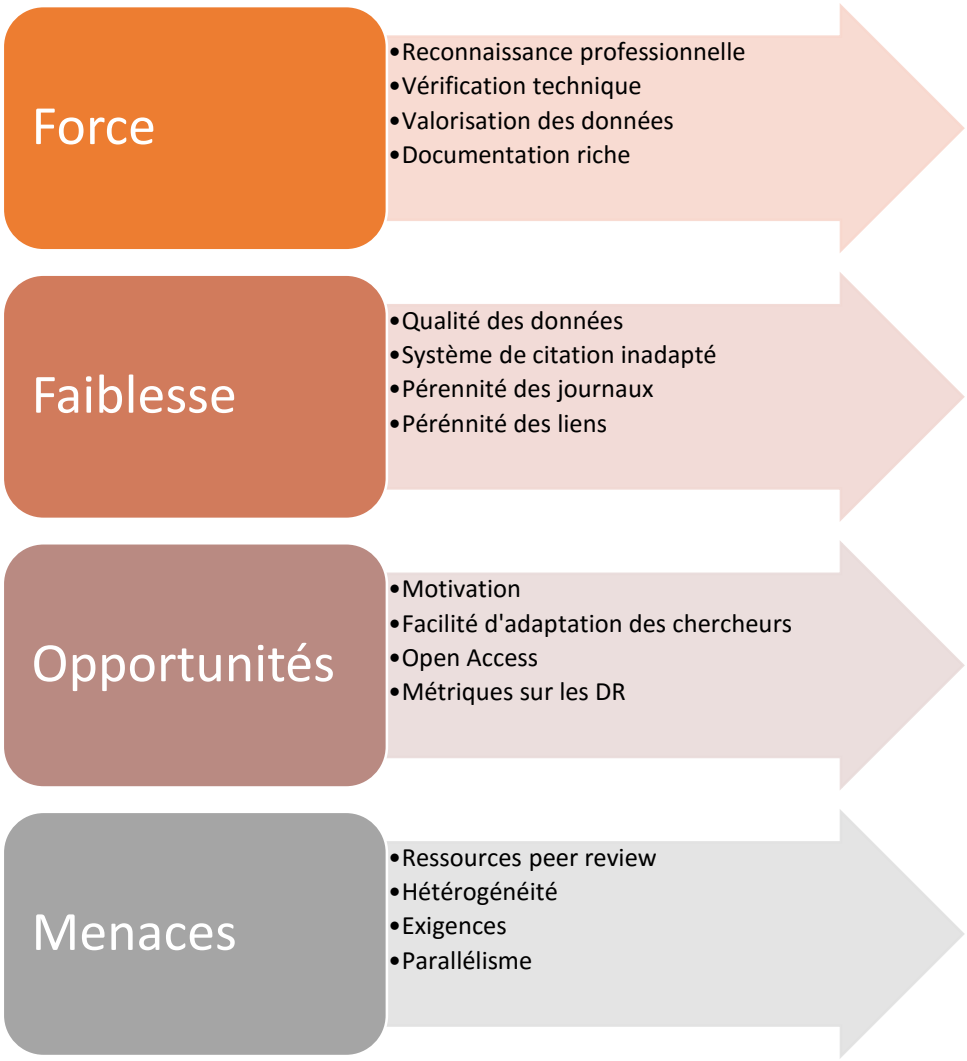
[115] Traduction par le rédacteur de : « A scholarly publication of a searchable metadata document describing a particular online accessible data set, […] published in accordance to the standard academic practices ».

[116] A titre d'exemple : F1000 Research : Disponible à l'adresse : http://f1000research.com/

[117] A titre d'exemple : Gigascience. Disponible à l'adresse : http://gigascience.biomedcentral.com/

6.3 Analyse

Figure 14 : SWOT pour les DR



6.3.1 Force

6.3.1.1 Reconnaissance professionnelle

La littérature s'accorde sur le fait que, la mise en place de ce type de solution de diffusion permet d'apporter une amélioration importante de la reconnaissance professionnelle apportée aux chercheurs pour les travaux en lien avec la gestion des DR. Ces productions scientifiques peuvent alors faire l'objet de citations.

- Ce qui implique que l'utilisation d'une telle solution de diffusion pour les DR pourrait servir de catalyseur permettant l'augmentation de ces types de pratiques au sein de la communauté scientifique.

6.3.1.2 Vérification technique

La séparation des DR de l'article qui les décrivent, permet également de vérifier les aspects techniques y relatif.

6.3.1.3 Valorisation des données

Il permet également de « montrer l'originalité et la portée d'un set de données […] c'est-à-dire leur potentiel de réutilisation dans le cadre d'autres recherches » (Dedieu 2016).

6.3.1.4 Documentation riche

La nature des data papers donne de nombreuses informations contextuelles sur les DR, de plus la présence de métadonnées enrichies permet de simplifier notamment les questions de propriété intellectuelle. (Huang, Hawkins, Qiao 2013, p 5)

6.3.2 Faiblesse

6.3.2.1 Qualité des données

La revue par les pairs est actuellement centrée sur les articles et non sur les données qui en sont l'objet. (Candela 2014, p. 14)

- Ceci entraîne la nécessité d'adapter le processus afin qu'il réponde aux exigences d'une diffusion performante des DR. Les éléments permettant de garantir la validité des données ne correspondent pas aux critères de qualité d'une publication classique.

6.3.2.2 Système de citation

S'ils permettent de valoriser à la fois le travail des chercheurs et les données elles-mêmes, les data papers souffrent de la considération, par les éditeurs que le système de citation construit sur la base du fonctionnement de la publication traditionnelle est suffisant (Candela 2014, p.14)

6.3.2.3 Pérennité des journaux

Le journaux scientifiques, même gratuits et libres d'accès, sont propriété d'entreprise et leur maintien n'est pas garanti, par exemple en cas de faillite, ce qui peut entraîner, à moyen terme, la disparition de DR.[118]

[118] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

6.3.2.4 Pérennité des liens

Le fait de disposer d'identifiants pérennes ne constitue pas, en lui seul, une garantie suffisante, si l'on ne tient pas compte des tâches dédiées à la gestion de l'archivage à long terme des DR.

6.3.3 Opportunité

6.3.3.1 Motivation

Les data papers sont considérés comme des articles scientifiques à part entière et correspondent, ainsi, au besoin de publication des chercheurs dans un contexte où leur rendement est jugé sur le nombre d'occurrences qu'ils produisent.

6.3.3.2 Adaptation des chercheurs

Le modèle de communication scientifique traditionnel, dispose de la confiance des chercheurs, ce qui constitue une opportunité pour la diffusion par le biais des data papers, au vu des similitudes des deux formules. (Anders, Elvidge 2013, p. 54)

6.3.3.3 Open Access

Les journaux publiant des data papers « ont opté en majorité pour un modèle [OA], reposant généralement sur le *Gold road* »[119] (Candela 2014, p.14) ce qui, associé aux éléments identifiés plus avant dans cette étude, représente un terreau fertile pour l'accessibilité des DR.

6.3.3.4 Métriques

Un tel modèle de diffusion permet d'envisager la mise en place de métriques performantes sur la publication des DR. (Gorgolewski, Margulies, Milham 2013, p. 2) Cela peut encourager les chercheurs à effectuer les démarches de partage des DR, celles-ci devenant mesurables.

6.3.4 Menace

6.3.4.1 Ressources peer review

La mise en application d'un tel système de publication peut mettre en difficulté les processus de revue par les pairs, qui disposent de ressources limitées. (Huang, Hawkins, Qiao 2013, p. 5)

[119] Traduction par le rédacteur de : « have embraced the [OA] model by generally reling on the […] Gold […] » model »

6.3.4.2 Hétérogénéité

Il n'existe pas de standard pour la publication de tels articles, ce qui peut constituer un obstacle majeur pour l'interopérabilité des DR. Cette menace peut concerner l'ensemble des parties prenantes du partage, tant il existe de différences dans les pratiques de gestion et de publication.

6.3.4.3 Exigences

Cette forme de publication nécessite une étroite collaboration entre les différents acteurs du domaine de la recherche, car elle implique à la fois les éditeurs, les gestionnaires commerciaux de base de données, les institutions et les chercheurs. Cette collaboration entre des entités dont les besoins en termes de gestion des DR sont hétérogènes peut représenter un équilibre fragile.

6.3.4.4 Parallélisme

Cette forme de diffusion peut être limitée dans le cas d'une parution d'un article scientifique classique disposant d'une description des données et dont les résultats sont basés sur les mêmes DR que celles décrites dans le data paper. Cela engendre, de fait, des questions sur la redondance de celles-ci.

7. Modèle de diffusion

7.1 Synthèse des entretiens

Développer des scénarios adaptés au contexte helvétique implique de disposer d'informations sur les spécificités réelles de ce dernier. Une approche incluant le contexte global de l'actualité et son influence sur les pratiques des parties prenantes du domaine de la recherche scientifique suisse constitue une première étape qui permet de définir les acteurs et projets principaux en lien avec celui-ci.

Une approche approfondie, menée par le biais d'entretiens semis directifs, permet d'apporter le point de vue de professionnels dont les activités sont ancrées dans l'utilisation ou le développement de solution en lien avec les DR.

Les entretiens semi directifs ont été réalisé avec la participation de :

- **Madame Eliane Blumer**, coordinatrice du projet DLCM et responsable du développement de services en lien avec les DR pour l'UNIGE. Son implication dans les processus mis en œuvre dans le cadre du programme de Swissuniversities et sa connaissance des instances académiques, sont en parfaite adéquation avec les réflexions menées.
- **Monsieur le Docteur Patrick Ruch**, professeur HES, responsable de la filière information documentaire de la HEG Genève. Sa connaissance du contexte de la HES-SO et sa qualité de directeur de recherche du projet BITEM et son statut de chercheur, en font un interlocuteur de choix pour la réalisation de cette étude.
- **Monsieur Jean-Blaise Claivaz**, coordinateur au sein de la DIS de l'UNIGE, responsable de l'archive ouverte et des questions relatives à l'OA, ainsi que de la gestion des DR pour l'institution est également impliqué dans les processus de collaboration entre les organismes engagés dans le projet DLCM. Il complète ce panel de professionnels en apportant une vision bibliothéconomique.

Une partie des informations récoltées durant ces interviews ont été utilisées, par le rédacteur, à des fins d'approfondissement ou d'illustration dans les différents chapitres de ce travail. Le solde, qui constitue l'essentiel des éléments dont l'importance est primordiale dans le développement, est décliné ci-dessous.

Les réponses des intervenants ont été classées selon deux thèmes : quelles sont les particularités de la recherche scientifique helvétique et quels sont les critères essentiels que doit respecter une solution de diffusion des DR.

7.1.1 Particularités helvétiques

Il est plus probant de présenter les particularités du domaine de la recherche suisse pour les intervenants de manière détaillée, ceci afin de disposer de l'ensemble de leurs raisonnements sur la question. Les éléments identifiés ci-dessous sont un amalgame de leurs interventions individuelles.

- Le découpage linguistique n'est pas spécifique à la Suisse.[120] La séparation des instances de hautes écoles, selon les voies académique et professionnelle constitue un découpage administratif, qui, de fait, n'a que peu d'incidences.
- Le mécanisme de financement de la recherche est l'un des éléments les plus spécifiques du contexte helvétique. En effet, les agences en charge de la promotion de la recherche scientifique, qui constituent les organes décisionnels sont peu nombreuses. Cela permet une rapidité de fonctionnement intéressante. En effet, chacune des recommandations de la FNS peut conduire à une modification structurelle importante de la manière de conduire les recherches.
- La relative opulence de la recherche suisse entraîne le désavantage le plus spécifique. L'importance des moyens entraîne le développement de solutions locales, dans le cadre de projets de recherche.
- La Suisse est une sorte d'îlot, à l'image de son positionnement sur l'échelle de la politique internationale. Ses organismes académiques collaborent à de nombreux projets de grande envergure, sans totalement en faire partie. Cette situation exige une adaptation de l'appareil scientifique helvétique pour lui permettre de travailler de manière interdisciplinaire et internationale.

[120] On retrouve en effet ce type de caractéristiques sur d'autres territoires, comme le Canada.

7.1.2 Critères essentiels du modèle

Cette partie de la synthèse est présentée sous forme de tableau, afin de mettre en perspective les éléments considérés par les intervenants comme indispensable au bon fonctionnement d'un modèle de diffusion des DR.

Tableau 2 : Critères essentiels

Critère	Intervenant	Information
Standardisation	P. Ruch	S'inspirer des communautés qui disposent des standards les plus élevés. Mise en œuvre de pratiques standardisées, que cela soit dans l'utilisation des métadonnées, des citations, etc.
	E. Blumer	Description normalisée et définition d'un minimum acceptable, mais sans être limitée pour permettre une description plus approfondie lorsque cela est nécessaire.
Accessibilité	P. Ruch, E. Blumer	*OA* : Cette forme de publication est un bon indicateur du partage libre des données.
	E. Blumer	Interopérabilité avec un maximum de systèmes. Interdisciplinarité. Simplicité d'utilisation
Politiques	P. Ruch	*Nationale* : modèle fort qui oblige les chercheurs à participer au développement des DR, à leur dépôt dans des entrepôts de données et à publier les résultats de recherche en OA. *Communautaire* : modèle qui repose sur les pratiques disciplinaires et dont les buts sont similaires. Institutionnelle : sur les exigences des instances académiques quant à la gestion et au partage des DR.
Infrastructures	P. Ruch	Disposer des financements nécessaires à la mise en place d'espace de stockage.
Structure des données	P.Ruch	Disposer de descripteurs, index matière
Chercheurs	J.-B. Claivaz	L'impulsion en faveur du partage des DR doit venir des chercheurs.
Rentabilité	J.-B. Claivaz	La diffusion devient rentable une fois les infrastructures utilisées.
	E. Blumer	Constituer une offre concurrentielle.

7.2 Scénarios

Le fonctionnement du modèle de diffusion présenté par la figure 16, n'est pas impacté par les scénarios faisant l'objet d'une présentation ci-dessous. En tenant compte des critères essentiels identifiés lors de cette étude, il paraît indispensable que les différentes propositions reposent sur une structure commune.

7.2.1 Structure commune

Afin de pouvoir mettre en place un modèle de diffusion centré sur le concept de data paper, il est fondamental de pouvoir s'appuyer sur un cadre définissant le type de DR que les scientifiques doivent partager.[121] Cela doit se faire en tenant compte des particularités liées aux pratiques institutionnelles et disciplinaires. Ceci permettra de garantir une reconnaissance professionnelle en lien avec les travaux effectués sur les DR, pour amener une prise de conscience, par les chercheurs, de leur valeur ajoutée.

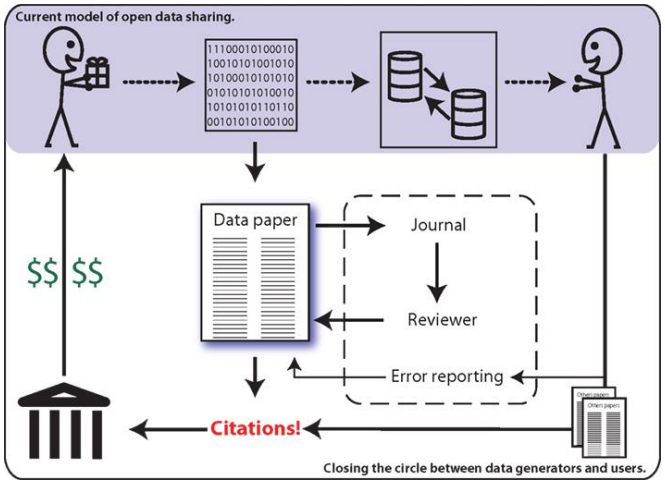
De plus, une politique claire des instances suisses, publiques et privées de financement de la recherche, en faveur du partage des DR et recommandant leur diffusion par le biais des data papers. En outre ce cadre doit tenir compte de celles mises en place par les bailleurs de fonds européens et internationaux, dans le cadre d'une collaboration étroite entre les acteurs de l'écosystème helvétique de la recherche, qu'ils soient institutionnels ou commerciaux.

Il faut également être attentif à la mise en place d'une gestion des DR maîtrisée et basée sur le cycle de vie des données et à la standardisation des pratiques de création de métadonnées et de citations. Le but est d'assurer une accessibilité et une reproduction optimale. Mais il est aussi nécessaire de veiller à disposer d'infrastructures de stockage de qualité, permettant de veiller au maintien de l'intégrité des DR et de garantir la pérennisation des identifiants permettant de lier les productions scientifiques entres elles.

Enfin, il apparait important de définir l'OA comme mode de publication privilégié, exiger une qualité de base minimale mais évolutive, intégrer la possibilité de corriger les DR après leur publication et mettre en place les outils nécessaires à leur structuration.

[121] Big data & Small data.

Figure 17 : Modèle de diffusion basé sur les data papers



Source : Gorgolewski, Marglies, Milham 2013. frontiersin.org [en ligne]. [Consulté le 13 mai 2016]. Disponible à l'adresse : http://www.frontiersin.org/files/Articles/39226/fnins-07-00009-HTML/image_m/fnins-07-00009-g001.jpg

7.2.2 Plateforme web

Cette première solution vise à créer un point d'accès central aux data papers produits par les chercheurs affiliés aux institutions académiques suisses. Elle doit permettre d'accéder aux articles sur les sets de DR disséminés sur les différentes publications, spécialisées ou généralistes et de les relier avec les DR elles-mêmes, quelles que soient leurs infrastructures de stockage.

L'accumulation des infrastructures locales, liée à la richesse identifiée dans les spécificités helvétiques est à prendre en considération dans cette option. Elle autorise en effet la mise en œuvre d'une solution de type cloud inter infrastructurelle,[122] qui augmente la visibilité de l'ensemble, car pris en considération de manière individuelle, les dépôts de données et archives institutionnelles ne disposent pas d'une masse suffisante pour devenir un partenaire incontournable.[123]

Ce modèle présente l'avantage de s'appuyer sur les publications existantes et de ce fait permet de tenir compte de l'ancrage disciplinaire des cultures de partage et d'éviter la

[122] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

[123] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

redondance. Un découpage en silo peut constituer une première forme de filtre permettant un accès plus aisé aux utilisateurs.

Le modèle permet d'identifier les DR faisant l'objet d'une description détaillée et de déterminer leur qualité, par la mise en place de métriques qui peuvent même faire l'objet de comparaison, pour établir une échelle de pertinence, voire l'attribution d'une récompense pour les articles les plus performants.

7.2.2.1 OAI-PMH

La création d'une interface d'accès unique à l'ensemble des articles disponibles par ce biais, permise par l'utilisation d'un moissonnage des métadonnées via le protocole OAI-PMH permet de répondre au besoin de simplicité d'accès. Cette version du scénario implique de conserver la responsabilité de validation scientifique des datas papers et des sets de DR qu'ils décrivent.

7.2.2.2 Web collaboratif

Une seconde version, permettant le libre dépôt des data papers, créés et mis en forme selon une politique éditoriale définie par une association d'éditeurs partenaires permet, quant à elle, de mettre en place une formule d'open post publication peer-review et d'intégrer des spécificités du web 2.0 dans le fonctionnement de la plateforme. Celles-ci permettent d'envisager d'intégrer les Altmetrics dans les outils de valorisation des DR.

7.2.3 Collaboration avec un journal existant

Ce scénario est orienté sur une collaboration étroite entre les instances académiques suisses et un ou plusieurs journaux spécialisés dans la publication de data paper. Afin de tenir compte des spécificités disciplinaires et des pratiques de partage, on peut envisager de conserver un découpage traditionnel.

La collaboration avec un data journal généraliste est envisageable, à des fins de simplification. Toutefois, les DR de certains champs scientifiques nécessitent l'usage d'informations spécifiques pour être réutilisables, ce qui constitue un argument en faveur d'un scénario tenant compte de plusieurs titres.

Ce scénario peut, dans une certaine mesure, faire l'objet d'un questionnement quant à son intérêt réel. En effet, le développement de la publication scientifique traditionnelle et de l'intégration de plus en plus importante des DR au sein des articles scientifiques et de la position de certains chercheurs qui pensent que la mise en place d'un nouveau

modèle de diffusion ne fait que complexifier le fonctionnement de la science et qu'une adaptation du modèle d'origine est une solution plus pertinente.

7.2.4 Création d'une publication

La dernière option envisagée est la création, par les instances académiques, d'un Swiss Data Journal, qui pourrait offrir aux institutions une position éditoriale en sus de celle d'organismes producteurs de résultats scientifiques.

Elle présente une obligation de disposer d'un élan d'implication des chercheurs qui sont au cœur des processus de développement des pratiques scientifiques. L'origine de l'OA l'illustre parfaitement : sans la participation de scientifiques intéressés par le développement d'une telle solution, aucun article ne sera proposé, ni revu, laissant les presses académiques sans matière.[124] Cela engendrerait un déséquilibrage important des investissements nécessaires au fonctionnement du modèle de diffusion.

8. Recommandations

Base de toute collaboration fructueuse, la communication entre les différents acteurs du domaine scientifique et au sein même des institutions est, selon le rédacteur, un facteur essentiel qui devrait être maîtrisé. En effet, le développement parallèle de projets tels que le DLCM et Openresearchdata.ch devrait pouvoir reposer sur un échange riche et ouvert sur les questions liées à la gestion des DR.

En somme, il faudrait prolonger l'ouverture des DR au fonctionnement structurel de l'écosystème scientifique, à l'image d'un débat télévisuel entre des astronomes de renom, lors duquel les interactions les plus fréquentes sont des hochements de tête affirmatifs et ou la parole est prise pour compléter le propos en cours et non pour le contrer.

Le rédacteur rejoint ici l'opinion soulevée lors des entretiens réalisés dans le cadre de ce travail. Il considère que les systèmes les plus avancés dans la culture du partage, comme l'astronomie ou les HEP, devraient servir de base empirique au développement des solutions de partage et de valorisation du travail effectué dans le cadre de la gestion

[124] Entretien avec Jean-Blaise Claivaz, coordinateur de l'archive ouverte pour la division d'information scientifique (DIS) de l'UNIGE, Genève, 19 mai 2016.

des DR[125]. Le rédacteur précise qu'une adaptation de celle-ci aux spécificités des disciplines devrait être envisagée.

L'utilisation du modèle de publication scientifique traditionnel pour développer des solutions correspondant aux spécificités des DR, est ancrée dans la routine pratique des chercheurs, qui évoluent en terrain connu, comme l'illustre les débats concernant la pertinence de l'usage d'un modèle de validation des publications scientifiques de toute nature avant ou après la publication de celles-ci.

La position du rédacteur sur la question repose sur la rapidité de disponibilité des DR et l'opportunité de leur appliquer des principes liés à l'intelligence collective, tel que la science citoyenne ou le crowdsourcing. Il relève également la nécessité de pouvoir gérer les corrections qu'entraîneront immanquablement la mise à l'épreuve d'une reproduction des expériences de recherche. Ceci recommanderait une solution de validation post publication.

Si les DR sont des données ayant valeur de preuve, il pourrait être intéressant, une fois les infrastructures crées, de considérer les données qui ne disposent pas de ce statut et de les intégrer dans le processus de gestion car, il est possible, qu'une fois combinées entre elles, celles-ci mutent à leur tour en quelque chose de probant.

Enfin, la simplicité des outils associés aux DR devrait permettre une utilisation modulaire, reposant sur la possibilité d'enrichir le contenu et les descriptions[126]. Son but est d'intégrer l'ensemble des utilisateurs dans le développement de la qualité des DR, voire l'ajout d'un aspect ludique à leur utilisation, se basant sur les fonctionnalités du web 2.0[127], selon une suggestion d'Eliane Blumer.

---

[125] Entretien avec le professeur Patrick Ruch, responsable de la filière information documentaire de la Haute école de gestion de Genève, Carouge, 19 mai 2016.

[126] Voir des solutions telles que tDAR, un entrepôt de données qui ne met en place que la validation technique des DR et offre la possibilité aux utilisateurs de les enrichir (Kratz & Strasser 2014 p.8).

[127] A titre d'exemple, on peut citer la solution Open Context (Kratz & Strasser 2014 p.8), qui permet un classement des DR avec une notation.

9. Synthèse

Selon les informations recueillies durant la réalisation de cette étude, on peut donc considérer que pour permettre de disposer de DR vérifiables, reproductibles et valides, véritables racines d'une amélioration nécessaire des processus de recherche, Il faut être capable de remplir plusieurs conditions.[128]

L'impulsion est d'abord politique. Elle repose essentiellement sur les décisions des instances en charge de la recherche et de l'innovation qui, sous l'influence d'une modification des sociétés civile et scientifique, se positionnent en faveur d'une science ouverte. En découle la mise en place d'une législation rendant obligatoire la transparence des activités financées par les fonds des collectivités publiques. Cela entraîne une nécessaire adaptation des processus de gestion des DR par les acteurs de l'écosystème de la recherche sur plusieurs aspects. Tant au niveau des agences de financement en charge du développement de celle-ci, qui mettent en place des politiques exigeant le partage des DR, qu'à celui des éditeurs commerciaux, dans la mise en place de leurs conditions contractuelles.

Ceci amène ensuite les institutions académiques à redéfinir leur politique de gestion, partage et diffusion des DR, ce qui entraîne des modifications importantes sur le rôle et les tâches des chercheurs et des bibliothèques académiques.

Afin de pouvoir répondre à cette évolution, il est essentiel de tenir compte de la nature même des DR, des cultures de partage disciplinaires et de la transformation du fonctionnement des processus de recherche, de par l'évolution importante des projets interdisciplinaires, l'intensification du volume de données produites. Et ce, quel que soit le champ de recherche concerné.

La première condition, sur laquelle repose l'ensemble d'une science ouverte est l'accessibilité des données. Celle-ci est construite sous la forme d'une infrastructure permettant de maintenir l'intégrité et la pérennité des données.

La seconde, réalisée par le biais d'une gestion du cycle de vie des DR, consiste à maîtriser l'ensemble des opérations ayant attrait à ces dernières, de leur création à leur destruction, en passant par la sélection des DR dont la conservation est intéressante.

[128] A des fins d'innovation, de respect des implications éthiques qu'engendre les investissements publics et de respect de droits fondamentaux tels que le droit à la vie privée.

Afin de garantir la reproductibilité et la possibilité de réutilisation des données, la mise en œuvre d'une vérification, à la fois technique et scientifique, est une étape obligatoire. Celle-ci peut s'effectuer par le biais des professionnels en charge des opérations d'archivage des DR et d'une revue par les pairs pour le second, qu'elle soit éditoriale ou basée sur un modèle d'intelligence collective. Ceci permet de valoriser professionnellement les tâches réalisées par les chercheurs dans ce contexte.

C'est dans le cadre de cette démarche de valorisation et de diffusion que s'ancrent les réflexions ayant pour résultat la rédaction de ce travail.

On observe qu'un modèle de diffusion des DR centré sur le concept de data paper ne constitue pas une solution permettant de répondre à l'ensemble des difficultés et des défis intrinsèques à une gestion performante des données. En revanche, elle peut apporter une plus-value certaine, si une attention particulière est maintenue sur les menaces et faiblesses de son fonctionnement. Cela notamment dans la qualité descriptive des DR diffusées et dans l'amélioration de l'impact positif du travail y relatif pour les chercheurs, sur qui, en finalité repose le modèle.

Les spécificités du domaine de la recherche helvétique, notamment le nombre restreint d'acteurs décisionnels permettent d'envisager une capacité d'adaptation rapide à cette évolution systémique, si une communication entre les parties prenantes et que les efforts de chacune d'entre elles se cristallisent autour de cet objectif. Ceci tout en permettant à l'écosystème scientifique suisse de maintenir son image de qualité et de performance.

Enfin, divers aspect permettent d'envisager de nouvelles perspectives d'évolution vers une meilleure utilisation du cœur de la science, les DR. Ceci regroupe la généralisation de l'adoption de la publication OA et le développement des TIC, la structuration des DR et la mise en place de solutions permettant le data mining ou le texte mining, comme l'application des principes du web sémantiques.

10. Conclusion

Selon Habert et Salaün (2015) cités dans Collet-Thireau et Thomas (2015) : « le document n'existe plus, mais est reconstruit en permanence « à la volée », via des algorithmes ». On peut s'appuyer sur son propos et considérer que le développement d'un modèle de diffusion des DR, centré sur le concept de data paper, même s'il dispose d'un certain nombre d'avantages, ne saurait répondre à lui seul aux enjeux induits par une science ouverte.

En effet, comme les éléments ressortant de cette étude de conceptualisation l'indiquent, la nouvelle unité de fonctionnement de l'appareil scientifique est la donnée.

Au terme de ce travail, Il apparaît au rédacteur de ces lignes, que les data papers ne constituent que l'une des étapes d'un développement qui, à terme, devrait permettre d'extraire du sens de données structurées. De plus il observe qu'une publication ne constituerait « qu'une seule image dans un plus vaste écosystème de partage des [DR] » (Parson & Fox 2013 cité dans Kratz & Strasser 2013, p.8).

Si les acteurs systémiques du domaine scientifique anglo-saxons et européens ont développé une conviction profonde du bien-fondé de cette mutation structurelle de la science, le microcosme que constitue la recherche helvétique, tente de les rejoindre, par nécessité. A l'image d'un train lancé à vive allure, il entraîne dans son sillage la transformation du paysage de la recherche suisse !

Bibliographie

ADBS, 2012. Vous avez dit "curation" ? (1) Définition, historique des pratiques, outils et usages. Adbs.fr [en ligne]. 13 mars 2012. [Consulté le 16 mai 2016]. Disponible à l'adresse :

http://www.adbs.fr/vous-avez-dit-curation-1-definition-historique-des-pratiques-outils-et-usages-115668.htm

ANDERS, Catie et ELVIDGE, Liz, 2013. Creative communication in a « publish or perish » culture : can postdocs lead the way ? In : The future of scholary communication. SHORLEY, Deborah et JUBB, Michael [éditeur]. London : Facet Publishing, pp.51-62. ISBN 978-1-85604-817-0

BELLAMY, Dorothée, 2014. L'ouverture des données de la recherche.*inshea.fr* [en ligne]. Juillet 2014. [Consulté le 16 mai 2016]. Disponible à l'adresse :

http://www.inshea.fr/sites/default/files/Les%20donn%C3%A9es%20de%20la%20recherche.pdf

BORGMAN, Christine L., 2012. The comundrum of sharing research data. *Journal of the american society for information science and technology*, 63 (6), pp. 1059-1076. DOI : 10.1002/asi.2634

CANDELA, Leonardo, et al., 2015. Data journal : a survey. Journal of the Association for Information Science and Technology. 1 February 2015. p.1. DOI 10.1002/asi.23358

CASSELY, Jean-Laurent, 2013. Big Data : ce n'est pas la taille qui compte. *Slate.fr* [en ligne]. 13 mai 2013. [Consulté le 16 mai 2016]. Disponible à l'adresse :

http://www.slate.fr/economie/72361/big-data-small-data

CHAVAN, Vishwas and PENEY, Lyubomir, 2011. The data paper a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics.15 December 2011. Vol. 12, no. Suppl 15, p. S2. DOI : 10.1186/1471-2105-12-S15-S2. PMID : 22373175

CLAIVAZ, Jean-Blaise, DIEUDE, Aude et KRAUSE Jan Brice, 2015. Données de la recheche : quèsaco ? *Hors-Texte*, 2015, pp. 24-29

CLEVELAND, William S. 2001. Data Science : An Action Plan for Expanding the Technical Areas of the Field of Statistics. *ISI Revue*. 26 avril 2001. Vol. 69- pp 21-26

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 ; p. CIDCR1-CIDCR75. doi.org/10.2481/dsj. OSOM13-043

COLLINS, Ellen, 2013.The Social media and scholarly communications : the more they change, the more they stay same. In : The future of scholary communication. SHORLEY, Deborah et JUBB, Michael [éditeur]. London : Facet Publishing, pp.89-102. ISBN 978-1-85604-817-0

CONSEIL DE L'UNION EUROPENNE, 2012. Communication de la commission au parlement européen, au conseil, au comité économique et social européen et au comité des régions : Pour un meilleur accès aux informations scientifiques : dynamiser les avantages des investissements publics dans le domaine de la recherche. *register.consilium.europa.eu* [en ligne]. 23 juillet 2012. [Consulté le 17.05.2016]. Disponible à l'adresse :

http://register.consilium.europa.eu/doc/srv?l=FR&f=ST%2012847%202012%20INIT

CONFEDERATION SUISSE, 2013a. La recherche et l'innovation en Suisse. *sbfi.admin.ch* [en ligne]. 01 janvier 2013. [Consulté le 16.03.2016]. Disponible à l'adresse :

http://www.sbfi.admin.ch/themen/01367/?lang=fr

CONFEDERATION SUISSE, 2015. *2017 à 2020.* [fichier PDF]. Commission pour la technologie et l'innovation CTI. Avril 2015.

Programme pluriannuel

CONFEDERATION SUISSE, 2016. Les origines de la CTI. *Kti.admin.ch* [en ligne]. 27 avril 2016. [Consulté le 30.04.2016]. Disponible à l'adresse :

https://www.kti.admin.ch/kti/fr/home/ueber-uns/die-kti--entstehung--leitbild-und-ziele-.html

CONFEDERATION SUISSE, 2013b. Programmes fédéraux d'encouragement à la coopération scientifique bilatérale avec des pays prioritaires. *sbfi.admin.ch* [en ligne]. 01 janvier 2013. [Consulté le 16.03.2016]. Disponible à l'adresse :

http://www.sbfi.admin.ch/themen/01370/01390/01420/index.html?lang=fr

CORNELL UNIVERSITY LIBRARY, 2016. arXiv.org. *arxiv.org* [en ligne]. 29.06.2016. [Consulté le 29.06.2016]. Disponible à l'adresse :

http://arxiv.org/#

CRAGIN, Melissa H., PALMER Carole L. CARLSON, Jacob R. et WITT, Michael, 2010. Data sharing, small science and institutional repositories. Phylosophical transactions of the royal society. A366, pp. 4023-428. DOI : 10.1098/rsta.2010.0165

DCC, 2016a. About the DCC [en ligne]. [Consulté le 1er juin 2016]. Disponible à l'adresse : http://www.dcc.ac.uk/about-us

DCC, 2016b. DCC Curation Lifecycle Model. dcc.ac.uk [en ligne]. [Consulté le 1er juin 2016]. Disponible à l'adresse : http://www.dcc.ac.uk/resources/curation-lifecycle-model

DEDIEU L. 2014. Rédiger et publier un data paper dans une revue scientifique en 5 points. *coop-ist.cirad.fr* [en ligne]. [Consulté le 29.05.2016]. Disponible à l'adresse : http://coop-ist.cirad.fr/content/download/5506/40945/version/4/file/CoopIST-Rediger-et-publier-un-data-paper-octobre2014.pdf

DLCM, 2016. First outcomes. dlcm.ch [en ligne]. 26 mai 2016. [Consulté le 30.05.2016]. Disponible à l'adresse :

http://dlcm.ch/2016/05/26/first-outcomes/

ENSSIB, 2013. Métadonnées. enssib.fr [en ligne]. 21 novembre 2013. Dernière mise à jour le 26 mars 2014. [Consulté le 29.05.2016]. Disponible à l'adresse :

http://www.enssib.fr/le-dictionnaire/metadonnees

ELSEVIER, 2016. What is peer review?. Elsevier.com [en ligne]. [Consulté le 15 mai 2016.] Disponible à l'adresse : https://www.elsevier.com/reviewers/what-is-peer-review

ETH ZURICH, EPFL. 2016. *Data Management Checklist* [fichier PDF]. Version 1.0. Mars 2016.

DLCM outcome

EU2016.NL. Tous les articles scientifiques européens en libre accès à partir de 2020. *français.eu2016.nl* [en ligne]. 27 mai 2016. [Consulté le 28.05.2016]. Disponible à l'adresse :

http://francais.eu2016.nl/a-la-une/actualites/2016/05/27/tous-les-articles-scientifiques-europeens-en-libre-acces-a-partir-de-2020

EUROPEAN COMMISSION, 2013. *Guidelines on Open Access to Scientific Publication and Research Data in Horizon 2020*. V1.0, p. 14. 11 décembre 2013

FAYET, Sylvie, 2013. Données de la recherche les mal-nommées. *Document numérique.* 25 novembre 2013. Edition scientifique, Gestion des connaissances (KM). pp. 18

FERGUSON, Adam R. et al. 2014. Big data from small data : data-sharing in the 'long tail' of neuroscience. *Nature.com*. [en ligne]. 28 octobre 2014. [Consulté le 13.04.2016]. Disponible à l'adresse :

http://www.nature.com/neuro/journal/v17/n11/full/nn.3838.html

FLAHAULT, François. Bien commun. *Participation-et-démocratie.fr* [en ligne]. Juin 2013. [Consulté le 23.03.2016]. Disponible à l'adresse : http://www.dicopart.fr/es/dico/bien-commun.

FMI, 2008. Pour une relance budgétaire ciblée : Réflexions de Dominique Strauss-Kahn. *Imf.org* [en ligne]. 30 janvier 2008. [Consulté le 13.03.2016]. Disponible à l'adresse :

http://www.imf.org/external/french/np/vc/2008/013008f.htm

FNS, 2016a. Partager le savoir : Open Access pour tous. *Snf.ch* [en ligne]. 10 mai 2016. [Consulté le 15.05.2016]. Disponible à l'adresse :

http://www.snf.ch/fr/pointrecherche/newsroom/Pages/news-160510-communique-de-presse-partager-le-savoir-open-access-pour-tous.aspx

FNS, 2016b. Portrait [en ligne]. 01 janvier 2016. [Consulté le 15.02.2016]. Disponible à l'adresse :

http://www.snf.ch/fr/leFNS/portrait/Pages/default.aspx

FNS, 2016c. Règlement des subsides [en ligne]. 01 janvier 2016. [Consulté le 15.02.2016]. Disponible à l'adresse :

http://www.snf.ch/SiteCollectionDocuments/fns-reglement_des_subsides_2015_f.pdf

GAILLARD, Rémi, 2014. De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ? *enssib.fr*. [en ligne]. Janvier 2014. [Consulté le 15.03.2015]. Disponible à l'adresse :

http://www.enssib.fr/bibliothequenumerique/documents/64131-de-l-open-data-a-l-open-research-data-quellespolitiques-pour-les-donnees-de-recherche.pdf

GORGOLEWSKI, Krzysztof J., MARGULIES, Daniel S., MILHAM, Michael P.. Making data sharing count : a publication-based solution. *Frontiers in Neuroscience*. 06 février 2013. Volume 7, pp. 1-7

GRAY, Jim. 2006. eScience : The Next Decade Will be Exciting.*signallake.com* [en ligne]. [Consulté le 17.03.2016]. Disponible à l'adresse :

http://www.signallake.com/innovation/JGrayETH_E_Science.pdf

HEIDORN, P. Bryan, 2008. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*. Vol. 57, No. 2, Fall 2008, pp. 280-299

HUANG, Xiaolei, HAWKINS Bardford A. et QIAO Gexia, 2013. Biodiversity data sharing : Will Peer-Reviewed Data Papers Works ?. *Bioscience.* Vol. 63, n°1, pp. 5-6

INIST, 2014. Les standards de métadonnées. Inist.fr [en ligne]. [Consulté le 14.03.2016]. Disponible à l'adresse :

http://www.inist.fr/donnees/co/module_Donnees_recherche_32.html

JACQUEMOT-PERBAL, M.-C. et COSSERAT F., 2015. Gestion et diffusion des données de la recherche. *inist.fr* [en ligne]. 11 juin 2015. [Consulté le 23.04.2016]. Disponible à l'adresse :

http://www.inist.fr/IMG/pdf/urfistrennes_20150616.pdf

JANSON, Nathalie. 2015. Le cas du Franc Suisse : quels Enseignements pour l'Euro. *Researchgate.net* [en ligne]. Octobre 2015. [Consulté le 17.03.2016]. Disponible à l'adresse :

https://www.researchgate.net/publication/282856729_Le_Cas_du_Franc_Suisse_quels_Enseignements_pour_l'Euro

COLLET-THIREAU, Katell, THOMAS, Jean-Paul. 2015. Big Data et Open Data : quel impact pour les professionnels de l'information ? *I2D-Information, données & documents*. 22 décembre 2015. N°4. pp 9-10

KIM, Jeonghyun, 2013. Datasharing and its implactions for academic libraries. *New Library World*. 12 juin 2013. Vol. 114, No. 11/12, pp. 494-506

KLING, R. & MCKIM G., 2000. Not just a matter of time : field differences and the shaping of electronic media in supporting scholarly communication*. Journal of the American Society for Information Science*. 14 décembre 2000. 51. pp. 1306-1320

KRATZ, John et STRASSER, Carly, 2014. Data publication consensus and controversies. [version 3 ; referrees : 3 approuved]. *S1000 research*. 2014 3 : 94, pp.1-21

La chromatographie, aspects généraux), [S.d.]. *Doping.chuv.ch* [en ligne]. [Consulté le 4 janvier 2015]. Disponible à l'adresse :

http://www.doping.chuv.ch/files/presentation-gc-fr.pdf

LA PRÉSIDENCE NÉERLANDAISE DE L'UE 2016. Plan d'action européen pour la science ouverte. *français.eu2016.nl* [en ligne]. 04 avril 2016. [Consulté le 15.04.2016]. Disponible à l'adresse :

http://francais.eu2016.nl/a-la-une/actualites/2016/04/05/plan-d%E2%80%99action-europeen-pour-la-science-ouverte

LEBIGDATA.FR, 2016. Qu'est-ce que le big data. *Lebigdata.fr* [en ligne]. [Consulté le 17.05.2016]. Disponible à l'adresse :

http://www.lebigdata.fr/definition-big-data

Loi fédérale sur la protection des données (LPD). Admin.ch [en ligne]. 1er janvier 2014. [Consulté le 10.03.2016]. Disponible à l'adresse :

https://www.admin.ch/opc/fr/classified-compilation/19920153/index.html

Loi fédérale sur le principe de la transparence dans l'administration (LTrans). Admin.ch [en ligne]. 19 août 2004. [Consulté le 10.03.2016]. Disponible à l'adresse :

https://www.admin.ch/opc/fr/classified-compilation/20022540/index.html

Loi fédérale sur la protection des données (LPD). Admin.ch [en ligne]. 1er janvier 2014. [Consulté le 10.03.2016]. Disponible à l'adresse :

https://www.admin.ch/opc/fr/classified-compilation/19920153/index.html

Loi fédérale sur le droit d'auteur et les droits voisins (LDA). Admin.ch [en ligne]. 1er janvier 2011. [Consulté le 10.03.2016]. Disponible à l'adresse :

https://www.admin.ch/opc/fr/classified-compilation/19920251/

Loi sur l'information (LInfo). rsv.vd.ch [en ligne]. 24 septembre 2002. [Consulté le 10.03.2016]. Disponible à l'adresse :

http://www.rsv.vd.ch/dire-cocoon/rsv_site/doc.fo.html?docId=5507

MCGRATH, Mike, 2013.The changing role of the journal editor. In : The future of scholary communication. SHORLEY, Deborah et JUBB, Michael [éditeur]. London : Facet Publishing, pp.117-129. ISBN 978-1-85604-817-0

MINISTERE DE L'EDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE, 2013. horizon2020.gouv.fr [en ligne]. [Consulté le 15.02.2016]. Disponible à l'adresse :

http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html

OCDE, 2007. Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publiques. *Oecd.org* [en ligne]. [Consulté le 10.03.2016]. Disponible à l'adresse :

http://www.oecd.org/fr/sti/sci-tech/38500823.pdf

OFFICE FEDERALE DE LA STATISTIQUE, 2012. Financement public de la recherche en Suisse : Evolution 2000 – 2010. [Fichier PDF]. Septembre 2012.

OPENDATA.CH, 2013. Manifeste : ouverture des données publiques en Suisse : le manifeste. *Opendata.ch* [en ligne]. Consulté le 11.03.2016]. Disponible à l'adresse :

http://fr.opendata.ch/manifeste/

OPENRESEARCHDATA, 2016. Openresearchdata.ch [en ligne]. [Consulté le 30 mai 2016]. Disponible à l'adresse :

http://openresearchdata.ch/

PRÉVISONMETEO.CH, 2016. Prévisionmeteo.ch [en ligne]. [Consulté le 29 juin 2016]. Disponible à l'adresse :

http://www.prevision-meteo.ch/meteo/modele-numerique

POLINE, Jean-Baptiste, et al.. Datasharing in neuroimaging reserach. *escholarship.org* [en ligne] 05 avril 2012. [Consulté le 11.04.2016]. Disponible à l'adresse :

http://escholarship.org/uc/item/01k5x5tg

QUAGLIA, Daniela, 2015. Data-publishing for better Science. Blogs.jobs.ac.uk [en ligne].[Consulté le 11.04.2016]. Disponible à l'adresse :

https://blogs.jobs.ac.uk/science-and-technology/2015/10/31/data-publishing-better-science/

RDA 2016. The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data. *Europe.rd-alliance.org/* [en ligne] Disponible à l'adresse : http://europe.rd-alliance.org/

SCHNEIDER René et PRONGUE Nicolas, 2015. Data papers as a catalyst for Open Access and the long tail of research data. *Zenodo.org* [en ligne]. [Consuté le 13 novembre 2015]. Disponible à l'adresse :

http://zenodo.org/record/30761#.V3KzNI6DNmA

SCHNEIDER René, 2015. Données de la recherche, module IV : Publication et services [fichier PPT].

Support de cours

STURGES, Paul, BAMKIN, Marianne, ANDERS, Jane, et al., 2014. Research Data Sharing: Developing a Stakeholder-Driven Model for Journal Policies. Journal of the Association for Information Science and Technology, 66(12)2445-2455, 2015. DOI: 10.1002/asi

SWISSUNIVERSITIES, 2015a. Swissuniversities. *Swissuniversities.ch* [en ligne]. 01.01.2015. 13.06. 2016. [Consulté le 05.03.2016]. Disponible à l'adresse :

https://www.swissuniversities.ch/fr/organisation/

SWISSUNIVERSITIES, 2015b. *White paper for a Swiss Information Provisioning and Processing Infrastructure 2020*. [fichier PDF]. 25 juin 2015. [Consulté le 03.10.2015]. Disponible à l'adresse :

https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/FR/UH/SUK_P-2/WhitePaper_V1.1-FR.pdf

TENOPIR, Carol et al., 2011. Data Sharing by scientists : Pratices and Perceptions. *Plos One*. Juin 2011. Vol. 6. Issue 6. e21101. pp. 1-21

VAN DEN EYNDEN, Veerle, CORTI, Louise, WOOLLARD, Matthiew, BISHOP, Libby et HORTON, Laurence, 2011. Managing and sharing data [en ligne] Colchester : UK Data Archive. ISBN 1-904059-78-3. Disponible à l'adresse :

http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

VINES, Thimothy H, 2014. The Availability of Research Data Declines Rapidly with Article Age. *cell.com* [en ligne]. [Consulté le 23 mai 2016]. Disponible à l'adresse :

http://www.cell.com/current-biology/abstract/S0960-9822(13)01400-0

UK Data Archives 2015.Create and manage data. data-archive.ac.uk/ [en ligne]. [Consulté le 23 mai 2016]. Disponible à l'adresse : www.data-archive.ac.uk/create-manage/life-cycle)

UNIGE, 2008. Copyright et Open Access. *Archive-ouverte.unige.ch* [en ligne]. Octobre 2008. Juin 2016. [Consulté le 05.04.2016]. Disponible à l'adresse :

http://archive-ouverte.unige.ch/pages/copyright_open_access

UNIL, 2014. Archives des savoirs : De la gestion des données de recherche vers une gestion des données pour la recherche. *Unil.ch* [en ligne]. 16 octobre 2014. Journées des archivistes des universités et hautes écoles suisses. [Consulté le 17.04.2016] Disponible à l'adresse :

https://www.unil.ch/uniris/files/live/sites/uniris/files/documents/public/JDAU14_6_UNIL_Gestion_Donnees_de_recherche.pdf

file:///C:/Users/sabak/Downloads/Mehrjahresprogramm_A4_fr_150504.pdf

Crise bancaire et financière de l'automne 2008. *Wikipedia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 3 mars 2016. [Consulté le 17.04.2016] Disponible à l'adresse :

https://fr.wikipedia.org/wiki/Crise_bancaire_et_financi%C3%A8re_de_l%27automne_2008

Intégrité (cryptographie). *Wikipedia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 30 novembre 2015. [Consulté le 17.04.2016] Disponible à l'adresse :

https://fr.wikipedia.org/wiki/Int%C3%A9grit%C3%A9_(cryptographie)

Sciences dures. *Wikipedia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 9 mars 2016. [Consulté le 17.04.2016] Disponible à l'adresse : https://fr.wikipedia.org/wiki/Sciences_dures

WILLET, Perry, 2016. ARK (Archival Resource Key) Identifiers. *confluence.ucop.edu* [en ligne]. Dernière mise à jour le 28 juin 2016. [Consulté le 29.05.2016]. Disponible à l'adresse :

https://confluence.ucop.edu/display/Curation/ARK

Annexe 1 : Entretien semi-directif, P. Ruch

Entretien réalisé avec Monsieur le Docteur Patrick Ruch, professeur HES, responsable de la filière information documentaire de la HEG Genève, le 19.05.2016. Les réponses de l'interlocuteur ont été retravaillées et classées selon les thèmes identifiés pour permettre de disposer d'éléments pertinent de comparaison entre les différents entretiens.

Acteurs

Les acteurs concernés par la gestion des données de la recherche ne sont pas uniquement institutionnels. Les entreprises dont la tâche est de gérer les données de la recherche ou de les stocker en sont un exemple.

Les recherches financées par le secteur privé peuvent engendrer des collaborations avec des entreprises commerciales, qui ne partagent pas les missions des institutions académiques.

Critères essentiels

S'inspirer des communautés qui ont atteint le niveau le plus élevé de standardisation pour mettre en place un système de diffusion des données de la recherche optimal est le procédé à suivre. Les étapes suivantes sont à suivre :

1) La publication des données en Open Access devrait être acceptée, car c'est un bon indicateur de partage libre des données. Cela permet de régler la problématique de l'accès en libérant la diffusion d'un pare feu financier.

2) Le problème des journaux est que même s'ils sont gratuits et propriété d'une entreprise privée, ils peuvent être amenés à disparaître, ce qui entraîne la seconde étape. Disposer de politiques communautaires, disciplinaires ou nationales pour forcer le dépôt institutionnel. Certains pays exigent déjà, de manière assez libérale, que la totalité des projets financés par les deniers publics soit déposé dans ce type d'archives.

Ce qui est notamment illustré par les prises de position de certaines agences de financements du Royaume-Uni, par exemple le Welcome Trust, un acteur important de la recherche du pays pour les exigences étatiques ou par les pratiques en cours dans le domaine de la physique des particules et des hautes énergies qui dépose l'ensemble de ses preprints sur ArXiv.org. Ces deux modèles permettre de garantir l'accès aux données.

Le premier est un peu plus puissant, car il exige que les données soient publiées en Open Access et qu'elles soient publiques se libérant ainsi de ce que les chercheurs feront des données, car quelques soient leurs options de publication, les conditions devront être respectées. L'utilisation du second devra faire l'objet de choix, soit opter pour la voie « Gold », soit « Green ».

Il faudra disposer de l'archive qui au départ, stockera des articles scientifiques, des publications, puis in fine, l'ensemble des productions académiques. En théorie, l'unique limitation et un problème de financement d'infrastructures.

3) Une fois cette étape réalisée, on pourra, à condition de développer des standards, des descripteurs, un index matière, etc. pour stocker des données structurées, alors réaliser l'ensemble des opérations qui découle de la disponibilité des données : curation, science citoyenne, reproductibilité, etc.

Data Paper

L'objectif de n'importe quel data paper est de décrire les processus de collecte, de développement de la méthode, pour permettre de donner une vision d'ensemble sur ce qui a été réalisé sur les données et de déterminer les conditions dans lesquelles elles ont été recueillies. Ça devrait même être l'objectif de tous les articles scientifiques. On ne peut pas imaginer une publication qui s'affranchirait d'une description des processus permettant de reproduire les expériences traitées.

On devrait présenter une section méthode qui serait probablement la partie principale de l'article et qui devrait être aussi précis que possible, car même si le travail effectué est rigoureux il est possible que lors de la reproduction d'une expérience, une étape ou un paramètre peut faire l'objet d'une description insuffisante. Il est en effet problématique, au niveau cognitif, de se mettre à la place de la personne qui sera amenée à reproduire une expérience.

L'objectif principal d'une publication scientifique est de contenir l'hypothèse, les données et leur documentation, que cela soit dans l'article lui-même, ou dans les annexes techniques, disséminés sur des espaces de stockage, l'essentiel étant de faire le lien entre les éléments composant cet écosystème. Le principe de ce dernier est que tout n'est pas contenu dans le rapport, une partie des éléments peut être disséminée dans des données supplémentaires stockées dans une archive institutionnelle.

La plupart des journaux qui s'appuient sur une politique éditoriale rigoureuse sont déjà devenus des data papers ou sont en passe de le devenir.

Les milieux académiques suisses ou autres, on besoin d'une solution qui ait la possibilité de mettre à disposition la donnée, de la conserver etc. Est-ce que cela passe par la mise en place de quelque chose qui s'appellera data paper, je l'ignore. C'est un problème de vision : soit on se positionne en faveur des data papers et tout le monde effectue la mutation, soit les écosystèmes évoluent vers quelque chose qui ne disposera pas de cette appellation, mais qui apportera la même chose.

L'important est de distinguer les données : publiques, à valeur de preuve, etc., c'est cela qui est fondamental. Ensuite, une partie des instruments qui permettront ce changement de paradigme seront le web 2.0, la science citoyenne et d'autres outils qui pourront selon leur spécificité être utilisés sur les différents types de données.

Données

Une donnée documentée au point de pouvoir la rendre indépendante de leur contexte de création n'existe probablement pas. Elle n'en est jamais indépendante, elle est toujours liée à un contexte d'acquisition et elle a un contenu intrinsèquement lié à un domaine scientifique. Il convient de la décrire, mais elle sera toujours spécifique. L'objectif n'est jamais de rendre la donnée complétement indépendante, mais de la mettre à disposition. Mais il faudra une compréhension assez fine de celle-ci pour pouvoir l'utiliser. On ne peut pas se passer d'une certaine expertise.

La diffusion des données de la recherche vise la reproductibilité, qui a plutôt un aspect de contrôle, mais également une volonté d'augmenter la connaissance en fusionnant les jeux de données et sont des objectifs légitimes.

Les données de la recherche ne peuvent pas être modifiées de manière continue, car si cette possibilité existait, aucune reproductibilité ne serait envisageable. Elles doivent être toujours figées, liées à l'instant T, ce qui n'empêche, ni la critique, ni leur mise à jour, ni la correction, ni leur revisite et la publication d'autres données. On peut par exemple effectuer un erratum en indiquant avoir fait une erreur dans l'échantillonnage ou dans la méthodologie. Ces pratiques sont autorisées par l'ensemble des journaux scientifiques.

Les données recueillies de manière continue, ne sont pas vraiment des données de la recherche, elles peuvent être utiles dans le cadre de recherche, mais les premières doivent être stables et stockées pour en permettre la reproductibilité.

On peut dissocier les données des hypothèses et considérer qu'une donnée de la recherche peut prendre à peu près toutes les formes, mais l'une d'entre elles est précise, la donnée associée à une expérience, qui dispose d'un statut particulier, elle a force d'évidence, de preuve. Elles peuvent également prendre une forme continue, comme des données climatologiques, qui sont collectées au fil des recherches et qui sont-elles, des données misent à jour en continu. Ces dernières pourront alors être utilisée dans le cadre d'autres expériences et pourront soit rester du même type, soit acquérir le statut d'évidence. Les grandes distinctions à prendre en compte dans les données de la recherche sont : publiques, non publiques, utilisée en qualité de preuves ou pas.

Facteurs de partage et de conservation

Les facteurs de partage des données de la recherche tiennent principalement aux exigences contractuelles des agences de financement et des journaux qui publient les données. Si les conditions indiquent que les données doivent être maintenues sur une durée de 24 mois, alors, vous allez faire en sorte que cela soit le cas.

L'usage est également un facteur important, si vous disposez de l'usage de données au-delà d'une période de 24 mois, vous essayerez de les maintenir au-delà du délai requis. Il se peut que certaines données représentent un intérêt qualifié ici de déontologique qui peut inciter le chercheur à la conservation et au partage.

On peut donc distinguer trois catégories de données : contractuelles, déontologiques et celles ne disposant d'aucune incitation au partage ou à la conservation. S'il y a une part de décision du chercheur dans la sélection des données à partager, mais elles ne sont pas explicitées. Elles sont souvent dictées par les cadres dans lesquels le chercheur évolue. Mais si vous n'avez pas d'indications, vous aurez tendance, si vous disposez d'un espace de stockage suffisant.

Les données sont conservées, car les échantillons peuvent être reproduits sur leur base, si l'on peut disposer de la totalité des processus permettant de refaire les expériences pour lesquelles elles sont conservées. Il est important de distinguer la donnée du processus.

Financements

La prise en compte des charges liées à la préparation des données en vue de leur publication (curation, documentation, etc.) est assez hétérogène et dépend fortement de l'agence de financement attribuant les fonds pour le projet de recherche et du type de ce dernier.

Il est assez rare de pouvoir déterminer le besoin financier exact nécessaire à la mise en forme des données ou à la préparation d'un rapport scientifique. Cela ne veut pas dire que la tâche est impossible à réaliser.

En règle générale, c'est implicitement lié aux exigences des organes de financements ou à celles des journaux dans lesquels vous êtes amenés à publier.

Par exemple, si vous obtenez des fonds européens, des livrables spécifiques vont vous être demandés : publication des résultats et des données dans des ouvrages scientifiques, mais ce, sans attribuer spécifiquement 40 heures ou un montant de 2000.- pour rédiger un article. En revanche il s'agira de la configuration minimale exigée.

Un autre exemple, si vous disposez d'une somme de 1000.- pour publier un article en Open Access, la facture présentée va être acceptée que vous déclinez de manière explicite ou non les informations concernant la gestion et la préparation des données de la recherche.

Donc on ne disposera pas forcément d'un paquet de travail intitulé : pérennisation des données de la recherche, mais on attendra de vous que vous en teniez compte dans les processus de recherche.

La recherche est également financée par les cantons, via les postes de professeurs, même si certaines places peuvent recevoir des fonds privés. Cela ne devrait entraîner aucune différence de fonctionnement, car en réalité, les directives viennent du FNS, quelle que soit l'origine des enveloppes budgétaires et les établissements exercent des contrôles rigoureux.

Long Tail of science & Data

En ce qui concerne les données de la recherche, il y a peu de différences entre les sciences qualifiées de dures3 ou de douces4. Il peut être arrangeant de les différencier lors d'une conversation, mais c'est en réalité très souvent une vision simplificatrice de la réalité scientifique. Séparer ces deux types de sciences n'est pas forcément souhaitable, en raison notamment de l'extrême hétérogénéité au sein même des disciplines scientifiques. En effet, on trouve certaines spécialités des sciences dites dures, dans lesquelles les données peuvent être dissemblables et peu standardisées, alors que dans certains champs de recherches en sciences humaines, on peut consulter des données normalisées.

Open Access

L'Open Access dispose de plusieurs significations : Il s'agit d'un modèle d'affaires pour de nombreux journaux, sous son aspect « Gold Road » qui vise à rendre libre l'accès aux publications scientifiques en finançant leur parution au moyen des APC.

Il s'agit également d'un modèle de diffusion, sous son aspect « Green Road » qui vise à ce que l'ensemble des publications scientifiques voient leurs preprints déposés dans des archives ouvertes qui permettent d'accéder librement à la forme non mise en forme par un éditeur de ces derniers.

Il s'agit également, en tout cas potentiellement de publier et de rendre accessible de manière ouverte.

Si toutes les données publiées en Open Access sont considérées comme publiques. Mais il est important de distinguer les deux types de données. En effet, si la donnée est publique ou a vocation à le devenir, on pourra alors la publier en Open Access. Alors que si la donnée n'est pas publique, pour des raisons légales ou éthiques (par exemple dans le cas de dossiers de patients), elle n'aura alors aucune raison d'être publiée sous cette forme.

Outils

Une fois les clés requises pour permettre de partager les données maitrisées, soit le stockage, la définition de la disponibilité, ainsi que la sélection du type. Il est nécessaire de disposer des instruments pour accéder aux données. Les moteurs de recherche pour les données textuelles, mais il est essentiel de développer quelque chose de plus performant pour les données structurées, des solutions de recherche plus intelligentes.

L'étape suivante est le développement d'outils permettant de donner du sens aux données, car si accéder à celles-ci et permettre la navigation au sein même des sets de données est un début prometteur, cela reste insuffisant. Il est nécessaire de pouvoir les utiliser réellement. Le web sémantique sera un bon début, mais ne répond pas à la totalité des problèmes posés par de tels développements, car le problème de l'ambiguïté des données n'est pas résolu par son utilisation.

On ne pourra pas non plus décrire la totalité des données par ce biais, car les descripteurs sémantiques sont limités et les données textuelles conserveront une certaine importance, ainsi que d'autres éléments qui ne sont ni textuels, comme des gloses qui permettent la description et l'explication, ni sémantiques continueront également.

Le DOI est un instrument qui permet d'attribuer un identifiant, mais ne garantit que partiellement l'accès aux données et qui ne donne aucune information sur celles-ci, d'autres éléments sont à prendre en compte pour le rendre possible, comme le maintien des sites d'hébergement par exemple.

Les ARK semblent être une alternative intéressante, car ils disposent de la même unicité, sont gratuits, ce qui n'est pas négligeable si vous devez générer beaucoup de DOI. De plus, ils sont maintenus par les bibliothèques, qui en termes d'infrastructures pérennes est une approche intéressante, car les probabilités qu'une bibliothèque institutionnelle se maintienne sont plus importante que celle d'un site web de chercheur.

Particularités helvétiques

Le découpage linguistique n'est pas spécifique à la Suisse et le découpage HES / Universités est administratif et, de fait, ne dispose pas d'une grande importance. Ce qui est spécifique au contexte helvétique est son mécanisme de financement de la recherche. En effet, il y a deux ou trois agences de financement et se sont-elles les actrices principales. Ce sont elles qui structurent la politique, qui dirigent et décident et

leur faible nombre implique que les décisions peuvent être prises rapidement, contrairement à des situations ou les environnements de financement de la recherche sont plus hétérogènes et fractionnés.

Le relatif désavantage du contexte helvétique est que la recherche suisse dispose de moyens importants et que l'on a tendance, de ce fait, à développer des solutions locales pour les problèmes rencontrés ou dans le cadre de projets. Pour illustrer, lorsque le FNS a décidé la mise en place d'archives institutionnelles en Suisse, plutôt que d'en avoir créé une par autorité, il a exigé que cette démarche soit effectuée par les institutions elles-mêmes, entrainant le développement de plus de dix solutions locales, ce qui est probablement autant de trop. Parce qu'ainsi, aucune ne dispose de la masse critique pour être suffisamment visible et sont toutes agrégées par Google Scholar et que ce type de développement coûte cher. On peut constater l'importance du FNS pour la recherche helvétique, dont chacune des décisions peut entraîner des modifications, mais son fonctionnement reste très dépendant des universités au travers de swissuniversities. Quant à la CUS, elle dispose d'un certain poids, mais si le FNS édicte des recommandations, celles-ci auront un impact beaucoup plus élevé.

Politique institutionnelle

La HES-SO12 fonctionne selon deux aspects :

Un groupe informel de travail institutionnel informel, sous la responsabilité du rectorat et dont la mission et de mettre en place une politique de gestion des données de la recherche et de collaborer aux projets nationaux dans le domaine.

Le second aspect, constitué de la gestion des données de la recherche est placé sous la direction des domaines HES-SO, qui par analogie, peuvent être comparés aux facultés des universités. Le domaine Economie et services, dont fait partie la HEG dispose probablement de ce qui se fait de plus avancé actuellement au niveau institutionnel dans la gestion des données. En effet, l'entité dispose d'une archive thématique : ArODES, qui permet de stocker des publications, peer reviewed ou preprint, des travaux d'étudiants et des données de recherche.

Annexe 2 : Entretien semi-directif, J-B. Claivaz

Entretien réalisé avec Monsieur Jean-Blaise Claivaz, coordinateur de l'archive ouverte et responsable de la gestion des données de la recherche pour la division de l'information scientifique pour l'université de Genève, le 19.05.2016. Les réponses de l'interlocuteur ont été retravaillées et classées selon les thèmes identifiés pour permettre de disposer d'éléments pertinent de comparaison entre les différents entretiens.

Acteurs

Il n'existe à ce jour, pas d'interlocuteur privilégié qui peut répondre aux problématiques soulevées par la gestion des données de la recherche dans le cadre de l'université de Genève. On est en effet dans une phase exploratoire. Trouver des interlocuteurs est complexe, plusieurs scientifiques considérant que la bibliothèque n'est pas concernée par le partage des données de la recherche, que cela concerne les chercheurs et non les spécialistes de l'information.

Mettre en avant des compétences potentielles que l'on peut acquérir en s'appuyant sur l'expertise des professionnels de l'information dans la gestion de métadonnées et sur la description, la curation et l'enrichissement des informations, ainsi que sur une certaine expérience de l'archivage à long terme, permettent de positionner les bibliothécaires comme candidat acceptable pour soutenir les scientifiques dans la gestion des données de la recherche. Mais cela reste une forme de pari, pourra-t-on, dans cinq ans, constater que ces pratiques seront acceptées et que des professionnels de l'information seront au centre de la recherche, dans la gestion des données ?

Pour les bibliothèques, il est important de se positionner face aux chercheurs et de « réinventer » une utilité pour eux, car si elles restent très pertinentes pour l'enseignement, elles sont de moins en moins considérées par les premiers, la presque totalité des informations nécessaires se trouvant en ligne. La présence de solutions telle que See Hub1 annule l'argument qui voulait que les parutions et production scientifiques soient disponibles par le biais des budgets de l'institution via les abonnements.

Les possibilités d'accès qui sont hors du domaine de l'institution amènent un changement de paradigme dans les pratiques informationnelles des chercheurs. Alors qu'il a quelque année, une panne dans l'accès aux bases de données provoquait une réaction très forte de leur part, des questions sur la situation, alors que récemment un éditeur a bloqué les droits de téléchargement de sa solution, ce sans provoquer une

seule réaction des scientifiques affilié à l'université. Cette évolution est également confirmée par la baisse des consultations sur les statistiques d'Elsevir.

Critères essentiels

Je ne sélectionnerai pas forcément le peer review, parce que ce type de solutions est scalable et donc non essentiel. Par analogie au lancement d'un véritable journal. Il faut que l'impulsion de base vienne des milieux académiques. En effet, d'après les expériences que j'ai pu observer et les différentes lectures que j'ai faites, le lancement d'une solution de type presse universitaire sans aucun chercheur dans le bateau n'amène rien, car la bibliothèque n'a alors rien à presser.

Donc il faut qu'un chercheur sollicite l'institution afin d'obtenir une infrastructure alors qu'il dispose de plusieurs auteurs qui sont prêts à écrire, de relecteurs et d'un prestige personnel qui entraînera un attrait pour la revue et une augmentation à terme du nombre d'articles proposés. A ce moment-là, l'entreprise devient rentable car la mise à disposition des infrastructures fait partie d'un partenariat entre les académiciens qui offrent le contenu et l'institution.

Données

L'une des questions irrésolues est quel type de données sera accepté. Car de nombreux chercheurs pensent que leurs données n'ont aucune valeur à long terme. Cela ne sert donc à rien de faire de l'archivage de données qui seront détruites dans dix ans. Il faudra donc prévoir des solutions de stockage temporaires, mais qui garantissent une reconnaissance et propose une réévaluation régulière. Indiquer au chercheur que le délai de destruction des données déposée est dans six mois et l'interroger sur la pertinence de leur conservation, ce en tenant compte que la volonté institutionnelle n'est pas de tout conserver, car c'est impossible.

L'archiviste décide ce qui doit être conservé, il est donc nécessaire de mettre en place des règles de conservation, des grilles de critères. Il faut indiquer que l'on conserve ce set de données pour telles raisons. Définir le fait de garder des données parce qu'elles permettent de vérifier les résultats d'un article est un principe théorique, cela ne s'est jamais fait jusqu'à présent car les données n'étaient pas conservées car non publiées.

Devons-nous les conserver pour garantir la reproductibilité de la science ? Pleins de gens estiment que ce sont des balivernes, que ce n'est de toute façon pas reproductible et que ça entraîne des coûts monstrueux. Par exemple, en génomique, certains

avancent que conserver des séquences de la mouche Tsé-Tsé n'a pas de sens, car si dix ans après le besoin d'une telle séquence se fait sentir, la reproduire alors se ferait dix fois plus rapidement, avec un résultat dix fois meilleur et dix fois moins cher en raison de l'évolution des technologies. Même si on peut arguer que la donnée produite peut être réutilisée, un certain nombre d'erreurs peuvent être présentes, car la méthodologie précédente n'était pas tout à fait cohérente.

Facteurs de partage et de conservation

Nous ne disposons pas d'une vision d'ensemble des pratiques liées au partage des données de la recherche pour les chercheurs affiliés à l'institution. Nous avons effectué des enquêtes avec des interviews auprès de chacune des facultés composant l'université. Trouver des collaborateurs pour réaliser ces entretiens n'a pas été simple et l'image dont nous disposons actuellement est loin d'être une vision d'ensemble.

Comme pour la perception de l'Open Access, je pense que le positionnement des chercheurs face au partage des données de la recherche relève des croyances individuelles et que cela n'a rien à voir avec leur discipline scientifique. Certains sont enclins à diffuser, d'autres pas et il ne semble pas que le clivage soit générationnel, car on retrouve des chercheurs en faveur de la diffusion dans toutes les catégories d'âge. Il s'agit plus de convictions personnelles et d'aptitude au partage qu'autre chose.

On peut notamment déterminer que cette différence existe également dans le cadre de l'attribution de récompenses pour la diffusion des données de la recherche afin d'en accroitre l'intérêt auprès des chercheurs. Car certains d'entre eux ne sont pas sensibles à ce type de levier et n'effectuent le partage que parce qu'ils considèrent que c'est bien !

Même si je ne fais pas partie du milieu académique a proprement parlé, mon expérience et les différentes collaborations que j'ai pu effectuer, me font penser que les chercheurs ne sont pas forcément soumis à la problématique du « publish or perish », même si l'environnement académique est très concurrentiel et que l'on peut le voir comme un « panier de crabes ».

Aux USA, le modèle exige, pour pouvoir être nommé, d'avoir publié un livre au moins, les projets de recherche sont très fréquemment temporaires et les scientifiques évalués chaque année, avec le risque d'être licencié. Alors qu'en Suisse, j'ai l'impression que de telles contraintes n'existent pas, que la gestion budgétaire est plus lissée sur la durée et

que la publication de 40 articles et la qualité des professionnels sont des critères suffisants pour obtenir des postes de travail.

Lors d'un entretien, un spécialiste de biologie nous a indiqué qu'il ne souhaitait pas échanger ses données, Car il considère qu'elles lui appartiennent et qu'il n'a pas confiance dans celles produites par des tiers. Ne sachant ni les conditions de température, ni qui a effectué le captage des données, ou encore si la chaine de conservation a été respectée et ne connaissant pas le degré de conscience professionnelle du manutentionnaire chargé de manipuler les matériaux liés à l'expérimentation.

Si certains groupes de recherche, ou certaines connaissances peuvent bénéficier de sa confiance et donc voir leurs données utilisées, ce n'est pas le cas de la plupart. Qui peuvent falsifier leurs résultats, en raison de l'environnement concurrentiel dans lequel il travaille. De plus cette méfiance s'étend aux autres chercheurs du domaine.

Une autre illustration nous a été fournie par une chercheuse. Elle faisait exécuter des tâches à un sujet devant un ordinateur et filmait le mouvement des yeux, ses activités et ses temps de réaction. Un des éléments identifiés par celle-ci est l'importance de la température des locaux dans lesquels les tests sont réalisés. Mais cette variable n'est jamais indiquée dans le protocole, ce qui compromet la reproductibilité de l'expérience, donc que reste-t-il de véritablement reproductible ?

Le CERN stocke énormément de données et chaque cinq ans, effectue une migration pour maintenir les performances des technologies utilisées (bandes, etc.) et garantir leur pérennité. Un contrôle au niveau du bit est effectué sur les données afin de contrôler la présence d'éventuelles corruptions. Cette opération dure trois ans. Grâce aux énormes quantités de données dont dispose l'organisme, il a été constaté que les logiciels réalisant les tests de vérification souffrent eux-mêmes de bugs, en raison du nombre de décimales important, ces derniers dysfonctionnent.

Long Tail of research data

Une certitude est que la solution développée par le projet DLCM ne s'intéressera pas à toutes les formes de données. Par exemple, nous n'intéresserons pas aux Big Data, ni à la physique, qui dispose déjà de modèle et d'outils de diffusion, comme les spécialistes des protéines avec Swiss Prot, où les astronomes et les climatologues. Nous allons nous occuper du « Long Tail of Science », toutes les petites disciplines qui ne disposent pas de solution.

Outils

Relier les différentes solutions existantes pour la diffusion des données de la recherche est quelque chose auquel nous devons réfléchir assurément, même si aucune démarche n'a été entamée dans cette direction, cela devrait être l'un des résultats du projet DLCM7. Mais fondamentalement, le besoin de faire le lien entre des données archéologiques et de médecine est-il réel ? A priori non, sauf dans le cas où l'appartenance institutionnelle des données est importante.

Aucune recommandation pour le dépôt des données de la recherche n'a été émise. Mais nous allons y venir, nous sommes en phase de construction de l'outil. Nous disposons d'un préfixe pour l'attribution des DOI, mais la forme n'étant pas encore définie, nous ignorons si les données se verront attribué un identifiant et seront déposée sur l'archive ouverte, ou si un fonctionnement sera sélectionné.

Deux logiques s'affrontent : l'une, institutionnelle, veut mettre en place une archive dans laquelle doivent être déposées l'ensemble des données produites par les recherches effectuées par les membres de l'institution et l'autre, disciplinaire, qui veut effectuer les mêmes opérations mais dont le découpage est orienté autour des spécialités scientifiques, illustrée par les physiciens qui souhaitent déposer leurs publications dans ArXiv.org et non sur l'archive ouverte de l'institution. La seconde présente plus de sens et est beaucoup plus parlant pour les chercheurs, alors que pour des raisons de simplification de gestion des solutions techniques, des financements, etc., les institutions sont favorables à la première, même si cette dernière implique une uniformisation des formulaires.

Ces deux logiques sont en confrontation et j'ignore si l'on aboutira à la mise en place d'une archive institutionnelle pour les données de la recherche. Va-t-on développer des

silos disciplinaires ou rendre accessible l'ensemble des données dans un seul « chapeau », en arguant, venez chez nous, LE dépôt, ça je l'ignore.

L'un des projets est de mettre en place un cloud repository au niveau suisse, parce qu'assurément, nous ne voulons pas répliquer une grosse infrastructure en plusieurs exemplaires dans l'environnement. Mais expliquer comment nous allons le vendre où la forme qu'il pourra prendre est aujourd'hui impossible. Fera-t-on une sorte de template avec une visibilité sur une seule discipline afin que les futurs utilisateurs ne voient pas qu'ils sont dans un système globalisé ? Ou une décision politique va-t-elle nous dire quelle forme elle devra prendre et celle-ci va-t-elle attirer l'intérêt des chercheurs Je ne sais pas, nous sommes plutôt dans le domaine du pari dans ce cas ?

On parle de deux problématiques distinctes. D'un côté, une infrastructure de stockage et de préservation des données, de l'autre des interfaces de recherche et d'agrégation, etc., ce sont des éléments complétement différents.

Peut-être que le seul succès du projet DLCM sera de parvenir à mutualiser les infrastructures de stockage, au niveau de la capture de données. Si l'on parvient à effectuer cela, ce sera déjà une réussite. Sans parler de diffusion, de réutilisation, de standardisation, de journaux, etc., qui sont pratiquement des projets en soi et qui visent à identifier des solutions pour valoriser cette première partie du DLCM, que la structure soit centralisée de manière nationale ou pas.

Dans l'absolu, DLCM devrait faire tout, mais j'ai un doute sur la réussite, sur une durée de quatre ans de l'ensemble des problématiques à traiter. De plus faire de la valorisation sans disposer de stockage n'est pas utile. Donc l'accent doit être mis en priorité sur la capture des données, en misant par exemple sur les données actives, traitées dans le track 2, comme une récupération automatique des données d'instruments de laboratoires et un stockage sur l'archive ouverte, sans travail supplémentaire des chercheurs.

L'idée est d'abord de sécuriser la couche accès et puis ça, c'est quelque chose qui a un réel intérêt de synergie nationale, parce qu'on a tous le même problème de stockage, de masse, de format, de compétences techniques et qu'il est intéressant de mutualiser ces démarches. Ensuite, l'exploitation peut amener des divergences de visions et qu'on doive le faire de manière disciplinaire et de manière transnationale et non de développer un projet suisse d'exploitation des données de n'importe quelle discipline a plus de sens.

Politique institutionnelle

La responsable de la DIS à une vision et estime que les bibliothèques ont un rôle stratégique dans la gestion des données de la recherche, qui permettra à terme, de justifier leur existence. Elle voit donc cela comme une reconversion d'une partie des missions traditionnelle de la bibliothèque, comme, par exemple, le catalogage, qui est vouée à disparaître. En qualité d'employé, je partage son point de vue. Actuellement les réflexions en cours sont : comment élaborer une politique, mettre en place des services pour ensuite vérifier qu'ils correspondent aux besoins identifiés.

Annexe 3 : Entretien semi-directif, E. Blumer

Entretien réalisé avec Madame Eliane Blumer coordinatrice du projet DLCM et responsable du développement de services en lien avec les DR pour l'UNIGE., le 22.05.2016. Les réponses de l'interlocutrice ont été retravaillées et classées selon les thèmes identifiés pour permettre de disposer d'éléments pertinent de comparaison entre les différents entretiens.

Acteurs

DLCM, Cet acronyme signifie Data Life Cycle Management et ma vision de ce projet et que nous essayons de proposer des solutions adaptées aux besoins des chercheurs tout au long du cycle de recherche pour ce qui concerne les données. Dès le dépôt de la proposition de recherche, en parallèle des tentatives pour l'obtention de financements, nous les accompagnons dans la rédaction d'un Data Management Plan, conformément aux exigences des bailleurs de fonds. On reste à leur disposition pour définir les modes de dépôts, de publication et de partage.

Le projet est planifié pour une durée de trois ans, financé par CUS-P2, nous avons commencé en septembre 2015 et une prolongation jusqu'au terme de 2018 nous a été accordées. Et son objectif est le développement de services avec les institutions partenaires, dans le cadre national, au vu de l'origine des fonds soutenant le DLCM.

Nous avons effectué des interviews avant de lancer le projet et les réactions étaient très mitigées. Certains étaient enthousiastes et très ouverts, notamment les scientifiques qui ont recours à des formats très hétérogènes, qui traitent de questions juridiques, des données géo spatiales, de stockage de masse ou de types de formats. Car les nombreuses questions auxquelles ils sont confrontés rendent pratique la présence de quelqu'un pour les soutenir dans la gestion de leurs données.

D'autres disciplines ont été peu abordées, comme les sciences de l'informatique, car les demandes qui nous ont été formulées étaient justes de l'ordre de la fourniture d'espace de stockage, les reproductions d'algorithmes sont directement gérées par les chercheurs. Il a été complexe d'entrer dans le domaine de la médecine mais les chercheurs spécialistes se sont montrés intéressés et collaboratifs, ce malgré la question de l'anonymisation des données et que nous pensions être un obstacle. On ne peut toutefois pas parler de généralisation aux seins même des domaines touchés par nos démarches.

Nous n'avons pas pu rencontrer Springer dans le cadre d'un entretien officiel, mais nous avons pu les croiser lors de workshops sur les données de la recherche. Et l'éditeur dispose malheureusement d'une position avantageuse, qui lui permet de rendre accessible tout ce que nous cherchons à réaliser avec la publication préalable de nos données.

Il a mis en place une solution pour les laboratoires, en cours de développement dans le cadre d'un cas pratique réalisé en collaboration avec un laboratoire de médecine, qui permet de relier les données et les publications au dépôt qui permet de disposer des fonctions de publication et d'analyse des données. Ce qui accélère l'automatisation de la publication des données.

C'est dommage que les vues Springer soient commerciales, car si l'éditeur dispose d'un système qui analyse les données et qui se montre suffisamment modulable pour être adapté aux autres disciplines scientifiques et propose un modèle qui autorise la publication d'un maximum de données de la recherche, combiné à une automatisation des processus une fois le cap des cinq citations franchies et à la qualité du H-index de ses publications, je pense que je l'utiliserai également.

Critères essentiels

Une description normalisée, mais modulable et une solution qui permet l'interopérabilité avec un maximum de systèmes. Le recours massif à la publication en Open Access. L'ajout de fonctions ludiques qui comprennent les métriques.

Une ouverture à l'interdisciplinarité, car sans cela la mise en place d'une solution de partage des données de la recherche n'aurait pas sens. Mais je maintiendrais une structure demandant un minimum de description, sans être limitant pour les utilisateurs qui souhaitent apporter plus de détails à leurs sets de données. La mise en place d'un pool de données libres d'accès, qui peuvent être déposées en attente de traitement et la mise à disposition d'un service de traitement pour les enrichir avec de la documentation, voir l'organisation d'activités tels que des Hack Haton sur les données. Pourquoi ne pas prendre des solutions telles que Github ou Jupiter notebook et leur ajouter des fonctionnalités de publication.

La simplicité des outils développés entraîne leur fonctionnalité, les éléments tels que la complexité des développements, qu'ils soient politiques, technologiques ou qu'ils

concernent la gestion des accès ne doivent pas transparaître. La solution doit être utilisable par des personnes qui n'ont rien de « scientifique ».

Data paper

Un Data paper est pour moi lié avec une publication, ou alors, ne dispose pas d'une valeur importante et j'ai de la peine à me dire que la mise en place d'un data journal suisse va être important. Nous savons que c'est un environnement très compétitif, mais souvent, les chercheurs suisses ne sont pas PI, ils sont englobés dans un groupe important qui effectue la recherche. Donc pourquoi ne pas développer un système qui permettrait d'effectuer de bonnes recherches n'importe où ? Peut-être est-on limités par des problématiques juridiques qui exigent de tenir compte du territoire helvétique ? Mais dans ce cas pourquoi ne pas faire une entrée suisse qui permet d'accéder aux contenus internationaux ?

Parce que ce qui m'intéresse c'est : quels chercheurs affiliés à l'UNIGE ont publiés en Open Access, sont engagés dans des projets de recherches n'importe où, dans le cadre d'Horizon 2020, etc. Pour cela je n'ai pas besoin d'une solution suisse. Ne serait-il pas plus intelligent de développer des instances locales de plusieurs data journals, de mettre en place des nœuds suisse de publications internationales afin de contourner certaines problématiques de droit ?

Facteurs de partage et de conservation

Changer les pratiques de publication est une problématique extrêmement complexe et j'ignore si indiquer à un chercheur qu'il doit publier ses données primaires sans qu'elles soient validées peut fonctionner. L'EPFL est en train de mettre en place un système de crowdsourcing des données et cela pourrait être une solution viable qui permettrait à chaque participant de bénéficier d'avantages comparables, contrairement au fonctionnement d'un modèle traditionnel de peer review commercial.

Je n'ai pas encore pu identifier beaucoup d'articles qui indiquent qu'une nouvelle découverte a été possible grâce à la publication de données de la recherche. Je me demande donc dans quels domaines est-il possible de réellement vérifier qu'une telle publication, validée au préalable, a eu un impact majeur. Pour les institutions, la motivation première du partage des données de la recherche est l'économie, qu'elle concerne l'échelle, ou la non recollection des données, en vue de diminuer les coûts. Si de tels moteurs sont au centre des démarches institutionnelles, la collaboration avec des

solutions existantes est plus pertinente que la mise en place de nouveau éléments. Dans ce cadre, il serait intéressant de connaître de manière précise les besoins des communautés scientifiques et de suivre, au moyen de métriques, les citations de telles productions académiques, afin de déterminer l'utilité réelle des données de la recherche.

Le gestionnaire de donnée est pour le dire « méchamment » un genre de secrétaire évolué, qui organise les données. C'est une tâche intéressante, car elles peuvent disposer de plusieurs formats, mais cela reste de l'organisation et non de la réflexion. Ce qui implique que le chercheur n'est pas forcément motivé à effectuer la gestion de ses données.

Financements

Chaque fois qu'un projet est financé par un fond, tel que le FP7 ou Horizon 2020, ce type de tâches est pris en compte dans les calculs d'attribution. Le chercheur doit donc en tenir compte, car il reçoit des financements pour. Nous disposons de suffisamment de fonds pour étendre ces pratiques aux recherches suisses.

Vu que le projet DLCM ne travaille pas sur du long terme, une réflexion sur les coûts est nécessaire. Le projet dispose donc d'un business analyst qui étudie les flux de financement et cherche à déterminer avec nous de quelle manière le projet pourrait atteindre une forma de rentabilité. Nous avons donc développé un modèle de coût pour le stockage des données afin de se démarquer de solutions comme Dryad, etc. Afin de disposer d'une offre plus avantageuse.

Open Access

Le DLCM, ne considère par l'Open Access comme une partie importante des réflexions à mener, le mouvement étant présent depuis longtemps dans le paysage de la publication scientifique et que nous ne souhaitons pas, pour le moment lutter pour la libre disposition des données. L'accès à celles-ci peut être fait via les métadonnées, comme dans Zenodo et qu'elles soient soumises à un embargo et qu'elles se libèrent ensuite est suffisant. Mais selon mon opinion, les deux éléments vont se rejoindre, car je lis beaucoup de documents Open Access depuis que je travaille dans le domaine des données de la recherche.

Outils

Un gap analysis a été entamé, pour chaque institution par rapport à leur dépôt institutionnel et en vue d'un archivage pérenne, les développements sont donc plutôt orientés autour de l'archivage que de la diffusion. Des DOI ont été commandés, afin de pouvoir répondre aux besoins des chercheurs et des services ont été mis en en place dans certaines institutions afin de pourvoir des conseils sur des outils tels que Zenodo, mais nous ne pouvons conseiller aucune solution suisse pour le moment.

Dans des domaines scientifiques très pragmatiques, comme la médecine, nous avons constaté qu'il existe des listes des endroits spécifiques où les productions peuvent être publiées et quelles sont les possibilités des chercheurs, etc. Mais nous espérons pouvoir rendre possible qu'un chercheur en sciences humaines qui cherche à comprendre comment fonctionnait le cerveau de Martin Bodmer puisse trouver d'autres données déjà présentes sur un autre dépôt pour en permettre la comparaison. De mettre à disposition une forme de méta moteur qui permettrait d'interroger l'ensemble des espaces de stockage et leurs documentations.

Je vois mal la mise en œuvre d'un nouveau système. Ajouter une couche de recherche sur chaque système existant, en quelque sorte, la mise en place d'un « Swisslib » pour les données me semble plus cohérant. Rendre son utilisation obligatoire par les institutions qui ne disposent pas d'une archive ouverte ne me semble pas envisageable, mais leur proposer d'intégrer un système qui offre de nombreux avantages, comme un méta moteur de recherche performant, devrait leur faire comprendre de l'importance de leur participation. La mise en place d'une communauté qui informe que les organismes sont impliqués dans les processus techniques ou non, me semble être important.

Nous essayons également de développer une collaboration avec l'Open Data Research depuis plusieurs semaines car cela nous semble très intéressant en raison de la possibilité d'usage des solutions développées par ce projet. Ses outils sont déjà fonctionnels ce qui permet d'utiliser les « méta mécanismes » existants, basés sur leur site. Nous pourrions utiliser ces derniers dans un futur service de diffusion ou reprendre le projet. Nous disposons de plusieurs idées, mais les efforts sont actuellement concentrés sur ce qui est déjà existant.

Particularités helvétiques

Cela fait maintenant deux ans que j'exerce dans le domaine des données de la recherche et je pense que l'on ne devrait pas parler d'un concept de nation, même si le fonctionnement de l'attribution des financements nous oblige à en tenir compte. Les besoins sont selon moi institutionnels, disciplinaires et communautaires. Le besoin actuel est plus une nécessité de faire correspondre le système à d'autres, comme la recherche en UK.

La Suisse est une sorte d'îlot, qui collabore à plusieurs projets de grande envergure, sans totalement en faire partie intégrante et nous devons donc adapter notre appareil pour pouvoir travailler de manière interdisciplinaire et internationale. Mais j'ignore si constituer un dépôt suisse est vraiment la bonne manière de suivre ces tendances. Il serait plus intéressant de suivre l'ensemble des projets européens et d'intégrer cette communauté.

Politique institutionnelle

Il n'existe à l'heure actuelle aucune politique institutionnelle qui définit ce qui doit être réalisé dans la gestion des données de la recherche. Nous disposons d'un modèle en phase terminale de développement et nous essayons de le valoriser sur la scène nationale. Nous savons que le FNS est également en cours de travail sur une proposition et il serait excellent de parvenir à mettre en place une collaboration.

Le modèle que nous développons se présente sous la forme d'une structure modulaire qui permet de sélectionner les phrases composant la politique institutionnelle suivant le degré de profondeur souhaité par l'organisme. Nous avons tenté de créer, sur cette base une politique de gestion des données de la recherche pour l'UNIGE et l'Université de Bâle suit une démarche identique. L'EPFL dispose d'une politique écrite, qui n'est pas basée sur la structure que nous avons imaginée. Nous savons que ces démarches présentent de nombreuses difficultés et qu'il n'est pas certain qu'une solution soit implémentée d'ici la fin de l'année 2016.