h e g

Haute école de gestion
Genève

Développement d'une offre de formation sur la gestion des données de la recherche en médecine et santé publique

IUMSP
Institut universitaire de médecine sociale et préventive, Lausanne

CHUV

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

par :

Vincent HUBER

Conseiller au travail de Bachelor :

René SCHNEIDER, Prof. Dr. HEG-GE

Lausanne, le 15 juillet 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 Bachelor Of Science HES-SO en information documentaire

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. »

Fait à Lausanne le 15 juillet 2016

Vincent Huber

Remerciements

J’adresse mes sincères remerciements à toutes les personnes m’ayant de près ou de loin aidé et soutenu durant la réalisation de ce travail, et tout particulièrement :

Mes mandants, Cécile Lebrand, responsable du soutien aux chercheurs, de la gestion des données et de l’Open Access à la BIUM, et Pablo Iriarte, bibliothécaire/documentaliste & webmaster à l’uDDSP, pour leur confiance en premier lieu mais aussi leur disponibilité pour les nombreux rendez-vous, leurs éclaircissements, leurs corrections ainsi que leurs encouragements tout au long de ce semestre.

René Schneider, mon conseiller pédagogique, pour sa disponibilité et ses conseils avisés qui m’ont guidé dans les méandres de ce travail.

Mon juré Jan Krause, Data Librarian à l'EPFL, qui a eu l’amabilité d’accepter cette responsabilité.

Florence Biegajlo pour sa relecture et ses encouragements.

Enfin, je tiens à remercier chaleureusement mes proches pour leur indéfectible soutien durant ces 4 années d’études à la Haute École de Gestion.

Résumé

Le présent mandat a été réalisé pour la Bibliothèque Universitaire de Médecine (BIUM) du CHUV ainsi que pour l'unité de Documentation et Données en Santé Publique (uDDSP) de l'Institut en Médecine Sociale et Préventive (IUMSP) de Lausanne. La problématique de la gestion des données de la recherche est actuellement l'un des axes de réflexion majeurs des universités, notamment de l'Université de Lausanne (UNIL). Les données de la recherche en médecine et santé publique, en ce sens, n'échappent pas à la règle et leur gestion est d'autant plus importante que leur partage et leur description sont des conditions sine qua non à la reproduction des expériences ainsi qu'à leur compréhension. Soucieuses de cette thématique, la BIUM et l'uDDSP ont décidé de proposer aux chercheurs de l'IUMSP ainsi que de la Faculté de Biologie et de Médecine (FBM) de l'UNIL deux modules de formation enseignant cette problématique aux scientifiques.

Ce travail s'articulera autour de trois axes. Le premier consistera à réaliser une réflexion sur l'importance de la gestion des données de recherche notamment dans le contexte de la recherche biomédicale effectuée à la Faculté de Biologie et de Médecine ainsi qu'en santé publique effectuée à l'IUMSP. Le deuxième axe consistera en une revue et une analyse de formations existantes à la gestion des données de recherche en Suisse mais aussi à l'étranger. Ces formations pourront être généralistes ou s'intéresser précisément aux données biomédicales. Le but étant de faire ressortir des exemples pertinents dont nous pourrons nous inspirer par la suite.

Le troisième axe consistera en la proposition de deux modules de formation complémentaires pour les chercheurs. Le premier se réalisera au début du processus de recherche et représentera une introduction à la thématique de la gestion des données de recherche. Cette introduction s'articulera autour du Data Management Plan, document indispensable dans la gestion des données de recherche. Le second module sera réalisé à la fin du cycle de recherche. Il s'intéressera plus précisément à la problématique du partage des données de la recherche, la mise en ligne sur des dépôts spécialisés ainsi que la création des métadonnées.

Ces modules ne seront pas obligatoires pour les chercheurs et aucun prérequis n'est demandé. La durée de chaque module sera de deux heures, rendant difficile la création d'un compromis entre exhaustivité et rapidité. Au terme de ce travail, nous espérons donc pouvoir réaliser une formation complète, mariant pratique et théorie qui sera adaptée aux besoins des chercheurs et des institutions concernées.

Table des matières

Liste des tableaux

Liste des figures

Liste des acronymes

BIUM	Bibliothèque Universitaire de Médecine du CHUV
CC	Licences Creative Commons
CHUV	Centre Hospitalier Universitaire Vaudois
DDI	Data Documentation Initiative
DCC	Digital Curation Center
DMP	Data Management Plan. Plan de gestion des données en français
EPFL	École polytechnique fédérale de Lausanne
ERSP	École romande de santé publique
ETH Zurich	Eidgenössische Technische Hochschule Zürich, École polytechnique fédérale de Zurich en français
FBM	Faculté de Biologie et de Médecine de l'UNIL
FNS	Fonds National Suisse
INIST	Institut de l'information scientifique et technique
IRM	Imagerie à résonnance magnétique
IUMSP	Institut universitaire de médecine sociale et préventive
NECDMC	New England Collaborative Data Management Curriculum
PET Scan	Positron Emission Tomography
RDM	Research Data Management. Gestion des données de la recherche en français
SSC	Sections des sciences cliniques de la FBM
SSF	Section des sciences fondamentales de la FBM
uDDSP	Unité de Documentation et Données en Santé Publique de l'IUMSP
UEL	University of East London
UNIL	Université de Lausanne
UNIRIS	Service des ressources informationnelles et archives de l'UNIL

1. Introduction

1.1 Problématique générale

Le processus scientifique standard est basé sur une hypothèse de départ à partir de laquelle le chercheur réalise une expérience, analyse les résultats obtenus, communique ses résultats auprès de la communauté scientifique qui elle-même tentera de reproduire l'expérience afin de vérifier la fiabilité des résultats originaux. Autrement dit, la qualité et la crédibilité de la science sont avant tout basées sur la reproduction des expériences jusqu'à atteindre un niveau de confiance certain, permettant alors d'affirmer la réponse à une hypothèse. Or, la reproduction et le contrôle d'une expérience sont intimement liés aux données issues de cette recherche et à leur partage.

À la lumière de cela, il est à noter que le XXIème siècle représente actuellement un paradoxe certain. En effet, grâce aux réseaux de communication liés à Internet, il n'a jamais été aussi aisé de transmettre une information à un tiers. Mais dans le même temps, la quantité de données de la recherche a tout simplement explosé durant la dernière décennie avec des productions en masse grâce à des outils informatiques de plus en plus développés (nous parlons alors *Big Data),* rendant le tout difficilement exploitable, du moins à sa juste valeur. Cette production massive peut être vue comme une chance pour la science. En effet, plus de données représentent plus d'informations donc, plus de connaissances. Mais la question qu'il est alors nécessaire de se poser est la suivante : **Comment trier et valoriser ce gigantesque contenu ?**

Dans ce contexte, la gestion des données de la recherche – Research Data Management (RDM)[1] en anglais – représente une nécessité pour les acteurs du monde scientifique de par les avantages qu'elle procure. Elle permet dans un premier temps de trier efficacement les données afin de ne conserver et de ne partager que les données pertinentes et essentielles. De même, elle encourage et guide la création de métadonnées standardisées afin de rendre les données plus compréhensibles et plus fiables. Le partage des données de la recherche permet aussi l'élément central qu'est la reproduction des expériences ainsi que leur validation grâce à un contrôle plus précis des résultats.

Les enjeux financiers y sont aussi importants car l'augmentation de la quantité de données engendre également une augmentation des coûts. Or, le partage et la bonne conservation des données permettent d'éviter la duplication des travaux de recherche,

[1] Nous utiliserons indifféremment le terme français, anglais ou l'acronyme durant ce travail

en d'autres termes, d'éviter une répétition d'une expérience due à la perte ou au manque d'informations. Cet avantage permet de faire de substantielles économies aux institutions scientifiques sans pour autant remettre en question la qualité des recherches effectuées. La gestion des données permet aussi de se conformer au cadre légal et éthique dans lesquels les scientifiques opèrent mais aussi à la réglementation et aux exigences des institutions et des organismes financeurs.

À ce stade de la réflexion, la nécessité de la gestion des données de la recherche n'est donc plus une simple hypothèse. Pourtant, sa mise en pratique n'est que récente, bien qu'elle soit présente depuis plusieurs années dans les pays anglo-saxons. Ce délai est certainement dû à un partage des responsabilités entre les acteurs de la recherche, menant désormais à une pluralité des actions en faveur de la gestion.

Les scientifiques eux-mêmes ont réalisé l'importance mais aussi les bénéfices qu'ils pouvaient tirer de cette problématique. En effet, le partage de leurs données permet notamment une augmentation de leur taux de citation dans les articles spécialisés mais améliore aussi leur méthodologie de recherche (due à une collaboration globale) et donc de leur crédibilité. Des initiatives allant dans ce sens ont été avancées au sein même des associations de scientifiques afin de promouvoir cette collaboration.

Les organes financeurs, les États et les institutions universitaires ont bien évidemment saisi eux aussi l'importance, notamment financière, de cette problématique. Ainsi, des directives et encouragements ont été mis en place au niveau international (avec Horizon 2020 notamment[2]) et national (programme CUS 2013-2016[3] de Swiss Universities[4]). Ces directives listent les actions que les chercheurs devront entreprendre dans la gestion de leurs données s'ils désirent être bénéficiaires d'un financement. Parmi ces actions, la création d'un Data Management Plan (DMP)[5] – plan de gestion des données en français – est l'un des points centraux.

Or, ces directives pouvant être parfois pointues, les chercheurs ne sont pas forcément à même d'y répondre avec efficacité. Ici entrent alors en jeu la responsabilité des professionnels en information documentaire qui sont, parmi d'autres, les garants de la qualité, du partage et de l'accès à l'information. De nombreuses initiatives ont déjà été réalisées dans cette optique-là. Soutiens auprès des chercheurs, créations de dépôts

[2] http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html
[3] https://www.swissuniversities.ch/fr/organisation/projets-et-programmes/programme-cus-2013-2016-p-2-information-scientifique-acces-traitement-et-sauvegarde/
[4] À noter que le programme a été reconduit pour la période 2017-2020 à hauteur de 30 mio CHF
[5] Nous utiliserons indifféremment le terme complet ou l'acronyme durant ce travail

institutionnels, aides à la publication, créations de standards dans l'optique d'uniformiser les outputs, les spécialistes ID n'ont de loin pas été inactifs sur ce sujet. De même, de nombreuses formations et modules de cours ont été créés dans le but de former et d'encourager les chercheurs à la gestion et au partage de leurs données. Ces formations ont globalement été acceptées avec enthousiasme par les scientifiques car elles répondaient à un besoin certain.

De même, les professionnels ID, conscients de l'importance de ces formations, ont rapidement décidé de partager leurs expériences mais aussi le contenu-même de leur travail, créant ainsi des formations « clé en main » en vue d'être réutilisées par leurs pairs.

C'est dans ce contexte que se réalise ce Travail de Bachelor. La recherche biomédicale est elle-même très sensible à la problématique de la reproductibilité de ses expériences. Celle-ci est actuellement mise à mal par le manque d'accès et de transparence des données de la recherche mais aussi par des méthodes de publication favorisant l'information rapide au lieu d'une information complète, ne rendant que 15% des recherches réellement utiles pour la communauté scientifique :

> *« These estimates fit remarkably well with estimates of 85% for the proportion of biomedical research that is wasted at-large » (Begley & Ioannidis 2015, p. 116)*

De même, la quantité de données biomédicales a crû exponentiellement durant cette dernière décennie, phénomène dû notamment au séquençage du génome humain produisant une quantité très importante d'informations.

L'unité de Documentation et Données en Santé Publique (uDDSP)[6] et la Bibliothèque de Médecine du CHUV (BIUM)[7], conscientes de ses enjeux, ont décidé de proposer deux modules de formations auprès des chercheurs dans les domaines du biomédical et de la santé publique en plus des divers services déjà existants dans ce domaine. La problématique sera donc de créer ces modules dans l'optique de répondre au mieux aux besoins et aux attentes des chercheurs.

1.2 Cadre général

Ce mandat est réalisé pour la BIUM ainsi que pour l'uDDSP, placé sur la supervision du Dr. Cécile Lebrand pour la première institution et de Pablo Iriarte pour la seconde. Il s'articule autour de trois axes principaux.

[6] Nous utiliserons indifféremment le terme complet ou l'acronyme durant ce travail
[7] Nous utiliserons indifféremment le terme complet ou l'acronyme durant ce travail

Le premier consiste en l'analyse et la revue de la problématique de la gestion des données de recherche à un niveau général mais aussi au niveau des sciences biomédicales ainsi que de la santé publique. Il a été alors nécessaire d'y décrire les problématiques générales qui en découlent mais aussi les problématiques spécialisées à ces domaines scientifiques. En outre, il a été nécessaire de détailler le contexte de la BIUM et de l'uDDSP, les services déjà existants ainsi que les données gérées par ces institutions. En effet, les chercheurs qui y sont affiliés produisent des données spécifiques qu'il sera important de comprendre afin de définir leurs besoins.

Le second axe de ce mandat s'articule autour d'une revue et d'une analyse des formations existantes dans la gestion des données de recherche. Ces formations peuvent être aussi bien généralistes que spécialisées dans le domaine médical. Le but étant de comprendre leur fonctionnement afin d'en tirer des exemples pertinents dont nous nous sommes inspiré par la suite dans notre propre formation. Ces formations existent sous diverses formes : vidéos, jeux, questionnaires ou simples présentations Powerpoint, la diversité est aussi due au fait que certaines sont cantonnées à Internet alors que d'autres peuvent être réalisées en présentiel.

Le troisième axe de ce travail consiste en la création des deux modules de formation proprement dits. Le but a été ici de proposer des formations « clé en main » avec minutage et exercices pratiques que les intervenants pourront livrer rapidement après le rendu de ce travail. Il s'agissait donc d'être le plus réaliste possible afin de proposer un contenu qui serait réellement applicable. Plusieurs institutions ont créé des formations « toutes faites » destinées aux professionnels ID. Il n'a pas été question de reprendre telle quelle l'une de ses formations, néanmoins, certains exercices ou concepts de présentation ont été les objets de notre inspiration.

Le premier module de formation consiste en une forme d'introduction à la problématique de la gestion des données de recherche. Celui-ci sera donc proposé aux chercheurs lorsqu'ils débutent un cycle de recherche afin qu'ils puissent démarrer leur projet dans les meilleures conditions possibles. Au-delà de cette introduction, cette partie de la formation comprend un large chapitre sur le Data Management Plan. En effet, celui-ci procure deux avantages : le premier est que les chercheurs affiliés à l'IUMSP et à la Faculté de Biologie et de Médecine reçoivent la directive de réaliser un DMP. Il nous semble donc nécessaire de leur expliquer les tenants et aboutissants de cet outil et une aide à sa complétion ne nous paraît pas superflue. Le second avantage est que le DMP recouvre l'entièreté du cycle de vie des données de recherche, de la création à la réutilisation en passant par la création des métadonnées et le dépôt en

ligne. Il agit donc comme un alibi pour expliquer les problématiques liées aux différents stades du cycle de vie.

Le second module de formation sera proposé à la fin du cycle de recherche et s'articule autour de la problématique de la mise en accès libre des données de la recherche (*Open Data)*, son importance et les bénéfices qu'il peut en résulter. Il aborde aussi diverses problématiques présentes à ce stade du cycle de vie des données comme de la gestion des métadonnées et leurs standards afin de rendre les données le plus compréhensible possible, la mise en ligne sur des dépôts spécialisés ainsi que la gestion des formats de fichiers pérennes.

Il a été défini que les formations seraient données pour un public de 8 personnes à la fois environ, mais que ce chiffre pouvait varier selon la demande. De plus, chaque module de formation doit être d'une durée d'environ deux heures et la fréquence de la tenue de chaque module serait idéalement d'une fois par mois.

1.2.1 Objectifs généraux et spécifiques

- Rédiger une synthèse relative aux enjeux de la gestion des données de la recherche dans le milieu académique et biomédical plus précisément.
    - Réaliser une revue de la littérature sur la problématique de la gestion des données de la recherche.
    - Définir les spécificités de la recherche biomédicale en termes d'enjeux mais aussi dans les types de données créés.
- Analyser le contexte dans lequel évoluent le CHUV et l'IUMSP ainsi que leur offre actuelle en termes de services et aides auprès des chercheurs.
    - Réaliser un état des lieux de la BIUM et de l'uDDSP. Définir les données générées par les chercheurs affiliés aux deux institutions ainsi que leurs besoins.
    - Analyser l'offre actuelle de la BIUM et de l'uDDSP en termes de formations et d'aides aux chercheurs.
- Réaliser une revue et analyser les formations à la gestion des données de recherche généralistes mais aussi spécialisées dans les sciences biomédicales (études précliniques et cliniques) et en santé publique déjà existantes.
    - Réaliser un état des lieux des formations existantes sur la gestion des données de la recherche et spécialement dans le domaine biomédical en Suisse et à l'international.
    - Sélectionner et analyser les exemples les plus pertinents desquels nous pourrons nous inspirer.
    - Préciser les services à valeur ajoutée que les bibliothèques universitaires proposent aux chercheurs.
- Formuler des propositions sur le contenu et la forme des modules de formation ainsi que des exercices pratiques qui seront réalisés par les deux services.

- Créer des séquences pédagogiques qui détailleront le contenu, les méthodes pédagogiques, la durée, le rythme et l'articulation des cours de même que la création d'exercices pratiques abordant les diverses thématiques des modules.

1.3 Méthodologie

Les moyens mis en œuvre dans le but de réaliser ce mandat ont été dans un premier temps théoriques. Il a, en effet, été réalisé une revue de la littérature professionnelle sur les thématiques abordées comprenant des ouvrages, des travaux de recherche mais aussi des articles publiés dans des revues spécialisées.

En outre, il a été nécessaire d'analyser les documents créés par les institutions universitaires dans le domaine de l'information documentaire. Ainsi, de nombreux sites internet, documents de références, plans, guides, formations en lignes, documents descriptifs accompagnant les formations et exercices pratiques ont été collectés puis décortiqués.

Enfin, la création des modules de formation a été précédée par des rendez-vous avec les mandants dans le but de cerner précisément le périmètre de la formation, mais aussi leurs besoins et ceux des futurs participants. Le but étant de produire une formation réaliste et pertinente, ces rendez-vous ont été nécessaires afin de ne pas dévier de la trajectoire préalablement définie. À partir de là, les modules ont été créés de manière autonome en s'inspirant des formations analysées et des besoins des mandants.

1.4 Limites et contraintes

La principale difficulté de ce travail correspond à la durée très limitée des modules de formation. En effet, quatre heures en tout et pour tout s'avèrent être relativement court pour un thème aussi vaste qui est par ailleurs en constante évolution. Lorsque nous avons analysé les autres formations existantes, les modules correspondaient généralement à des cours de six à neuf heures. Il a donc été nécessaire d'aller à l'essentiel et de ne pas se perdre dans les détails afin de pouvoir survoler tous les points importants sans pour autant dépasser dans le minutage. Il a fallu donc trouver un compromis entre exhaustivité et généralité.

Une seconde contrainte a été de satisfaire deux branches scientifiques certes toutes deux médicales, mais dont les données créées et les besoins des chercheurs pouvant relativement différer. De même, les services proposés par la BIUM et l'uDDSP ne sont actuellement pas alignés et il en a résulté une adaptation aux besoins différenciés des deux institutions. Ainsi, il a été nécessaire de créer des formations qui se concentrent

sur le domaine médical afin d'intéresser les participants et de les rendre les plus pertinentes possibles tout en laissant une certaine marge de manœuvre pour les différences existantes.

2. La donnée de la recherche

2.1 Qu'est-ce qu'une donnée de recherche ?

La difficulté lorsque l'on tente de définir une donnée de recherche est que, de par sa diversité, la recherche scientifique rassemble des formats et des types de données en elles-mêmes très différentes. Le dénominateur commun ne doit donc pas être le contenu proprement dit d'une donnée de recherche mais le processus lié à sa création et à son utilisation.

Selon l'Université de Bristol citée par Rémi Gaillard, il s'agit :

> *« Des données, ou unités d'information, qui sont créées au cours d'une recherche, subventionnées ou non et qui sont organisées ou formatées de telle sorte qu'elles soient communicables, interprétables et adaptées à un traitement souvent informatisé ».* *(Gaillard 2014, p. 15)*

De son côté, l'OCDE (l'Organisation de Coopération et de Développement Economique) dans ses *Principes et lignes directrices pour l'accès aux données de la recherche financées sur fonds publics* les définit comme des « *enregistrements factuels […] source principale pour la recherche scientifique […] nécessaires pour valider des résultats de recherche ».* (OCDE 2007, p. 18)

À la lumière de ces définitions, nous pouvons synthétiser la donnée de recherche en un enregistrement factuel produit dans le cadre d'une recherche scientifique, étant à la base de l'information, de nature et format très diversifié, étant communicable et transférable afin de pouvoir valider la recherche scientifique en question.

De par leur côté communicable, les données de recherche sont fortement liées au format numérique car, toujours selon l'OCDE,

> *« C'est en effet ce format qui offre le plus de possibilités d'améliorer la distribution efficiente des données et leur application pour la recherche, dans la mesure où les coûts marginaux de la transmission de données via l'internet sont pratiquement nuls ».* *(OCDE 2007, p. 18)*

Quelques limitations doivent toutefois être amenées afin de ne pas y inclure certains éléments entrant aussi en compte dans le processus de recherche mais qui ne font pas foncièrement partie des données de recherche. L'Université de Leicester nous précise que « *Research data does not typically include data generated in the course of personal activities, desktop or mailbox backups, or data produced by non-research activities such as University administration or teaching. »* (Burnham 2013, p. 7). De plus, il importe de ne pas confondre les données en elles-mêmes et les résultats qui en

découlent. En effet, ceux-ci découlent d'une analyse *a posteriori* qui tient compte de l'interprétation du scientifique. La notion factuelle n'est alors plus garantie.

Une donnée est-elle pour autant forcément qu'à l'état brut ? Selon l'article *Data issues in the life sciences*[8], Thessen et Patterson cités par Rémi Gaillard nous disent que :

> *« Pour quelques-uns le terme "data" doit être limité aux données brutes, pour d'autres la notion inclut n'importe quel type d'information ou d'opération qui aboutit à une idée. Nous préférons limiter l'usage du terme aux données brutes, neutres, objectives, qui ne dépendent pas de leur contexte de création, d'une analyse ou de leur producteur. Dès lors qu'elles sont délimitées, filtrées et sélectionnées, elles acquièrent ou se voient donner un sens particulier dans le contexte auquel elles s'appliquent. C'est là une partie du processus qui transforme les données en information. Il n'y a pas de point clair de transition »*
>
> *(Gaillard 2014, p. 17)*

Les données brutes, appelées aussi données primaires, sont donc vierges de toute modification garantissant ainsi une certaine neutralité dans leur contenu. Néanmoins, tout traitement ne rend pas la donnée inutilisable pour autant.

Les données traitées représentent des données brutes après formatage et correction (Jacquemot-Perbal & Cosserat 2015). L'utilité principale de ces données est que le traitement permet de les rendre plus compréhensibles pour une personne tierce et permet de les normaliser en vue d'une possible comparaison avec d'autres données. Les données traitées sont donc utiles en vue de l'accompagnement d'une publication et de leur réutilisation.

Enfin, les données dérivées sont une présentation spécifique des données brutes (Jacquemot-Perbal & Cosserat 2015). Elles permettent ainsi d'avoir une vision généralement résumée, simplifiée ou compilée mais qui se veut objective des données brutes. Cela aide aussi à la compréhension de ces dernières en vue de leur reproduction.

2.2 Contenu et caractéristiques des données de recherche

Les données de recherche peuvent se distinguer de par leur nature. Il peut s'agir aussi bien d'images, de vidéos ou de fichiers audio mais aussi d'algorithmes, de données chiffrées, de statistiques, de simple texte, entre autres.

De même, leur format varie en fonction de la nature des données. Elles peuvent donc être sous format PDF, XML, Word ou Rich Text Format (RTF) pour les données textuelles, sous format XLS et ODS pour les données de tableurs, JPG, PNG, TIF pour

[8] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234430/pdf/ZooKeys-150-017.pdf

les images, etc[9] (UK Data Archive 2002c). Il est évident que certains formats sont à conseiller pour chaque type de données notamment pour préserver leur qualité (ex : résolution d'image, perte d'information due à la compression) et vis-à-vis de la problématique de leur obsolescence et donc de la pérennisation des données à long terme. (Pour approfondir ce sujet, voir le Tableau 1 : Tableau récapitulatif des formats)

Les données de recherche peuvent aussi se distinguer de par leur origine. Ainsi, la Boston University[10] ainsi que le National Science Board[11] (National Science Board 2005, p. 19) s'accordent pour en distinguer cinq catégories :

- **Les données d'observation**. Ces données sont capturées à un instant précis et son généralement irremplaçables. Puisqu'elles dépendent d'un contexte donné, une description complète est nécessaire.
- **Les données expérimentales**. Ce sont des données créées lors d'une expérience scientifique. Ces données sont généralement reproductibles mais il n'en reste pas moins que le processus peut se révéler difficile, car très onéreux (notamment pour les séquençages de gènes) mais aussi au niveau technologique ou même humain.
- **Les données de simulation ou computationnelles**. Ces données sont créées informatiquement à partir de modèles de test. Ces données peuvent être reproduites si elles sont accompagnées d'une description précise du modèle informatique. Par exemple, les simulations climatiques ou économiques.
- **Les données dérivées ou compilées**. Il s'agit de données brutes qui ont été préalablement traitées. Elles *dérivent* donc des données précédentes.
- **Les données de référence ou canoniques**. Données qui ont été organisées voire déjà publiées afin de représenter une référence pour d'autres recherches scientifiques.

Nous l'avons vu, il est nécessaire, pour que les données de recherche soient reproductibles, de définir de manière exhaustive le protocole qui a servi à les obtenir. Une fois ces conditions remplies, il est également nécessaire pour que les données soient réutilisables, d'y ajouter une documentation complète relative au contenu, au contenant, au contexte ainsi qu'aux sources de la donnée. On parle alors de **métadonnées**.

Une certaine sélection, tri et agrégation des données de recherche est nécessaire afin de « *permettre la recherche et la récupération, le traitement et la réorganisation*[12] » (CODATA-ICSTI 2013, p. CIDCR11). On parle alors d'un **jeu de données** ou **dataset**

[9] Les formats sont bien évidemment plus nombreux et plus diversifiés. Il s'agit ici que d'exemples afin de montrer l'étendue des possibilités.
[10] http://www.bu.edu/datamanagement/background/whatisdata/ (consulté le 25.03.2016)
[11] Le NSB est un conseil composé de 25 membres qui conseille la politique scientifique du gouvernement américain https://en.wikipedia.org/wiki/National_Science_Board
[12] « *to permit search and retrieval or processing and reorganizing* »

qui est une collection cohérente, ou paquet de données sélectionnées (Gaillard 2014). Les métadonnées jouent ici un rôle central dans le sens où le jeu de données, généralement téléchargeable, doit être décrit avec le plus de précision quant à son contenu, son contexte et son utilité.

2.3 Problématique et enjeux

2.3.1 Importance de la gestion des données de la recherche

Au-delà d'une notion déontologique qui encouragerait la gestion des données de recherche, leur partage et leur promotion, il est important de définir les avantages concrets que ces pratiques ont pour le chercheur, pour les institutions mais aussi pour les États, premiers pourvoyeurs de fonds pour la recherche. UNIRIS, le service des ressources informationnelles et archives de l'Université de Lausanne, distingue cinq arguments (UNIL, UNIRIS 2014, p. 8) :

- La gestion des données est **scientifiquement incontournable** : afin d'assurer la reproduction de la recherche.
- Elle est **financièrement obligatoire :** car l'octroi de fonds est généralement conditionné par la mise en place d'une gestion des données. Bien que ce point ne soit pas encore mis en pratique par le Fonds National Suisse
- Elle est **techniquement indispensable :** la gestion d'un volume de données toujours plus important nécessite une réponse organisationnelle adéquate.
- Elle est **démocratiquement essentielle :** de par la nécessité d'une transparence et un contrôle sur l'information et les données.
- Elle est **juridiquement nécessaire :** car certaines lois dans le domaine de l'archivistique interdisent la destruction non justifiée d'informations produites dans le secteur public.

Selon Begley et Ioannidis, la problématique de la reproductibilité des expériences est intimement liée à la gestion et au partage des données de recherche, chiffres à l'appui. Ainsi,

> *« An analysis of data from high-profile microarray publications that appeared between 2005 and 2006 in Nature Genetics reproduced the results of only 2 of 18 reanalyzed papers. The principal reason for failure was lack of availability of original raw data. […] An analysis of 500 papers in the 50 top journals across scientific fields (those with highest impact factors) revealed only 9% deposited full primary data online »* *(Begley et Ioannidis 2015, p. 119)*

De même, dans l'article *Out of Cite, out of Mind*, la CODATA-ICSTI[13] insiste sur le rôle de la gestion des données de recherche. Par exemple, elle nous dit que :

> *« Data have always been the cornerstone of science as it is not possible to replicate experimental findings, perform observational research, or test*

[13] « CODATA, the Committee on Data for Science and Technology, is an interdisciplinary Scientific Committee of the International Council for Science (ICSU) » http://www.codata.org/about-codata/our-mission

assertions without them. Because data often have a longer lifecycle than the research projects that create them, understanding the role of data in the research lifecycle is vital. […] Data are integral to the modern practice of research. When data are not included as part of the dissemination of scientific research, the link from the published research results back to them is broken, and the provenance and trustworthiness of the results may be in question.»

(CODATA-ICSTI 2013, p. CIDCR10)

Les données de recherche étant donc le centre de la pratique scientifique, définissant les analyses, résultats et les conclusions, une gestion responsable est désormais un enjeu central afin d'améliorer la qualité de la recherche scientifique actuelle.

2.3.2 Importance de l'Open Science

Avoir une science financée par les fonds publiques accessible à tous les chercheurs, telle est la signification du mouvement *Open Science* qui vise à améliorer globalement la diffusion du savoir scientifique. Cette diffusion se traduit par diverses actions possibles, décrites notamment dans *l'Initiative de Budapest :*

« Par « accès libre » à cette littérature, nous entendons sa mise à disposition gratuite sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et l'utilisation d'Internet »

(Institut de l'information scientifique et technique, 2004)

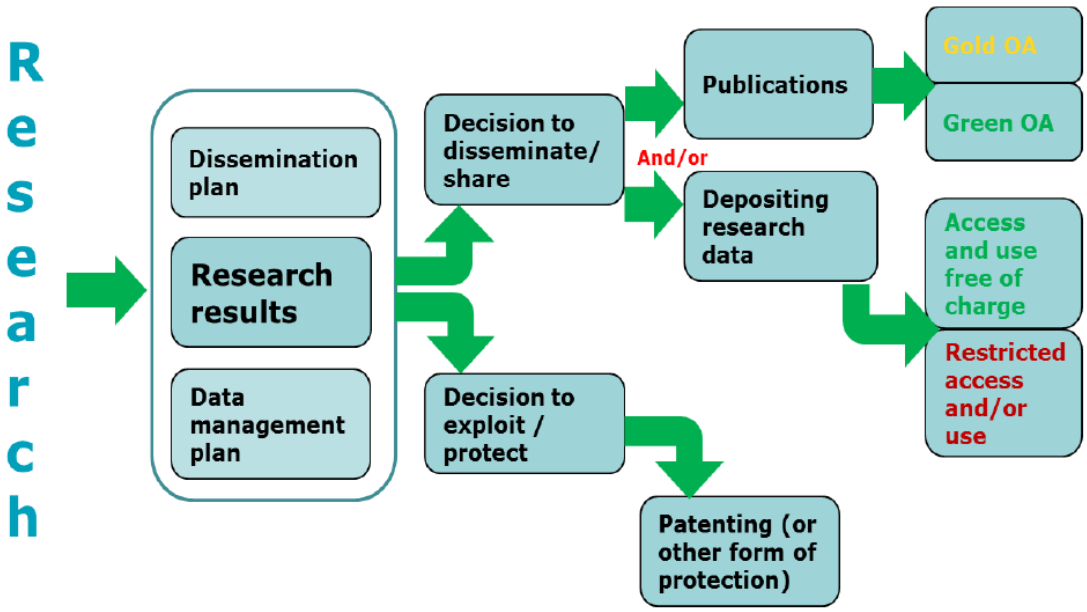
L'Open Science prend en compte le partage des articles publiés par les chercheurs (on parle alors d'*Open Access*) mais concerne également le partage des données de recherches proprement dites, que l'on nomme *Open Data*. Concernant les articles, les visées de l'Open Science peuvent être atteintes avec la publication dans une revue fonctionnant sur le modèle de l'Open Access ou encore en déposant la version de l'article « author manuscript[14] » dans une archive ouverte ou une base de données institutionnelle qui serait accessible à quiconque. Ces deux pratiques se nomment respectivement le *Gold Open Access* ou la *voie dorée* et le *Green Open Access* ou la *voie verte* (Horizon 2020 2013c).

Concernant l'Open Data, qui est le partage des données de recherche de manière libre, le but est finalement de publier ses données sur une plateforme accessible à la communauté scientifique. Pour cela, il est nécessaire de stocker ses données sur un dépôt en ligne qui existe en deux versions principales. Il s'agit des dépôts

[14] La version *author manuscript* d'un article est celle qui a déjà été « peer reviewed », c'est-à-dire contrôlée par les pairs et acceptée par la revue scientifique en question. Le contenant est donc strictement le même que la version publiée, seule la forme peut varier légèrement.

institutionnels, créés et gérés notamment par les universités, et les dépôts privés. Ces dépôts garantissent la consultation de ces données mais aussi et surtout leur téléchargement permettant ainsi la reproduction de l'expérience (Horizon 2020 2013c).

Figure 1 : Schéma de l'Open Access pour les publications et les données de recherche



(Horizon 2020 2016a, p. 3)

L'ouverture des données de la recherche est un réel enjeu pour les financeurs de la recherche scientifique. Les États paient de grandes sommes afin d'avoir une recherche scientifique dont les résultats puissent être de qualité, vérifiables et reproductibles. De ce fait, l'Open Access et l'Open Data sont indispensables et les seuls moyens efficaces pour permettre la transmission transparente des données et leur vérification. L'Open Data permet non seulement la vérification, mais aussi une recherche plus rapide en se basant sans entrave sur les données d'une recherche précédente. De plus, il apporte aussi une plus grande transparence dans la recherche pour le grand public (Horizon 2020 2013c). Enfin, il favorise la collaboration permettant, de ce fait, d'éviter les recherches à double, c'est-à-dire des recherches similaires réalisées parce que les données produites précédemment n'étaient alors pas accessibles. (Horizon 2020 2013c)

Face à cet enjeu, les États ont créé plusieurs projets visant à favoriser l'Open Science, voire à en faire une obligation. C'est notamment le cas d'*Horizon 2020*[15], le programme

[15] Le programme est doté d'un budget de 80 milliards d'euros. http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_v1.0_fr.pdf

pour la recherche et le développement de l'Union Européenne pour la période 2014-2020. Dans celui-ci, l'Open Science y est inclus de manière concrète à travers l'article 29 du *Modèle général de convention de subvention multi bénéficiaire*. Celui-ci explicite que les bénéficiaires provenant de ce programme doivent divulguer les résultats de leur recherche de manière appropriée. Ces résultats, comprennent notamment les données de recherche, ce qui est précisé au point 29.3 :

> *« En ce qui concerne les données numériques de la recherche produites dans le cours de l'action, les bénéficiaires doivent :*
>
> *(a) Les déposer dans une banque de données de la recherche et prendre des mesures afin de permettre aux tiers d'accéder aux éléments suivants et de les explorer, exploiter, reproduire et diffuser, gratuitement pour l'utilisateur […]*
>
> *(b) Fournir des informations, par la banque de données, sur les outils et les instruments à la disposition des bénéficiaires et nécessaires pour la validation des résultats (et, si possible, fournir les outils et instruments eux-mêmes). »*
>
> *(Horizon 2020 2013a, p. 69)*

Il est important de préciser que bien qu'Horizon 2020 encourage vivement l'Open Data, cette obligation n'est applicable, jusqu'en janvier 2017, qu'au projet pilote de libre accès aux données de recherche (Open Research Data Project). Or, ce projet pilote « *correspond à environ 3 milliards d'euros, soit 20 % du budget total d'Horizon 2020 pour 2014 et 2015* » (Horizon 2020 2013c, p.9).

2.4 Cycle de vie des données de recherche

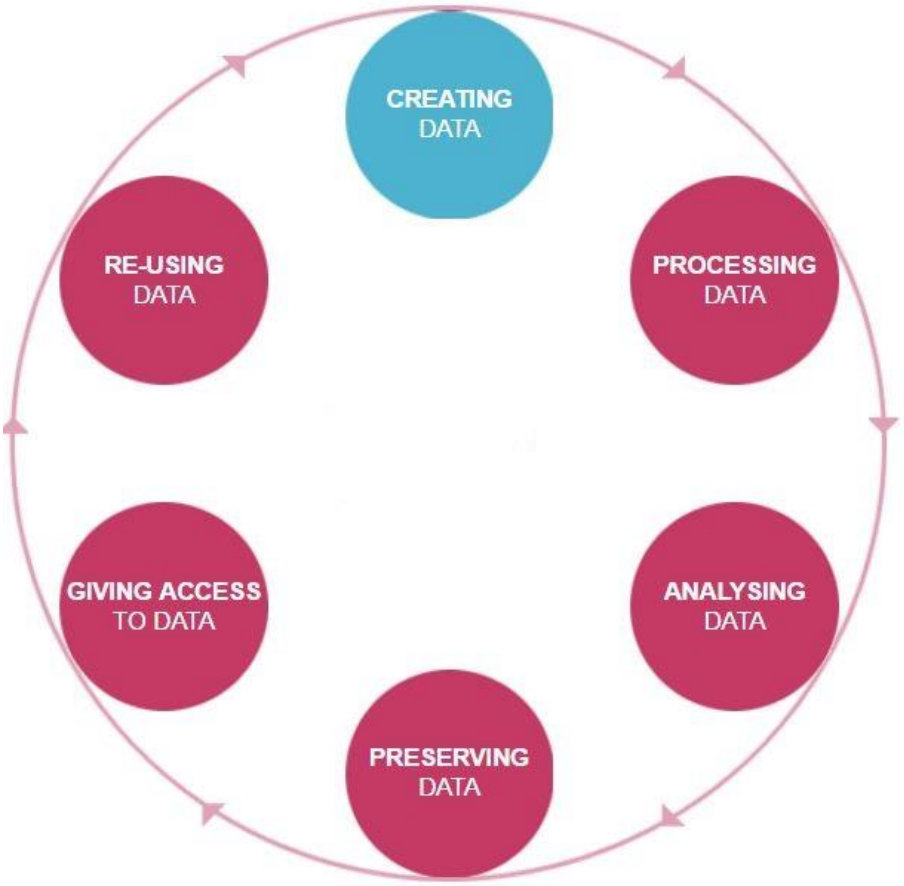
Le site *UK Data Archive*[16] a créé un modèle comprenant six étapes dans le cycle de vie des données de recherche.

- **La création de données** : comprend aussi le stade de *pré-création* des données avec l'établissement d'un Data Management Plan ainsi que la définition de la recherche proprement dite. Par la suite, cette étape comprend la collecte des données à travers les expériences (on parle alors des données brutes).
- **Le traitement des données** : comprend l'entrée, la traduction et la transcription des données. Par la suite, il s'agira de contrôler la validité des données et leur tri. C'est aussi à cette étape que l'on anonymise les données. Les données peuvent aussi être enrichies lors de cette étape et les métadonnées peuvent y être aussi ajoutées.
- **L'analyse des données** : l'interprétation des données se fait à ce moment-là afin d'infirmer ou de confirmer l'hypothèse de départ, tout comme la création des données dérivées. Le chercheur produira des résultats de recherche notamment en vue d'une possible future publication.

[16] http://www.data-archive.ac.uk/create-manage/life-cycle

- **La préservation des données** : migration dans un format idéal pour la pérennisation, création des métadonnées et de la documentation, archivage et dépôt des données sur un serveur adapté.
- **L'accès aux données** : partage des données, contrôle de l'accès aux données (Open Access total ou une mise en place d'une certaine limitation ?), définir les licences de droit d'auteur liées aux données. Cette étape comprend aussi la promotion des données.
- **La réutilisation des données** : suivi des recherches, réutilisation afin de faire de nouvelles recherches, évaluation par les pairs, croisement avec d'autres données.

Figure 2 : Le cycle de vie des données selon UK Data Archives



(UK Data Archive, 2002b)

Au regard de ce modèle, nous pouvons en déduire plusieurs éléments. Premièrement, la gestion des données ne cesse pas avec la recherche en question et l'archivage de celles-ci. Au contraire, l'utilité première du RDM ne se réalise qu'ensuite, lorsque le partage de celles-ci est effectif et que leur réutilisation permet la reproductibilité des expériences.

Deuxièmement, la responsabilité du chercheur est ici prépondérante, d'où l'importance qu'il se forme afin de savoir gérer parfaitement ses données de recherche dès les premières étapes de ce cycle de vie. L'anticipation (notamment avec la création du

Data Management Plan) semble être la clé afin d'avoir des données de recherches complètes et compréhensibles, accessibles aux autres chercheurs.

Enfin, d'autres entités telles que les institutions universitaires et scientifiques ont aussi leur part de responsabilité dans ce cycle de vie. Au-delà des formations et aides aux chercheurs, ces dernières se doivent d'avoir une démarche proactive pour trouver des solutions permettant les dépôts de données mais aussi dans la protection, la promotion et le partage de celles-ci.

2.5 Politique de gestion des données de recherche

La mise en place de ces bonnes pratiques suivies par le chercheur peut être renforcée par une démarche proactive de l'institution grâce à l'établissement d'une politique de gestion des données. Celle-ci est non seulement un appui institutionnel sur lequel il est nécessaire de compter mais est aussi importante pour la mise en place de directives concrètes pour chaque acteur intervenant dans la gestion des données de la recherche.

Mais avant même d'établir une politique de gestion des données, Graham Pryor met en évidence qu'il est nécessaire de faire évoluer les mentalités du monde académique. Pour cela, il cite les six changements dans les pratiques générales qu'a reportés la *Royal Society*[17] afin d'exploiter de manière réellement fructueuse les résultats de la recherche scientifique (Pryor 2014, p. 2) :

- Un changement de culture chez les chercheurs où les données étaient jusqu'alors vues comme une propriété privée uniquement.
- Un élargissement des critères utilisés pour évaluer la recherche afin que des crédits puissent être alloués pour le partage de données et les nouvelles formes de collaboration.
- La création de normes communes pour le partage des données.
- Une mise en libre accès des données pertinentes accompagnant les articles scientifiques publiés.
- La création d'une plateforme professionnelle intervenant dans la gestion et l'utilisation des données digitales.
- Le développement et l'utilisation de nouveaux outils logiciels afin d'automatiser et de simplifier la gestion et l'exploitation des jeux de données.

Ces remarques ne s'appliquent pas à une institution propre mais représentent les changements de mentalité nécessaires afin d'acquérir une recherche scientifique globale plus efficiente.

[17] Institution britannique promouvant les sciences naturelles notamment https://royalsociety.org/

Concernant les politiques de gestion de données proprement dites, chaque université est encouragée à en créer une, afin de générer une impulsion top-down et les moyens requis à une action efficace et coordonnée.

2.5.1 Incitations et politique des institutions

La première initiative des pays ainsi que des institutions en termes de politique est de s'engager à travers divers accords internationaux. Rappelons notamment Horizon 2020, qui, à travers son projet pilote sur le libre accès des données de recherche, impose aux chercheurs désirant un financement de leur part de déposer leurs données sur un site de stockage accessible aux tiers de manière gratuite, mais aussi de réaliser un Data Management Plan complet afin d'assurer la qualité, l'échange et la sécurité des données.

Les politiques institutionnelles, elles, déterminent un cadre comportant des directives que les acteurs devront suivre. Par exemple, l'Université d'Edinburgh a réalisé une politique en dix points[18] qui résument efficacement une gestion responsable des données de recherche, pour le chercheur mais aussi pour l'Université. En résumé, il est dit que (Edinburgh 2015b) :

- *« Les données doivent être gérées selon les normes les plus rigoureuses »*
- *« La gestion des données de recherche incombent en premier lieu au chercheur »*
- *« Toute proposition doit comprendre un plan de gestion »*
- *« L'Université fournira l'aide et le conseil nécessaires auprès du chercheur, ainsi que les mécanismes de stockage et de conservation »*
- *« Il est nécessaire d'enregistrer auprès de l'Université tout ensemble de données enregistré ailleurs »*
- *« Le chercheur doit garantir l'accessibilité et la réutilisation des données »*
- *« Le chercheur doit protéger les intérêts légitimes des individus »*
- *« Les données seront évaluées en vue d'une conservation dans un domaine national, international ou dans un dépôt de l'Université »*
- *« Les droits libres d'accès doivent être conservés »*

Il est important de noter que ces politiques de gestion des données de recherche ne sont pas obligatoires au sein des institutions universitaires. Elles sont fortement recommandées, mais des disparités existent selon les pays. Ainsi, les pays anglo-saxons semblent être ceux qui ont développé les politiques les plus exhaustives,

[18] http://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy

notamment le Royaume-Uni et les États-Unis. En Suisse, l'EPFL[19], l'UNIGE ou l'ETH Zürich[20], entre autres, sont en phase active de création et d'activation d'une telle politique.

Les bibliothèques et services de conseils académiques représentent, de par l'expertise de leurs spécialistes, dès lors une passerelle entre la politique de l'institution et le chercheur.

2.5.2 Impulsions venant des chercheurs

Au-delà des incitations provenant des institutions elles-mêmes (top-down), il est intéressant d'avoir une réflexion autour d'une incitation *bottom-up*, c'est-à-dire venant des communautés scientifiques. Une culture du partage existe préalablement dans certaines branches et les diverses associations scientifiques peuvent être de réels moteurs envers la diffusion et la réutilisation des données de recherche. Prenons comme exemple l'*International Virtual Observatory Alliance*[21] qui tente d'établir des normes techniques dans le domaine de l'astronomie. De même, *Science Open*[22] qui est un réseau ouvert permettant non seulement de publier en Open Access mais qui favorise les commentaires et les discussions autour des articles proposés.

2.6 Data Management Plan

Le Data Management Plan (DMP) ou plan de gestion des données, est l'un des outils mis en avant dans le RDM. Comme nous l'avons déjà dit, Horizon 2020 en fait l'une de ses conditions imposées au chercheur dans l'optique du projet pilote *Open Research Data*. De même, selon le *UK Data Service*[23] :

> *« Good data management is fundamental for high-quality research data and research excellence. A data management plan helps you consider at the planning stages of research how data will be managed throughout the research process and shared afterwards. »* *(UK Data Service 2002a)*

Cette mise en avant du DMP est symptomatique de son importance. En effet, il permet d'expliciter de manière concrète comment seront gérées les données produites durant tout le processus de recherche mais aussi ultérieurement à celui-ci. Le DMP couvrant une grande partie du cycle de vie des données, il doit être pensé et rédigé au commencement de la recherche afin de s'assurer d'une conformité avec la politique de gestion des données de son financeur ainsi que celle de son institution (UK Data

[19] http://research-office.epfl.ch/financements/international/horizon-2020/open-research-data-pilot
[20] http://www.library.ethz.ch/en/ms/Digital-Curation-at-ETH-Zurich
[21] http://www.ivoa.net/
[22] https://www.scienceopen.com/
[23] https://www.ukdataservice.ac.uk/manage-data/plan/planning

Service 2002a) (Horizon 2020 2013b). Selon la version les *Lignes directrices pour la gestion des données dans Horizon 2020*, les recherches prenant part au projet pilote *Open Research Data* devront fournir une première version de leur plan de gestion des données dans les six premiers mois (Horizon 2020 2013b). De plus, il est nécessaire de préciser que le DMP n'est en aucun cas un document fixe, figé dans le temps. Il se doit d'être évolutif, mis à jour au gré des changements dans la recherche concernée.

Le Data Management Plan est donc fortement lié aux principes de l'Open Access et de l'Open Data dans le sens où il les prévoit dans sa structure-même mais il ne s'agit en réalité que du résultat d'une politique responsable de gestion des données de recherche.

Concrètement, le Data Management Plan comporte plusieurs champs que le chercheur devra remplir selon les caractéristiques de sa recherche. Le modèle de DMP peut varier selon l'institution qui le crée, mais nous y trouvons généralement plus ou moins les mêmes rubriques que l'on peut résumer à partir des exemples (Deboin, 2014), (Horizon 2020 2016b), (Cartier, Moysan, Reymonet, 2015), (University of Edinburgh [sans date][24]) ainsi que le NIH (National Institutes of Health, 2003) :

- **Informations sur le projet** : présentation rapide du projet. Comprend aussi les informations administratives
- **Copyright, cadre légal et éthique** : détermine la licence sous laquelle sont créées les données. Précise aussi s'il y a un embargo, ainsi qu'une possible protection sur les données sensibles.
- **Jeux de données**
    - **Description des jeux de données** : le type de données créées, leur origine, leur format, leur volume, la possible existence d'autres données semblables réutilisées
    - **Collecte des données** : méthode de collecte et de création de données
    - **Stockage, accès et sécurité des données** : Lieux de stockage, sauvegarde, gestion des droits d'accès des créateurs.
    - **Métadonnées** : quelles sont les métadonnées créées et sous quels standards.
- **Partage des données** : description des modalités d'accès, les méthodes de diffusion. Précision sur les possibles restrictions. Durée d'un possible embargo.
- **Sélection, archivage et conservation** : informer quelles sont les données retenues et celles qui vont être supprimées. Préciser la durée de conservation si possible, le coût de stockage et le volume approximatif des données.

En conclusion, selon Rémi Gaillard, les plans de gestion des données ont une importance sur au moins trois points :

[24] http://www.ed.ac.uk/files/imports/fileManager/Edinburgh_DMP_template_web.pdf

« Ils établissent contractuellement une obligation de dépôt (à des fins de conservation et de diffusion), définissent le cycle de vie des données collectées (en statuant sur leur « valeur », et donc sur la nécessité, ou non, d'une conservation pérenne), et jouent un rôle essentiel dans la chronologie de l'ouverture en précisant dans quels délais les données pourront être diffusées »

(Gaillard 2014, p. 39)

3. Les données de la recherche biomédicale

3.1 Les fondements et caractéristiques de la recherche biomédicale

La recherche biomédicale peut se diviser en plusieurs disciplines rattachées à la recherche fondamentale, la recherche préclinique et la recherche clinique.

La recherche fondamentale représente les actions à la production de savoirs dans un domaine précis. Dans le domaine du biomédical, elle tente de comprendre le corps humain, son fonctionnement et les maladies existantes, étant ainsi à la croisée des différentes sciences du vivant.

> *« Dans les sciences du vivant, elle s'attache à connaître et comprendre les systèmes biologiques qui régissent la vie, sans se préoccuper immédiatement des applications éventuelles à court, moyen ou long terme. Elle désigne généralement une recherche menée par des équipes et organismes de recherche non industriels, le plus souvent publics »*
>
> *(Direction de la Prospective et du Dialogue Public 2013, p. 7)*

En effet,

> *« On considère d'habitude que la recherche fondamentale se consacre à l'acquisition des connaissances sans but défini d'utilité ou d'objet spécifique. »*
>
> *(OMS 2003, p.2).*

Cependant, la recherche fondamentale peut aboutir à une utilisation des connaissances créées. Il s'agit ici du rôle de la recherche translationnelle, qui agit comme un chaînon entre la recherche fondamentale et la recherche clinique en analysant l'applicabilité des résultats (Direction de la Prospective et du Dialogue Public 2013) (Q-CROC 2010).

La recherche préclinique consiste à tester et à évaluer l'utilité, l'efficacité mais aussi la toxicité d'un traitement ou d'une molécule révélés parfois grâce à la recherche fondamentale. Ces tests peuvent se faire de manière dite *in-vitro*, c'est-à-dire sur des organismes cellulaires mais aussi *in-vivo*, c'est-à-dire sur des animaux tels que les rongeurs ou les primates (Direction de la Prospective et du Dialogue Public 2013) (Jouis, Guery & Vicaut 2011). Les essais précliniques comportent trois axes (Bencheikh [sans date]) (les entreprises du médicament 2013) :

- **Les études pharmacologiques** : les composantes de la molécule et son potentiel
- **Les études pharmacocinétiques** : son assimilation et son élimination par le corps permettant de déterminer les possibles dosages à employer.
- **Les études toxicologiques** : s'intéresse aux effets secondaires mais aussi à la possible accumulation de la molécule dans les tissus du corps. Ces

expériences tentent aussi de déterminer les doses de toxicité mais aussi les possibles effets cancérigènes de la molécule notamment.

La recherche clinique fait suite à la recherche préclinique. Il s'agit de la phase où les essais précliniques ont été validés et peuvent être testés sur l'être humain. Ceux-ci sont encadrés légalement en Suisse par *l'Ordonnance sur les essais cliniques dans le cadre de la recherche sur l'être humain*[25].

La recherche clinique est divisée en quatre phases distinctes (Direction de la Prospective et du Dialogue Public 2013, p. 19-21) (Université Paris Descartes 2010, p. 1-3) (Jouis, Guery & Vicaut 2011, p. 2-5) :

- **Phase 1** : la phase 1 étudie la tolérance de l'être humain à la molécule en question. On teste celle-ci sur un nombre restreint de patients sains afin d'évaluer sa toxicité et son action sur l'organisme. Lors de cette étape, le chercheur calcule aussi la pharmacocinétique de la molécule chez l'Homme, c'est-à-dire la quantité d'absorption idéale ainsi que son élimination.
- **Phase 2** : la phase 2 examine l'efficacité du médicament auprès d'un petit nombre de volontaires malades. Lors de cette phase, il est aussi important de déterminer les effets secondaires selon la quantité de médicament donné. On obtient alors des données préliminaires qui sont importantes afin de valider l'efficacité du médicament, permettant ainsi de démarrer la phase 3.
- **Phase 3** : il s'agit ici de l'étude de l'efficacité du médicament à grande échelle. Pour cela, les essais se réalisent notamment *en double aveugle* où une moitié des volontaires testent le médicament en question et l'autre moitié testent un placebo. Cette phase est déterminante car « *Elles doivent apporter la preuve de l'intérêt thérapeutique et de l'innocuité du produit afin d'obtenir son autorisation de mise sur le marché* » (Direction de la Prospective et du Dialogue Public 2013, p. 20). C'est à l'issue de la phase 3 que les données sont soumises aux organismes décisionnaires dans le but d'obtenir une autorisation de commercialisation.
- **Phase 4** : cette dernière phase constitue les études continues faites alors que le médicament est dès lors commercialisé. Le but est ici d'approfondir les connaissances sur la molécule, notamment dans les cas rares qui n'ont pas été détectés lors de la phase 3.

3.2 Caractéristiques des données biomédicales

De par leur diversité et de leur complexité, il est difficile de faire une liste exhaustive des données de recherche biomédicales. Afin d'avoir un aperçu pertinent dans le cadre de ce travail, nous avons décidé de nous intéresser tout particulièrement aux types et aux formats de données utilisés entre autres à la faculté de biologie et de médecine de l'Université de Lausanne et au CHUV.

[25] https://www.admin.ch/opc/fr/classified-compilation/20121176/index.html

3.2.1 Les Omics

Les Omics sont les champs d'étude correspondant aux gènes, aux ARNs, aux métabolites, aux protéines entre autres (Micheel, Nass & Owen 2012). Les Omics sont des données générées en recherche fondamentale, translationnelle, préclinique mais aussi en recherche clinique. Utilisant des nouvelles technologies informatiques, les Omics sont donc des disciplines récentes, créant des données qui sont caractérisées par leur volume extrêmement important, notamment dû au séquençage, permettant de créer notamment des données computationnelles (Micheel, Nass & Owen 2012). Ces *Big Data* nécessitent donc des outils particuliers pour être gérées dû notamment à leur diversité et à leur volume. Il est nécessaire d'employer des plateformes dédiées ainsi que des outils de bioinformatiques performants afin de croiser les données en vue d'obtenir des résultats probants. Ces données sont donc lourdes à gérer mais aussi à stocker. Les institutions se doivent donc d'anticiper ce phénomène en proposant des lieux de stockage de taille suffisante ainsi que des procédures de traitement bien définies (Micheel, Nass & Owen 2012).

Les types d'Omics sont nombreux. Harel et al. proposent une liste non-exhaustive des champs d'étude :

> *« **Genomics**: The genes' sequences and the information therein*
> ***Transcriptomics**: The presence and abundance of transcription products*
> ***Proteomics**: The proteins' sequence, presence and function within the cell*
> ***Metabolomics**: The complete set of metabolites within the cell*
> ***Localizomics**: The subcellular localization of all proteins*
> ***Phenomics**: High-throughput determination of cell function and viability*
> ***Metallomics**: The totality of metal/metalloid species within an organism*
> ***Lipidomics**: The totality of lipids*
> ***Interactomics**: The totality of the molecular interactions in an organism*
> ***Spliceomics**: The totality of the alternative splicing isoforms*
> ***Exomics**: All the exons*
> ***Mechanomics**: The force and mechanical systems within an organism*
> ***Histomics**: The totality of tissues in an organ »*
>
> *(Harel et al. 2011, p. 79)*

De par leur importance, le **versioning** des données Omics est prépondérant. Celui-ci se traduit par des mises à jour incrémentales ou complètes des bases de données permettant néanmoins d'accéder aux versions précédentes des data. Toujours selon Harel et al., les bases de données deviendraient rapidement obsolètes si des mises à jour n'étaient pas régulièrement effectuées. La planification du versioning est rendue

difficile par le fait qu'elle doit prendre en compte la multiplicité des sources mais aussi par les caractéristiques mêmes des Omics :

« Incremental updates of this sort, along with triggered reports to users, are very attractive. In practice, this is often extremely difficult to implement in Omics applications, due to widespread data complexities, exacerbated by unpredictable source data format changes, as well as the interdependencies of many of the major data sources ». *(Harel et al. 2011, p. 83)*

3.2.2 L'imagerie

L'imagerie est l'une des composantes les plus répandues dans le type de données de recherche biomédicales. Celle-ci peut être réalisée à tous les stades de la recherche.

Lors des études fondamentales et précliniques, il s'agit alors généralement de microscopie, autrement dit des images prises par microscopie optique ou électronique, d'éléments de petite taille.

Lors d'études cliniques, nous parlons alors de l'imagerie médicale. Cette dernière est composée des IRMs (imagerie à résonnance magnétique), des radiographies, des échographies ainsi que des PET Scans (Positron Emission Tomography), qui permet d'avoir des images en trois dimension d'un organe en utilisant un traceur radioactif entre autres. (Bruel & Gallix 2007)

L'imagerie médicale peut être animée (il s'agira alors de vidéos) ou figée. D'une manière générale, ce type de données est lui aussi très volumineux et nécessite de grosses capacités de stockage.

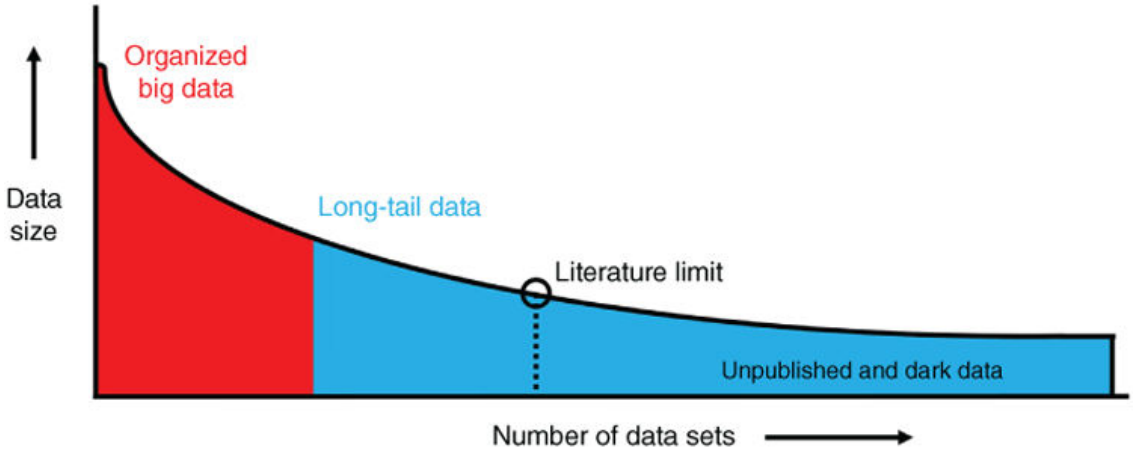
3.2.3 Données de laboratoire « long tail »

Les données de laboratoire sont les données produites par les chercheurs individuellement. Il s'agit de données qui sont beaucoup plus modestes en volume, générées à petite échelle, mais qui sont beaucoup plus fréquentes et donc beaucoup plus nombreuses au total. Cette myriade de données créées à petite échelle mais formant la majorité des données de recherche est un phénomène qui peut être catégorisé comme une longue traîne, ou « Long Tail » en anglais. Ces données ont été historiquement traitées comme un simple supplément au *Big Data*, dont l'utilité n'était pas aussi importante et dont la gestion et le partage était secondaire. Or, potentiellement, ces données constituent un savoir aussi important que les Big Data. Leur gestion est toutefois difficile car ces données sont rarement standardisées (Ferguson et al. 2014).

Toujours selon Ferguson et al., la longue traîne de données est aussi composée des « Dark Data », c'est-à-dire des données non publiées dues à l'échec de l'expérience

ou à des résultats non probants (Ferguson et al. 2014). Le partage de ses données apporte aussi quelques bénéfices, permettant notamment de ne pas refaire des expériences que l'on sait dorénavant inutiles ou de ne pas reproduire des erreurs dans la méthodologie.

Figure 3 : Principe de la longue traîne pour les données biomédicales



(Ferguson et al. 2014, p. 15)

3.2.4 Données médicales - biobanque

Les données liées aux tests cliniques courants, nommées données médicales, sont composées des prélèvements directs auprès de patients, notamment de tissus mais aussi d'échantillons de sang. Ces prélèvements représentent, cumulés, une quantité très importante de données qui peuvent être utilisées pour la recherche. Ces données peuvent être génétiques, permettant le séquençage du génome des patients mais aussi non-génétiques, s'intéressant aux cellules cancéreuses ou aux protéines (Jacquemont 2016). Ces données sont conservées dans une biobanque, accessible aux chercheurs de l'institution en question à condition que le projet de recherche y a été validé. Étant donné que les chercheurs travaillent sur des données médicales, c'est-à-dire appartenant à des personnes, des lois strictes régulent leurs autorisations (Jacquemont 2016).

3.2.5 Formats

La problématique des formats de données de recherche est intimement liée à leur pérennisation, leur transmission et à leur qualité. Il est donc encouragé d'utiliser des formats non-propriétaires, qui ne dépendront d'un logiciel ou d'une compagnie mais aussi qui pourront être lus par le plus de matériels possibles. Quant à la qualité, la question se révèle importante pour les données sous forme de médias puisqu'il n'est pas rare de sacrifier une partie de la qualité, et donc de l'information due à la compression, afin de réduire le poids du fichier. Concernant ce point, le choix est

toujours une affaire de compromis selon les besoins et les capacités des services. Enfin, les formats se doivent de supporter les métadonnées afin de pouvoir décrire de manière précise le contenu mais aussi le contexte de création (University of Edinburgh 2015c).

3.2.5.1 Imagerie

Selon *The University of Edinburgh*[26],le site *Digital Preservation*[27] et *UK Data Archive*[28], les formats d'images conseillés sont les formats non propriétaires tels que le TIFF et le JPG 2000. Le TIFF est un format utilisé par les professionnels de l'images permettant d'avoir une image de très bonne qualité ainsi que des métadonnées complètes mais dont les fichiers sont très volumineux.

D'autres formats existent pour l'imagerie tels que le JPG, le PDF ou le PSD. Ceux-ci sont parfois utilisés mais déconseillés car la pérennisation et la qualité des données n'est pas garantie, en raison d'une trop forte compression pour certains ou le fait qu'ils soient sous formats propriétaires pour d'autres (notamment Photoshop) (UK Data Archive 2002c) (University of Edinburgh 2015c).

L'imagerie est également composée de vidéos. Le format conseillé dans le but de garantir la qualité et la pérennisation des données est le MPEG-4

3.2.5.2 Tableurs

Les formats conseillés pour les données sous forme de tableurs sont Comma-Separated Values (CSV) ainsi que le Tab-Delimited Values (.tab).

Malgré tout, les formats Excel (.xls) et OpenDocument Spreadsheet (.ods) sont les plus répandus. Ils sont généralement acceptés de par leur diffusion importante et leur utilisation très fréquente mais ils ne sont pas conseillés car trop dépendants des logiciels pouvant les lire (UK Data Archive 2002c) (Université of Edinburgh 2015c).

3.2.5.3 Fichiers Textuels

La problématique des fichiers textuels est la même que pour les tableurs. Les fichiers les plus largement utilisés sont créés dans les formats de Word (.docx) et OpenDocument text (.odt). Or, on y rencontre les mêmes défauts que pour les tableurs, soit un problème de pérennisation dû aux logiciels nécessaires à leur lecture.

Ainsi, les formats conseillés sont le Rich Text Format (.rtf) ou encore le Text (.txt) (UK Data Archive 2002c) (University of Edinburgh 2015c).

[26] http://www.ed.ac.uk/files/atoms/files//recommended_file_formats-apr2015.pdf
[27] http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml
[28] http://www.data-archive.ac.uk/create-manage/format/formats-table

Tableau 1 : Tableau récapitulatif des formats de fichiers

Type de données	Formats conseillés	Autre formats existants
Images	TIFF, MPEG-4, JPEG2000	JPG, PNG, PDF, PSD, AVI
Tableurs	CSV, TAB	XLS, ODS, MDB
Fichiers textuels	TXT, RTF, XML	DOC, ODT
Fichiers audio	FLAC	MP3, WAV

3.2.6 Données de recherche en santé publique

La recherche en santé publique est à la croisée entre les sciences sociales et les sciences médicales. Le principe est de récolter des données médicales à grande échelle, c'est-à-dire au niveau de la population afin d'établir des statistiques sur la santé de ladite population. Les statistiques en santé publique représentent alors un outil indispensable dans la création d'un système de santé adapté et des mesures efficientes (Haut conseil de la santé publique 2009).

En Suisse, la collecte et le traitement de ses données sont le fruit de plusieurs institutions publiques telles que l'Office fédérale de la statistique, les hôpitaux ou les instituts de santé publique[29] mais aussi diverses institutions privées telles que les assurances maladies, autant de sources statistiques indépendantes les unes des autres (Santé Publique Suisse 2013). Afin d'obtenir des statistiques cohérentes et reflétant le plus fidèlement possible la réalité au niveau national, il est indispensable de pouvoir partager et croiser les données recueillies. Or, selon le manifeste *Des données de meilleure qualité pour augmenter l'efficacité du système de santé* publié par *Santé publique Suisse*[30], il existe actuellement quatre domaines où des améliorations doivent être réalisées afin d'avoir un partage efficient de ces données (Santé Publique Suisse 2013) :

- **L'appariement des données** : Les données des divers producteurs doivent être regroupées. Cet appariement peut se faire de manière ciblée selon les projets de recherche où seule une statistique peut être utilisée.
- **La compatibilité des données** : Les outils et formats utilisés pour la collecte et le traitement des données ne sont actuellement pas homogénéisés provoquant un problème de compatibilité. Il est alors indispensable de standardiser la collecte de données.
- **La disponibilité des données** : Les données provenant des instituts privés ne sont que rarement disponibles pour les chercheurs. *Santé publique Suisse* propose ainsi un pacte de solidarité entre les chercheurs et les assurés où ces derniers peuvent décider s'ils veulent que leurs données personnelles (entre

[29] http://www.public-health.ch/logicio/pmws/publichealth__institut__fr.html
[30] Santé publique Suisse est l'organisation indépendante nationale qui défend le point de vue de la santé publique

temps anonymisées) puissent être utilisées dans la recherche en santé publique.

- **L'intégralité des données** : Il manque actuellement en Suisse une enquête à très grande échelle, c'est-à-dire à au moins 100'000 personnes et à long terme dans ce domaine. Les échantillons représentatifs n'étant pas toujours suffisamment pertinents pour élucider certaines causes de maladies, il est indispensable de réaliser des statistiques à une échelle globale.

Selon Pisani & AbouZahr, un obstacle au partage des données proviendrait des politiques publiques elles-mêmes, réticentes à afficher les possibles échecs de leur politique de santé. Ce frein hypothétique n'est cependant pas forcément présent en Suisse et doit être relativisé car selon les auteurs, celui-ci surtout significatif pour les statistiques internationales (Pisani & AbouZahr 2010).

3.3 Partage des données biomédicales : bénéfices et inquiétudes

Le domaine de la recherche biomédicale est relativement conservateur dans le partage des données de recherche (Anderson et al. 2007). Selon les auteurs, les chercheurs maintiennent une certaine tradition protectrice de leurs données freinant ainsi leur partage. De plus, l'utilisation répandue d'outils informatiques limités comme les tableurs de bureautique, prisés pour leur souplesse d'utilisation, ne facilite pas une future homogénéisation et classification des données en vue de leur conservation et de leur partage (Anderson et al. 2007). Ce phénomène est aussi dû à plusieurs facteurs liés aux caractéristiques mêmes des données de recherche biomédicale. Toujours selon ces auteurs, la recherche biomédicale voit actuellement un accroissement constant de la quantité mais aussi de l'hétérogénéité des données, rendant le travail de traitement plus fastidieux pour le chercheur (Anderson et al. 2007).

Historiquement, la compétitivité entre les chercheurs est très élevée dans ce domaine de recherche. Le partage de données et leur contrôle n'y est alors pas toujours pris comme une qualité mais comme une ingérence dans leur travail :

> *« Data sharing may lead to the fear that others will uncover errors in the data or question the validity of the analysis. Unlike the open source software community, where error correction is encouraged and welcomed, uncovering errors in scientific data may be perceived as an attack on a researcher's reputation. In the hypercompetitive environment of biomedicine, such attacks may lead to hard feelings, finger pointing and a competitive disadvantage. In recognition of such potential abuses, data sharing has contributed to the advocacy for development of normative practices as to how researchers raise issues of errors in a manner that encourages open dialog ».* *(Ferguson et al. 2014, p. 5)*

Au-delà des avantages communs au partage de tout type de données vus précédemment, l'une des spécificités argumentant fortement en faveur du partage des

données biomédicales, est leur dépendance au contexte dans lequel elles ont été récoltées. Autrement dit, plus encore que pour d'autres domaines scientifiques, les méta-analyses permettent de confirmer ou d'infirmer une conclusion selon des variations de ce contexte et permet de mettre en évidence des particularités, ou des effets secondaires inédits. Ainsi, la zone géographique, le type de population testée mais aussi le climat sont des critères vérifiables qu'avec le partage des données (Bull et al. 2015).

De plus, toujours selon Bull et al., le domaine de la santé étant directement connecté à la population (tests cliniques, soins, etc.) le partage permet une certaine démocratisation, une certaine transparence nécessaire à la confiance en ce domaine scientifique (Bull et al. 2015).

La singularité des données biomédicales vis-à-vis d'autres types de données sont les questionnements d'ordre éthique que peut amener leur partage, notamment lors de recherches cliniques. Le partage de données de patients nécessite un accord de la personne concernée. Cet accord peut se traduire en trois degrés de consentement de la part du patient qui permettent une utilisation plus ou moins libre de ses données personnelles. Ainsi, selon Hate et al. :

> *« We discussed three types of consent. Broad consent implied that participants were told that **their data might be shared after use** in the index study and that they would not be contacted for an opinion on sharing. The research organization would generally stand proxy for the participant in deciding whether sharing was appropriate. Middle consent implied that participants were told that **their data might be shared with people working in specific research areas** related to the study. Explicit consent implied **that participants would be contacted for an opinion** whenever there was a request for sharing ». (Hate et al. 2015, p. 245)*

Le deuxième questionnement d'ordre éthique est l'anonymisation des données avant partage. Cette anonymisation est, en effet, un enjeu primordial dans la préservation du secret médical. Elle doit donc être garantie, nécessitant un travail supplémentaire de la part du chercheur. Ainsi :

> *« This literature, in both academic and policy domains, has highlighted not only the fact that data sharing presents a range of ethical challenges not previously encountered but also the challenges of taking seriously both ethical arguments for sharing data and those supporting the development of appropriate governance models and mechanisms to ensure the protection of the interests of participants, communities, and the scientists who produce and share data […] These include concerns about the effectiveness of measures to de-identify data, and to protect the privacy of participants ». (Parker and Bull 2015, p. 218)*

Plusieurs éléments des données doivent être anonymisés. Abou El Kalam et al. citent non seulement les éléments identitaires de la personne (nom, prénom, date de naissance, numéro de sécurité sociale, adresse) mais aussi certains éléments indirects

qui permettent d'identifier la personne « *par simple rapprochement* » tels que les dates d'entrée et de sortie de l'hôpital, les dates d'accouchement, etc. (Abou El Kalam et al. 2004a).

L'anonymisation des données peut être réalisée sur trois niveaux décrits :

> « ***Réversibilité****: cacher les données par un simple chiffrement des données. Dans ce cas, il y a possibilité de remonter depuis les données chiffrées jusqu'aux données nominatives originelles.*
>
> ***Irréversibilité*** *: c'est le cas réel de l'anonymisation; une fois remplacés par des identifiants anonymes, les identifiants nominatifs originels ne sont plus recouvrables. […]*
>
> ***Inversibilité****: c'est un cas où il est impossible en pratique de remonter aux données nominatives, sauf en appliquant une procédure exceptionnelle sous surveillance d'une instance légitime (médecin-conseil, médecin inspecteur) garante du respect de la vie privée des individus concernés; il s'agit cette fois-ci d'une pseudonymisation au sens des critères communs ».*
>
> *(Abou El Kalam et al 2004a, p. 4-5)*

La réversibilité est donc un encodage des données permettant un retour un arrière, une forme de dé-anonymisation. Celle-ci se fait à l'aide d'une clé algorithmique et peut être très importante pour le patient, dans le cas où l'on désirerait retrouver sa trace dans le cadre d'une thérapie liée à sa maladie. À contrario, l'irréversibilité, comme son nom l'indique, ne permet pas un retour en arrière. L'anonymisation y est donc assurée *ad aeternam* mais rend aussi impossible la récupération de l'identité d'un patient en vue d'une possible thérapie.

Les données primaires vont alors être anonymisées selon l'un de ces trois stades et vont être regroupées formant ainsi des données agrégées pouvant être étudiées statistiquement.

Une inquiétude supplémentaire consiste en la stigmatisation d'une communauté ou d'une population donnée. Dans leur article consacré au partage de données biomédicales indiennes, Hate et al. insistent sur le fait que « *Data use might also result in a pejorative presentation of a community, despite anonymization* » (Hate et al. 2015, p. 243), notamment pour les différentes castes indiennes. En effet, de simples critères géographiques permettent de mettre en avant une population. Or, ce même problème peut exister dans les pays occidentaux où certaines communautés minoritaires peuvent être mises en avant pour de bonnes ou de mauvaises raisons, notamment dans la médecine sociale qui utilise ces critères sociogéographiques.

De nombreux bénéfices et inquiétudes peuvent résulter du partage des données biomédicales. Bull et al. proposent un tableau complet des éléments qu'ils ont pu noter selon leur expérience dans le domaine.

Figure 4 : Tableau récapitulant les bénéfices et les inquiétudes du partage des données biomédicales

Reasons to share individual-level data	Concerns about sharing individual-level data
To improve science	May hamper science
• Enable verification, replication, and expansion of research results • Address biases, deficiencies, and dishonesty in research • Enable novel analyses and increase study power • Improve meta-analyses • Maximize data use, particularly for datasets that cannot be replicated • Inform research design and research funding • Improve teaching resources • Increase primary data producers' academic profiles and collaboration opportunities	• Reputational harms of critical secondary analyses • Consequences of flawed/poor quality secondary analyses • Reduction of incentives for primary research • Increased incentives to conduct short-term research rather than long-term research • Opportunity costs of curating and sharing data
To improve health	May hamper health
• Inform health care planning and allocation • Inform regulatory review • Improve evidence base for clinical decision making • Improve use of health care resources • Improve patient care	• Effects of flawed secondary analyses on scientific evidence base • Burden of evaluating validity of secondary analyses • Effects of second-guessing regulatory procedures, policies, and processes
Explicit moral claims	Explicit ethical issues
• Importance of maximizing the value and utility of data • Promotion of scientific values • Promotion of best practices in research conduct, analysis, and reporting • Demonstration of respect for research participants • Promotion of the public good	• Protection of participants' privacy and confidentiality • Validity of consent, including broad consent • Potential harms of secondary research for research participants including discrimination and stigma • Researchers' ability to fulfill commitments made to research participants during data collection • Effects of moral distance and limited awareness of the context in which data were collected • Potential impacts on public trust and confidence of conflicting analyses • Balancing the interests of differing stakeholders in data sharing • Making best use of limited research resources
	Barriers to sharing
	• Costs of developing and maintaining appropriate expertise and infrastructure • Curation costs • Ownership, intellectual property rights, and commercial confidentiality • Lack of policies and processes

(Bull et al. 2015, p. 226)

4. Contexte des institutions

4.1 IUMSP

L'Institut universitaire de médecine sociale et préventive (IUMSP) est un institut rattaché au CHUV et à la Faculté de biologie et de médecine de l'université de Lausanne, ainsi que membre de l'école romande de santé publique (ERSP)[31], le pôle romand de la Swiss School of Public Health[32]. Sa mission est « *d'élaborer les réponses adéquates aux besoins de santé de la population et accompagner leur mise en œuvre* »[33]. Pour cela, les activités de l'IUMSP s'articulent autour de trois points définit sur son site internet[34] :

- ***Générer de nouvelles connaissances*** *: à travers la recherche scientifique qui concernent les domaines suivants :*
  - *Epidémiologie et la prévention des maladies chroniques en particulier les cancers et les maladies cardio-métaboliques.*
  - *Organisation des systèmes de santé, concernant en particulier la population âgée et l'évaluation des soins.*
  - *Développement de méthodes quantitatives en santé publique et en médecine, en particulier les méthodes biostatistiques.*
- ***Transmettre les connaissances et assurer un haut niveau d'expertise*** *: à travers l'offre de formations universitaires ainsi que les formations continues et post graduées.*
- ***Accroître les interactions entre recherche et pratique*** *: à travers la réalisation de mandats (IUMSP sans date)*

L'IUMSP comprend deux pôles distincts, le premier concernant les unités recherche proprement dites, et le second concernant les unités de service (IUMSP 2015b).

4.1.1 Formations et services existants

L'unité de documentation et données en santé publique (uDDSP) est l'une des unités de service, en relation directe avec les unités de recherche. Sa mission est de « *faciliter l'accès aux informations et données (publications, indicateurs, données brutes issues de la recherche) relatives à la santé publique* » (IUMSP 2015b, p. 36) et gère aussi la bibliothèque et le site web de l'IUMSP.

L'accès aux données de recherches générées à l'IUSMP se fait via une plateforme internet créée et gérée par l'uDDSP nommée Data@IUMSP[35]. Cette plateforme est à la fois un dépôt pour les données des chercheurs de l'Institut ainsi qu'un service de

[31] https://www.iumsp.ch/fr/ersp/domaines-activites
[32] http://www.ssphplus.ch/sharepoint/startseite_fr.html
[33] https://www.iumsp.ch/fr/a-propos
[34] https://www.iumsp.ch/fr/a-propos
[35] https://data.iumsp.ch/home

partage de ces données. Il est à noter que ces dernières ne sont pas en Open Access, les modalités d'accès étant choisies par le responsable de chaque recherche au niveau du set de données déposé. Les niveaux d'accès possibles sont les suivants :

- **Données non accessibles** (Data not available)
- **Accès direct aux données** (Direct Access Data Files) **:** les données peuvent être téléchargées sans restriction.
- **Utilisation publique des données** (Public Use Data Files) **:** nécessite que la personne voulant les télécharger soit enregistrée et accepte les conditions d'utilisation. Le dépôt conserve une trace de la personne qui les télécharge.
- **Données autorisées** (Licensed Data Files) : nécessite que la personne voulant les télécharger soit enregistrée et soumette un formulaire détaillant précisément les raisons de l'accès aux données. Les personnes responsables décideront alors si elles acceptent de partager leurs données ou non.
- **Données disponibles dans une « enclave »** (Data available in an Enclave) : aucune donnée n'est partagée à partir du dépôt. Les utilisateurs soumettent une demande d'accès aux données à une installation sécurisée à laquelle le producteur a accès.
- **Données disponibles via un dépôt externe** (Data available from external Repository) : seules les métadonnées sont accessibles sur le dépôt accompagnées d'un lien vers un dépôt externe.

L'unité de documentation assure les formations auprès des chercheurs dans le domaine documentaire. Actuellement, aucune formation standardisée n'est proposée et celles-ci sont réalisées au cas par cas, lorsque le chercheur manifeste un besoin d'approfondir ou découvrir une pratique documentaire. Il s'agit notamment[36] :

- Aide à l'acquisition, la gestion et à la publication des données de recherche
- Aide à l'anonymisation des données
- Aide à la création des métadonnées
- Aide à la création du Data Management Plan
- Aide à la recherche documentaire dans les bases de données spécialisées
- Utilisation des logiciels bibliographiques

De plus, l'uDDSP gère bien évidemment la bibliothèque de l'IUSMP et l'accès aux revues spécialisées entre autres.

L'aide à la création des métadonnées est proposée à partir du standard DDI (Data Documentation Initiative)[37], au format XML, qui définit les besoins en description pour les enquêtes et sondages en sciences humaines et sociales. Les descriptions des métadonnées vont donc être réalisées à deux niveaux. Un niveau global dans l'optique

[36] https://www.iumsp.ch/fr/uddsp/services-iumsp
[37] http://www.ddialliance.org/

de créer un DMP, ainsi qu'à un niveau fin afin de respecter le standard DDI. En effet, Le document de travail du standard DDI, nommé *Codebook :*

> *« Includes the same information as a traditional codebook, describing variables, question text, and the categories and codes used as response domains and the values of variables. It also captures some other information about the data set »*
>
> *(Arofan 2011, p. 2)*

Ce *Codebook* va donc détailler les données d'enquêtes et les informations nécessaires à leur compréhension mais aussi assurer la fiabilité des traitements statistiques liés à ces données. Chaque variable sera alors précisée par son format, son périmètre de recherche, ses conditions de création, etc. Enfin, le standard DDI permet d'avoir un vocabulaire contrôlé précisant davantage la nature des données (DDI 2015). À noter enfin que l'outil informatique de gestion des métadonnées conseillé par l'uDDSP est **Nesstar Publisher**[38] permettant d'éditer facilement ses données et évitant aux chercheurs de travailler en XML brut.

Toujours au niveau informatique, l'unité est responsable du site Web et des outils d'acquisition, de partage ainsi que de l'anonymisation des données de recherche. Cette dernière est réalisée avec diverses techniques (dé-identification, cryptage, Statistical Disclosure Control) réversibles ou irréversibles selon les besoins.

Concrètement, le service encrypte les données afin d'ôter toute information et identifiant personnels avant la publication. L'anonymisation consiste en la suppression d'informations sensibles ou en un ré-encodage irréversible, en utilisant des algorithmes pour « hacher » les données afin de rendre illisibles les informations personnelles (Abou El Kalam 2004a). Toutefois, une réversibilité de l'anonymisation et souvent nécessaire pour pouvoir réutiliser les données et les lier avec celles d'autres études en cas de besoin. Ainsi, la dé-identification est aussi utilisée. Dans ce cas-ci, les données sont certes anonymisées, mais un lien rétroactif peut être effectué afin de retrouver les identifiants par les personnes autorisées (TransCelerate BioPharma 2013).

Enfin, l'uDDSP met en place l'infrastructure web pour la collecte de données dans les projets de recherche avec un outil de gestion d'enquêtes en ligne nommé *LimeSurvey*[39].

4.1.2 Recherche existante et données créées

Les recherches produites à l'IUMSP sont à la croisée entre les sciences humaines et les sciences dites dures. La production de savoirs se fait via deux processus : il s'agit

[38] http://www.nesstar.com/software/publisher.html
[39] https://survey.iumsp.ch/

soit de mandats pour des services publics de santé tels que l'Office fédéral de la sécurité alimentaire et des affaires vétérinaires (OSAV), le service de la santé publique du Canton de Vaud (SSP) ou encore l'Office fédéral de la santé publique (OFSP), soit de la recherche pure financée notamment par le Fonds national suisse (FNS) (IUMSP 2016b).

Les réflexions présentes dans ces projets abordent les thématiques suivantes :

> *« Contribuer à prévenir les maladies chroniques, notamment par la recherche universitaire sur les maladies cardiométaboliques comme les affections cardiaques, le diabète ou le cancer, ainsi que par la réalisation d'expertises. »*
>
> *(IUMSP 2016b, p. 22)*

> *« Développer la recherche et les connaissances sur les services de santé, et diffuser les résultats obtenus afin qu'ils soient pris en compte dans les décisions de politique sanitaire. »* *(IUMSP 2016b, p. 24)*

> *« Favoriser le transfert de connaissances scientifiques fiables et actualisées auprès des professionnels de la santé, de la population, des patients et des décideurs ; évaluer la qualité des soins et la sécurité des patients ; développer et évaluer de nouvelles modalités de prévention et de prise en charge de soins coordonnés des personnes vivant avec une maladie chronique. »*
>
> *(IUMSP 2016b, p. 26)*

> *« Comprendre, maîtriser, développer, appliquer et enseigner les méthodes statistiques ; étudier les façons optimales de traiter des données, à la fois d'un point de vue mathématique et pratique. »* *(IUMSP 2016b, p. 28)*

La spécificité de la recherche à l'IUMSP se reporte dans les types de données créées, relativement différentes de la recherche biomédicale. Ces données sont produites à partir de questionnaires soumis aux patients/sondés ou par des tests médicaux, notamment concernant les données biologiques des maladies chroniques.

Les formats en découlant sont généralement des tableurs synthétisant les réponses aux questionnaires, des données biologiques ou encore des données au format image. Cela signifie que jusqu'à récemment, les données de recherche produites par l'IUMSP n'étaient pas encore massives et pouvaient être gérées relativement facilement. Ainsi, les données étaient gérées par les différents services, responsables de leurs projets de recherche provoquant alors une certaine décentralisation de la gestion des données.

Cela est toutefois en train de changer car de nouveaux types de données sont désormais produits au sein de l'IUMSP. Il s'agit notamment de données génomiques qui sont beaucoup plus volumineuses. Il est, dès lors, nécessaire de centraliser la gestion des données via le dépôt unique sous la responsabilité de l'unité de documentation.

4.2 BIUM – CHUV

La Bibliothèque universitaire de médecine du CHUV, située dans le bâtiment principal du centre hospitalier, est rattachée à la faculté de biologie et de médecine de l'université de Lausanne. Ses services sont principalement destinés au personnel médical du CHUV, aux étudiants et professeurs en médecine ainsi qu'aux chercheurs appartenant à la FBM (Besse 2015). Sa collection est donc principalement consacrée aux domaines de la médecine fondamentale, de la médecine clinique et des sciences biomédicales (Pochon 2000). Les chercheurs de la FBM se répartissent entre les sections des sciences cliniques (SSC) et les sections des sciences fondamentales (SSF).

4.2.1 Formations et services existants

Dans ce contexte, plusieurs formations et soutiens ont été créés afin de répondre aux besoins des divers publics de la bibliothèque. La BIUM propose ainsi des formations à la recherche documentaire, générale ou dans les bases de données spécialisées, une formation aux logiciels de gestion des sources bibliographiques (Endnote et Zotero) et divers tutoriels et guides.

À la vue de l'importance des publications sur la production scientifique de la FBM, la BIUM a créé une nouvelle unité d'aide aux chercheurs *FBM Publication Management Unit* qui propose dorénavant une aide à la gestion des articles et des données de recherche ainsi qu'une aide à la valorisation de la recherche effectuée. Cette aide se traduit par :

- Le développement de projets favorisant une meilleure gestion de l'information scientifique pour la faculté (ex : ORCID institutionnel, nouveau développement sur le serveur institutionnel Serval, etc.).
- Le développement d'une stratégie Open Access et Open Data au niveau de la faculté en collaboration avec l'UNIL.
- Une aide personnalisée sous forme de consultations et de gestion de dossiers pour les chercheurs de la FBM.
- Des présentations d'information scientifique dans tous les départements de la FBM
- Une documentation écrite et des pages internet expliquant les divers services proposés et les enjeux sous-jacents.

Concernant plus spécifiquement le domaine de la gestion des publications scientifiques et des données associées à la publication, l'unité propose diverses aides et met en place des solutions pratiques telles que :

- Veille des publications scientifiques en Open Access de la FBM et gestion des références des chercheurs dans le dépôt institutionnel Serval

- Soutien pour la réalisation de l'Open Access et la mise en dépôt dans le partage des articles et des données selon les diverses possibilités offertes au chercheur (Gold Road ou Green Road).
- Aide aux chercheurs pour le reporting des publications en Open Access aux agences de financements (FNS et H2020).
- Aide pour assurer les droits du chercheur et leurs copyrights grâce à l'utilisation des licences Creative Commons.
- Explication sur le cycle de vie des données
- Description des standards pour les données de recherche et leurs métadonnées.

Les aides au chercheur se réalisent également en présentiel sous la forme de formations du type workshop en groupe restreint mis en place de manière régulière à la BiUM. Ces formations sur mesure permettent d'accompagner le chercheur à tout moment dans le cycle de vie de ses recherches selon ses besoins, de la création du Data Management Plan à la mise en ligne en accès ouvert d'articles ou de sets de données.

4.2.2 Recherche existante et données créées

A la FBM, les données en science fondamentales sont, dans notre contexte, créées par les sections des sciences fondamentales (SSF) de l'UNIL alors que les données précliniques et cliniques sont créées par les sections des sciences cliniques (SSC) du CHUV.

Les données Omics produites dans le cadre de la recherche concernent à la fois le fondamental, le préclinique et le clinique et étudie trois champs d'étude scientifique que sont les *genomics*[40], les *metabolomics*[41] et enfin les *proteomics*[42], c'est-à-dire respectivement l'étude des gènes, des métabolites[43] et des protéines dans les cellules. Chacune de ces diverses plateformes institutionnelles génèrent un grand nombre de données sous divers formats, stockés par les serveurs universitaires et accessibles pour les chercheurs de la FBM.

Ces *big data* nécessitent un outil puissant afin d'être analysées et comparées. Pour cela, la FBM utilise le projet *Vital-IT*[44]*.* Celui-ci est un centre de compétence bio-informatique qui est le fruit de la collaboration entre le Swiss Institute of Bioinformatics (SIB), les universités de Lausanne, Genève et Bâle, le Ludwig Institute for Cancer Research, l'EPFL, Hewlett Packard et enfin Intel. Les données brutes amenées par les

[40] https://www.unil.ch/cig/home/menuinst/research/core-facilities/dr-harshman---gtf.html
41 https://www.unil.ch/cig/home/menuinst/research/core-facilities/mef.html
42 https://www.unil.ch/cig/home/menuinst/research/core-facilities/dr-quadroni---paf.html
[43] Molécules produites ou transformées au sein des cellules.
[44] https://www.unil.ch/cig/home/menuinst/research/core-facilities/vital-it.html

instituts universitaires et médicaux seront alors analysées grâce aux compétences des bio-informaticiens du SIB.

Les données provenant de l'imagerie microscopique sont, quant à elles, générées par trois plateformes *Cellular Imaging Facility* (CIF)[45] distribuées sur trois sites différents afin d'assurer des services pour les divers départements du CHUV et de l'UNIL[46]. Celles-ci assurent aux chercheurs le stockage, les back-ups et un accès simplifié à leurs données de microscopie mais, par contre, ne donnent pas d'indication spécifique aux chercheurs quant à la manière de gérer l'organisation et le versioning de leurs données.

Les données en imagerie clinique sont, elles, produites par divers services spécialisés du CHUV, notamment par le *Service de médecine nucléaire et imagerie moléculaire*[47], par le *Département des neurosciences cliniques*[48] ainsi que par le *Service de radiodiagnostic et radiologie interventionnelle*[49]. Le premier produit les PET-Scans ainsi que les IRM. Le second produit l'imagerie liée au cerveau, alors que le dernier produit les radiographies.

Quant aux données médicales, c'est-à-dire les données liées aux tests courants, elles sont générées par les différents service cliniques du CHUV et ensuite gérées avec la mise en place d'un *data warehouse*[50] par la *biobanque institutionnelle de Lausanne*[51]. Cette biobanque étant la plus grande de Suisse[52], « *offre un réservoir exceptionnel d'échantillons biologiques* » (Jacquemont 2016, p. 8). La biobanque offre divers services dont la protection et sécurité des données, la mise à disposition et le partage contrôlé des données récoltées lors de la pratique clinique à toute fin éventuelle de recherche à la FBM. Les chercheurs n'ont accès qu'aux données qui ont été dépersonnalisées, même si une réversibilité est possible par des personnes dûment autorisées à la biobanque garantissant ainsi le secret médical du patient (Jacquemont 2016).

[45] https://www.unil.ch/fbm/fr/home/menuinst/recherche/plates-formes/cellular-imaging-facility-ci.html
[46] http://cifweb.unil.ch/
[47] http://www.chuv.ch/medecine-nucleaire
[48] http://www.chuv.ch/neurosciences
[49] http://www.chuv.ch/rad
[50] Entrepôt de données en français
[51] http://www.chuv.ch/biobanque/bil_home.htm
[52] https://uniris.unil.ch/files/researchdata/document/05_Biobanque_BIL_Nathalie-JACQUEMONT.pdf

5. Formations générales et biomédicales sur les données de la recherche – exemples

Afin de réaliser des formations adaptées aux chercheurs et au domaine de la santé plus particulièrement, ce chapitre aura comme rôle de lister mais surtout de décortiquer des formations existantes (générales ou spécialisées dans la santé). Le but étant de préciser le fonctionnement de celles-ci afin de relever les points positifs dont nous pourrons nous inspirer mais aussi les points négatifs, qui auront reçu un accueil plus mitigé, qu'il nous sera utile d'éviter.

Afin d'atteindre cet objectif, il sera nécessaire de décrire précisément le contenu du cours, la formation théorique mais aussi et surtout les exercices pratiques qui ont été utilisés.

5.1 MANTRA – Research Data Management Training

Le MANTRA – Research Data Management Training[53] est un portail anglophone proposant un cours en ligne complet sur la gestion des données de recherche. Ce portail a été réalisé par le JISC Managing Research Data Programme[54] ainsi que par the University of Edinburgh[55] qui le gère dorénavant. Le portail est orienté pour les géosciences, les sciences sociales et la psychologie clinique mais la formation proposée se veut généraliste et peut être suivie par quiconque désirant s'auto-former sur la gestion des données de la recherche.

5.1.1 La formation du MANTRA

Créé en 2011, MANTRA offre ses cours en Open Access, permettant à tout un chacun de se former aux diverses problématiques de la gestion des données. Pour cela, le cours est divisé en neuf modules distincts ayant chacun ses objectifs :

- **Research Data Explained :** Expliquer la nature des données de recherche, leurs spécificités ainsi que les problématiques qui en découlent.
- **Data Management Plans :** Expliquer le principe d'un Data Management Plan, comment le créer selon les exigences des financeurs et enfin comment le gérer tout le long du cycle de la recherche (notamment au niveau des mises à jour).
- **Organising Data :** Traite du nommage des données ainsi que du versioning de celles-ci.

[53] http://datalib.edina.ac.uk/mantra/
[54] https://www.jisc.ac.uk/
[55] http://www.ed.ac.uk/information-services/about/organisation/edl/data-library-projects/mantra/team

- **File formats and transformation :** Expliquer la problématique des formats ouverts et propriétaires dans les données ainsi que comment les transformer dans un format plus adéquat.
- **Documentation, metadata, citation :** Expliquer pourquoi la documentation de ses données est importante dans l'objectif de la réutilisation de celles-ci, quelles métadonnées doivent impérativement être présentes et quels sont les standards en la matière.
- **Storage and security :** Enseigner les méthodes de sauvegarde des données, comment les stocker ainsi que la gestion des mots de passe.
- **Data protection, right and access :** Expliquer l'importance de la confidentialité des données, de leur anonymisation ainsi que les protections juridiques dont jouissent celles-ci.
- **Sharing, preservation and licensing :** Déterminer sous quelle licence Creative Commons les données seront publiées, expliquer l'importance du partage de celles-ci et du choix du dépôt en vue de leur partage.
- **Data handling tutorials :** Ce dernier module concerne des exercices pratiques sur la gestion de ses données à travers différents logiciels.

Chacun de ses modules a été élaboré dans l'optique de représenter une heure de cours environ. Bien que généraliste, ce cours en ligne, de par ses différents modules, s'avère être relativement complet et didactique. En effet, celui-ci couvre l'entièreté du cycle de vie des données, mettant l'accent sur les points importants de la *Digital Curation*. Cette approche est en accord avec le public visé, les chercheurs, à qui ce cours doit rapidement apporter des solutions d'ordre pratique.

La forme des cours est diverse et complète. Lorsque l'on choisit un module, celui-ci commence par un diaporama dont le style est proche de *Slideshare.* L'on y décrit au départ les objectifs du module en question et ce que la personne devra être en mesure de comprendre d'ici la fin de celui-ci.

Les explications se font sous la forme de paragraphes textuels accompagnés d'images pour illustrer le propos, notamment des captures d'écran. Les informations peuvent être complétées et illustrées par des vidéos hébergées sur *Youtube*. L'interactivité avec l'utilisateur se concrétise sous la forme de petits questionnaires à choix multiples ou de textes lacunaires. Chaque exercice comprend une correction rapide avec une note explicative. Néanmoins, ces questionnaires ne sont pas obligatoires et il est possible de ne pas en tenir compte.

Les modules se terminent en décrivant le module suivant ainsi qu'en proposant une liste de sources si d'aventure l'utilisateur désirait approfondir le sujet.

Le dernier module, nommé *Data handling tutorials,* est pour sa part relativement différent des autres. En effet, il ne représente pas un cours textuel mais un module

pratique, basé sur des exercices sur différents logiciels tels que *SPSS*[56] et *R*[57] pour les analyses statistiques, *ArcGIS*[58] pour les données géographiques ou encore *NVivo*[59] pour l'analyse des données qualitatives. Il est donc composé de dossiers au format zip, téléchargeables, qui comportent des données à traiter dans le cadre des exercices. Les formats des données dépendent du logiciel utilisé et de la nature des exercices. Il peut s'agir de simples fichiers Word ou Excel ou des formats plus spécifiques comme le XML entre autres. Des tutoriels au format pdf accompagnent les données afin de proposer une marche à suivre dans le but de compléter les exercices.

5.1.2 Kit prêt à l'emploi pour les bibliothèques

Outre ses modules de cours dont nous pourrions nous inspirer, du moins en partie, l'intérêt du portail MANTRA est la proposition d'un kit *clé en main* de formations pour les bibliothèques nommé **Do-It-Yourself Research Data Management Training Kit for Librarians[60].** Celui-ci a été créé en 2013 suite à la tenue de formations pour les bibliothécaires académiques dans la gestion des données de recherche. Ce matériel, destiné aux professionnels ID, est diffusé sous la licence CC BY et peut donc être utilisé et modifié par quiconque tant que l'on cite les auteurs de ce kit. Selon les auteurs, le DIY Training Kit a été créé afin de pouvoir réaliser des formations similaires à ce qu'ils ont eux-mêmes donné.

Le kit comprend six modules distincts dont cinq cours à proprement parlé. À l'exception du premier, la composition des modules est sensiblement homogène. On y retrouve :

- Une présentation Powerpoint sur la thématique abordée.
- Un podcast audio suivant la présentation afin de donner un exemple de celle-ci.
- Un document pdf proposant des questions à poser oralement durant le cours afin de sonder l'avancée des connaissances des participants.
- Divers exercices pratiques.

Les six modules proposés sont :

- **Pré-training :** Ce matériel introductif explique le contenu des cours mais aussi donne certains conseils aux professionnels sur la tenue des formations. On y explique notamment les bonnes pratiques sur le nombre de participants idéal et sur les exigences concernant des connaissances préalables notamment.
- **Session 1 : Data management planning :** Ce cours concerne les modules 1 et 2 du cour MANTRA reprenant l'importance de la gestion des données de

[56] http://www.ibm.com/analytics/us/en/technology/spss/
[57] https://www.r-project.org/
[58] https://www.arcgis.com/home/index.html
[59] http://www.qsrinternational.com/nvivo-french
[60] http://datalib.edina.ac.uk/mantra/libtraining.html

recherche, leur cycle de vie ainsi que le Data Management Plan. Le déroulement du cours se fait comme suit :

- I. What is Research Data Management (RDM)?
- II. Why does RDM Matter?
- III. What do the funders expect?
- IV. What does Edinburgh University expect?
- V. Creating Data Management Plans
- VI. Questions ?

L'exercice proposé demande aux participants de remplir un schéma de cycle de vie des données à partir des recherches qu'ils vont réaliser. Le but étant ici qu'ils puissent mettre en pratique la théorie vue dans la présentation.

- **Session 2 : Organising & documenting data :** Ce cours concerne les modules 3 et 5 du cours MANTRA sur le nommage, le versioning et la création de métadonnées. La session 2 comprend trois exercices. Le premier consiste en un renommage de plusieurs fichiers proposés, le deuxième exercice un questionnaire à choix multiples et, enfin, le troisième exercice est une mise en situation où selon un contexte précis, on demande à la personne formée de définir les éléments importants qu'elle désirerait retrouver dans les métadonnées qui en découleraient.
- **Session 3 : Data storage & security :** Ce cours correspond au module 6 du cours MANTRA. Le déroulement du cours se fait comme suit :
  - I. Where to store data
  - II. Data backup
  - III. Data security
  - IV. Usernames and Passwords

Deux exercices pratiques sont proposés. Le premier demande aux participants de définir cinq brèches de sécurité potentielles dans leur gestion des données et le second est un questionnaire à choix multiple sur la sécurité du stockage des données.

- **Session 4 : Ethics & copyright :** Ce cours concerne le module 7 du cours MANTRA. Il s'intéresse à l'éthique quant au stockage de données sensibles, l'anonymisation des données et l'utilisation des données d'autrui. La notion de copyright y est également évoquée.

Ce cours propose deux exercices pratiques. Le premier évoque plusieurs scénarios liés au copyright qui pourraient ressembler au vécu des participants. Chaque scénario soulève des questions auxquelles les participants devront répondre. Le second exercice reprend trois formulaires de consentement que les chercheurs devront fournir aux patients. Plusieurs questions sont posées aux participants afin de connaître leur avis sur chacun des formulaires, lequel est le plus complet, quels sont les éléments manquants, etc.

- **Session 5 : Data sharing :** Ce cours reprend le module 8 du cours MANTRA, s'intéressant à la problématique du partage des données de recherche, le principe de l'Open Data, etc.

Deux exercices pratiques y sont proposés. L'originalité de ces exercices est qu'ils se réalisent sous forme *négative*. Ainsi, le premier demande aux participants de décrire 5 obstacles au partage des données, des raisons pour

lesquelles les chercheurs seraient réticents au partage. Le second exercice reprend les raisons à ne pas partager les données et demande aux participants de répondre et d'argumenter contre chacune de ces raisons. À noter que le second exercice est proposé avec un corrigé.

5.2 University of East London

L'University of East London[61] et sa bibliothèque sont contributrices au projet du *Data Curation Center*. L'université a suivi une politique proactive dans la gestion des données de recherche en définissant des directives[62] (University of East London [sans date]) à appliquer ainsi que la création de modules d'entraînement ciblés pour les chercheurs mais aussi pour les professionnels de l'information. Ces modules, au nombre de six, sont trouvables sur Zenodo[63]. Ils comportent un cours au format Powerpoint, des annotations et bonnes pratiques pour les formateurs ainsi que des exercices à proposer aux participants. Ces modules s'intéressent aux thématiques suivantes :

- **Introducing research data management** (Jones, Murtagh & Grace, 2014)

L'exercice principal de ce module comprend une série d'exemples de données, de tout type (image, statistique, graphique) et demande aux participants à quel article publié (compris dans une liste avec titre, auteurs et sujet) ces données correspondent-elles. Le but étant de montrer la pertinence de chaque type de données selon ce que l'on désire transmettre et la grande diversité des formats utilisés.

- **Guidance and support for researchers** (Grace 2014a)

Ce second module a été créé uniquement pour les professionnels de l'information et les formateurs à la gestion des données de recherche. Il y explique les facteurs de succès à la *Monash University*[64] pour le service de gestion des données, de la formation au soutien quotidien, en passant par l'offre de services divers. Ainsi, nous pouvons résumer les points suivants :

- Une politique proactive de l'université
- Une responsabilité partagée entre les services de chercheurs et le service de documentation
- Prendre en compte tous les formats qui peuvent être créés afin de ne laisser personne démuni face à ses données de recherche produites
- Adopter rapidement les nouvelles pratiques et être constamment à l'affut des nouvelles offres

[61] http://www.dcc.ac.uk/tailored-support/institutional-engagements/east-london
[62] https://www.uel.ac.uk/wwwmedia/services/library/lls/resources/rspresearchtools/Research-Data-Management-policy-for-UEL-FINAL.pdf
[63] https://zenodo.org/search?ln=en&p=uel&action_search=
[64] http://www.monash.edu/library/researchdata

- Amener diverses propositions en termes d'outils et de structures
- Importance de s'adapter aux thématiques et besoins de l'université pour laquelle nous travaillons

- **Data management planning** (Jones 2014)

Ce module consiste en une étude de cas de plusieurs services en lien avec le soutien à la création d'un DMP. On y explique les bonnes pratiques des autres universités, notamment à l'université de Leicester où le DMP est intégré au système informatique (Grants Application System) et où les chercheurs doivent remplir des cases telles que dans un CMS ce qui permet d'avoir des documents uniformisés.

- **What data to keep and why** (Grace 2014b)

L'exercice pratique de ce module consiste en la création d'une checklist afin d'aider les participants à définir quelles sont leurs données qui méritent d'être conservées ou non. Cette checklist, d'une trentaine de questions pouvant toutes être répondues par Oui/Non, s'articulent autour des points suivants :

  - *Legal/statutory considerations*
  - *Policy*
  - *Scientific or historic value*
  - *Origin*
  - *Condition*
  - *Storage and preservation*
  - *Access/use*
  - *Formats/technical limitations*

- **Cataloguing Data** (Duke 2014a)

L'exercice pratique consiste en l'utilisation de dépôts de données où les participants notent les métadonnées qui leur sont utiles afin de comprendre le set de données qu'ils sont en train d'étudier → ils se demanderont donc « de quelle information ai-je besoin afin de comprendre le contenu de ce set ? »

- **Sharing Data** (Duke 2014b)

Les participants remplissent un tableau à partir de la question suivante : « quelles sont les contraintes et les limites au partage des données de recherche ? ». Le but étant bien entendu de créer une discussion autour de ces inquiétudes afin de les atténuer une par une.

5.3 DATUM for Health

Le projet *Research Data Management Training materials (RDMTrain)*[65] est un projet britannique créé par le *JISC Managing Research Data programme*[66] en 2010. Ce projet de gestion des données de recherche regroupe cinq programmes de formations aux chercheurs, chacun dans un domaine spécifique.

Le *DATUM for Health*[67] est le programme créé pour le domaine de la santé, fruit de la collaboration entre *The Northumbria University*[68], le *Digital Curation Centre* et la *Digital Preservation Coalition*[69] entre autres. Le but de ce projet collaboratif était de :

> *« Promote the research data management (RDM) skills of postgraduate research (i.e. doctoral) students in the health studies discipline via a specially developed training programme. It focused on the management of qualitative, unstructured data. »* *(McLeod 2011, p. 4)*

Vingt-cinq chercheurs ont pris part à ce programme durant l'année 2010, vingt-deux étudiants post-gradués, autrement dit doctorants, et trois chercheurs diplômés.

Le projet *DATUM* comprend plusieurs *outputs* qui ont chacun leur utilité dans la gestion des données de recherche. Les plus importants sont la création d'un moteur de recherche Google customisé pour la gestion des données, la création d'une formation complète pouvant aider les chercheurs, l'évaluation de l'efficacité et de l'utilité de cette formation ainsi que l'établissement d'une politique volontariste dans la gestion des données de recherche dans toute l'université afin que la formation soit intégrée automatiquement à tous les projets de recherche (Wells 2013).

5.3.1 La formation du DATUM

Bien que la formation s'adresse avant tout aux doctorants et non pas aux chercheurs diplômés proprement dit, il nous paraît intéressant de l'étudier car elle semble avoir été ambitieuse et couronnée d'un certain succès.

Le matériel produit et fourni dans le cadre de ces cours a été publié sur la plateforme *Jorum*[70] sous la licence CC BY-NC-SA.

La formation du DATUM était, au départ, composée de quatre sessions distinctes[71]. Après avoir sondé la volée de la première mouture, il s'est avéré que la quatrième

[65] http://www.webarchive.org.uk/wayback/archive/20140614021623/http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain.aspx
[66] https://www.jisc.ac.uk/rd/projects/managing-research-data
[67] http://www.webarchive.org.uk/wayback/archive/20140614071433/http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain/datum.aspx
[68] https://www.northumbria.ac.uk/research/research-data-management/
[69] http://www.dpconline.org/training
[70] http://find.jorum.ac.uk/resources/18276

session était relativement redondante, donc inutile. Les organisateurs ont donc décidé de réduire la formation aux trois premières sessions et une introduction à la préservation de l'information digitale (thématique de la quatrième) a été incorporée à la deuxième session :

- **Session 1**. Introduction to Research Data Management (durée 2h30)
    - What is research data?
    - Where is your research data?
    - Why manage research data?
    - How to manage research data?
    - The research data lifecycle
    - Creating a DMP
- **Session 2**. Digital Curation (durée 2h30)
    - What is data curation?
    - Why curate?
    - DCC Data Curation Lifecycle Model
- **Session 3**. Problems and Practical Strategies and Solutions (durée 2h30)
    - What problems are there?
    - Conflicts
    - Data security and storage
    - Metadata
- (**Session 4**. Data4Life – Digital Preservation for Health) (durée journée entière)

Chaque session de cours est composée d'une présentation Powerpoint, d'un document de notes à l'attention du formateur au format pdf, comprenant un résumé du discours donné aux participants, une séquence pédagogique succincte, des remarques importantes à souligner pendant le cours, la liste des questions que le formateur posera ainsi que la proposition d'un débat qu'il serait opportun d'installer avec les participants. De plus, divers exercices, faits durant la formation, sont proposés afin d'illustrer mais aussi dynamiser la formation. Enfin, les formateurs ont créé un document pdf à l'attention des participants résumant le but et les points importants de la formation, permettant de les guider durant les trois sessions (Northumbria University 2011).

En plus de la présentation orale en elle-même, nous pouvons résumer les exercices et questions de chaque session ainsi :

Session 1 : (Northumbria University 2011a)

[71] https://code.soundsoftware.ac.uk/projects/sodamat/wiki/DATUM_for_Health

- Exercice pratique : les participants doivent répondre à ces questions après la présentation du premier chapitre : (2 min)
    - *What data are you using (i.e. that you've got from elsewhere)?*
    - *What data are you creating?*
  - Question ouverte posée à tout le groupe concernant la localisation des données (médias, systèmes de stockage virtuels ou physiques) et inscrire les réponses proposées sur un tableau.
  - Question ouverte à propos du danger de la perte de données. Qui en a déjà souffert ? Rebondir sur des exemples.
  - Exercice pratique : les participants vont remplir leur Data Management Plan à partir d'un exemple mis à leur disposition. Il est nécessaire de prévoir au minimum 30 minutes pour cette activité. Il est utile de grouper les participants par 3 ou 4 afin qu'ils puissent s'entraider, bien qu'il soit important de passer dans les rangs, afin de voir si tout se passe correctement et si des incertitudes demeurent.
  - Il est aussi nécessaire de réserver du temps après l'exercice pour un feedback, corriger leurs réponses et les points qui ont été incompris ainsi que pour répondre aux éventuelles questions.
    - *What do you have for…? (a few Qs to select)*
    - *Any difficult areas?*
    - *Would you say you had already addressed all of the points in this DMP for your research?*
      - *If so why? (experience, someone suggested you did, you had to i.e. it was a requirement – who made it a requirement?)*
      - *If not then which questions / aspect haven't you considered and / or need to consider further?*

Session 2 : (Northumbria University 2011b)

- Exercice pratique (2 min env.) : à partir du modèle du cycle de vie des données présenté brièvement précédemment, demander aux participants qu'elles sont les activités qu'ils devront accomplir à chaque étape du cycle. Cette question sert d'introduction à la suite du cours.
  - Questions ouvertes à tout le groupe après avoir décrit chaque étape du cycle de vie, reprendre le premier exercice comme une synthèse. Ont-ils pensé à toutes les tâches évoquées ? par lesquelles ont-ils été surpris ? lesquelles ont-ils ajouté qui ne sont pas dans le cours ?
  - Exercice pratique à propos de la préservation des données (env. 15 min.) : Grouper les participants par 3 ou 4. Leur demander de réfléchir et discuter sur les données brutes qu'ils vont créer ainsi que les données associées (e-mails, protocoles, etc.). Une fois cela fait, leur demander quelles données :
    - Peuvent être gardées uniquement pour une courte période ?
    - Peuvent être gardées pour une longue période ?
    - Doivent absolument être conservées le plus longtemps possible ?
  - Réserver une plage de 20 minutes environ pour discuter de leurs réponses, les corriger si nécessaire et répondre aux éventuelles questions.

- Reprendre le Data Management Plan que les participants avaient rempli lors de la première session. Le but étant qu'ils puissent le compléter à la lumière des informations qu'ils ont acquises durant cette session. Les participants devront répondre aux questions ci-dessous : (env. 5 min.)
    - *What do I need to add?*
    - *What do I need to change?*
    - *What further questions do I have?*
- Réserver 10 minutes afin de corriger leurs réponses.

Session 3 : (Northumbria University 2011c)

- Exercice pratique d'anonymisation : les participants se voient recevoir des documents biomédicaux qu'ils doivent anonymiser. (Env. 15 min.)
- Correction : discussion ouverte sur l'exercice, quels éléments ont-ils dû anonymiser ? Donner aux participants un corrigé. (Env. 20 min.)
- Exercice pratique sur la gestion des fichiers : basé sur des scénarios[72] de recherche, les participants devront répondre aux trois questions ci-dessous :
    - *1. What research data and documentation would you expect to exist, including raw data and analysed data? Create a list.*
    - *2. Taking this list of data can you develop a list of potential file names?*
    - *3. Can you structure the file names into a file plan with folder names?*

  Compter environ 40 minutes pour la réalisation de l'exercice et la correction de celui-ci.

5.3.2 Évaluation

Selon le rapport de Julie McLeod[73], les auteurs du programme DATUM ont réalisé une évaluation de leur cours auprès des participants. Ces derniers ont trouvé le style et le contenu appropriés et ont mis en évidence la grande utilité du cours pour la suite de leurs recherches. Ils auraient néanmoins souhaité recevoir cette formation dès le début de leur cursus doctoral afin de gagner du temps.

La majorité des participants aurait préféré un discours et des exercices spécialisés dans leur discipline afin de rendre le cours plus compréhensible et utile. Cette opinion doit toute de même être contrebalancée par l'avis des formateurs eux-mêmes. Selon eux, une formation complètement spécifique nécessite un trop gros investissement en temps et ne serait pas forcément plus utile. Selon eux, 80% de la formation pourrait être traitée de manière générique sans pour autant perdre en pertinence. Ainsi,

> *« A pragmatic and sustainable way of delivering the disciplinary focus and contextualisation is to 'tailor' generic materials through (a) discussion about research philosophy/epistemology; (b) covering specific requirements of qualitative or quantitative data; and (c) incorporating discipline-specific examples, case studies, exercises and references. »* *(McLeod 2011, p. 10)*

[72] https://www.northumbria.ac.uk/static/5007/ceispdf/scenarios.pdf
[73] http://nrl.northumbria.ac.uk/3864/2/report.pdf

Les participants ont souligné l'importance de la réalisation du Data Management Plan en cours, avec l'aide du formateur. De même, les formateurs notent qu'il est utile d'adapter le DMP à son public, afin qu'il réponde au mieux à leurs besoins.

5.4 New England Collaborative Data Management Curriculum

Le projet du New England Collaborative Data Management Curriculum (NECDMC)[74] est dirigé par la *Lamar Soutter Library*[75] de la *University of Massachusetts Medical School* en partenariat avec la *George C. Gordon Library*[76] du *Worcester Polytechnic Institute.*

Tout comme le MANTRA et le DATUM, ce projet consiste en la création d'une formation pour chercheurs sur la gestion des données de recherche qui ont été mis sous licence CC BY-NC-SA disponibles donc au téléchargement et à la réutilisation[77] Cette formation s'adresse aux étudiants, doctorants mais aussi chercheurs dans le domaine de la santé notamment. Elle a été élaborée en sept modules reprenant une progression classique dans la formation aux données de recherche.

5.4.1 La formation du NECDMC

Chaque module de cours proposé par le projet comprend un texte explicatif à l'attention du formateur, une présentation Powerpoint ainsi que divers exercices et cas à traiter. Tout au long de la formation, les créateurs ont intégré en celle-ci des *Demo Research Teaching Cases*[78], utilisés en tant qu'exercices. Il s'agit de cas réels de recherche, repris et adaptés avec l'aide des départements concernés qui comprennent donc un processus, des données et des métadonnées qui vont être utilisées et sur lesquelles il va être possible de questionner les participants et leur demander leur avis. Parmi ces cas étudiés, nous avons notamment :

- Clinical health study:
    - Outcomes from Orthopedic Implant Surgery
    - Studying Vitamin D as an Augmentation of Treatment for Bipolar Depression
    - Health study in lab using derived data from multiple projects: Combining Data from 10 Years of Research for Retrospective Studies on the Effects of Exercise and Diet on the Risk of Diabetes
- Biomedical engineering lab research:
    - Regeneration of Functional Heart Tissue in Rats

[74] http://library.umassmed.edu/necdmc/index
[75] http://library.umassmed.edu/
[76] https://www.wpi.edu/academics/library.html
[77] http://library.umassmed.edu/necdmc/modules
[78] http://library.umassmed.edu/necdmc/research_cases

        - Case Excerpt: Regeneration of Functional Heart Tissue in Rats Designing a Mobile and Compact Optical Mammography Instrument
    - Designing a Mobile and Compact Optical Mammography Instrument

Module 1 : Overview of Research Data Management (Creamer et al. 2015)

Ce module représente l'introduction classique à la gestion des données de recherche, leur utilité, les bonnes pratiques, etc.

L'exercice pratique consiste en la création classique d'un DMP à partir de 7 questions basiques que les chercheurs doivent se poser quand ils désirent le réaliser :

- *Types of data*
  - *Contextual Details (Metadata) Needed to Make Data Meaningful to others*
  - *Storage, Backup and Security*
  - *Provisions for Protection/Privacy*
  - *Policies for re-use*
  - *Policies for access and sharing*
  - *Plan for archiving and preservation of access*

Module 2: Types, Formats, and Stages of Data (Ferguson 2015)

Informant sur les types de données, leur format et de leur standardisation, le deuxième module propose de s'intéresser à l'une des *Demo Cases.* Après s'être informé sur le cas étudié à l'aide d'un document récapitulatif, les participants doivent répondre à une vingtaine de questions sur celui-ci. Elles concernent les informations à conserver dans le carnet de laboratoire, les données à conserver et comment les nommer.

Module 3: Contextual Details Needed to Make Data Meaningful to Others (Coburn, Furfey & Walton 2015)

Le troisième module forme les participants à l'importance des métadonnées et liste les standards à adopter dans leur contexte. Dans le cadre de l'exercice pratique, la formation reprend les *Demo Cases* et demande aux participants quelles métadonnées seront, d'après eux, susceptibles d'être conservées, lesquelles intéresseront leurs collègues et seront nécessaires dans la description de leurs données de recherche.

Module 4: Data Storage, Backup, and Security (Canavan, McGinty & Reznik-Zellent 2015)

Le but de ce module est d'expliquer l'importance de la sauvegarde des données ainsi que leur sécurité en définissant les droits d'accès. Reprenant l'un des *Demo Cases*, le participant devra répondre aux questions concernant les bonnes pratiques à faire pour que les données soient conservées à long terme et de manière sécurisée. En supplément, le module propose un QCM concernant cette thématique.

Module 5: Legal and Ethical Considerations for Research Data (Kafel, Palmer & Piorun 2015)

Le module 5 s'étend sur la notion de copyright lié aux données de recherche ainsi que la notion d'anonymisation qui y est aussi évoquée. À partir des cas de démonstration, l'exercice pratique consiste à demander quelles sont les données qu'ils devraient anonymiser, quels copyrights entrent en jeu, etc. Un second exercice consiste à fournir une série de données à anonymiser à partir de documents théoriques ; sous cette forme « changer *nom original de la donnée* en *nouveau nom* »

Module 6: Data Sharing & Reuse Policies (Sheridan et al. 2015)

Le module 6 s'intéresse à l'importance du partage des données, des notions d'Open Acces et d'Open Data, en termes de bonnes pratiques mais aussi d'éthique. La problématique de la citation y est aussi évoquée car elle est un corollaire du partage. L'exercice pratique consiste à citer des données correctement à partir d'informations que l'on a sur celles-ci.

Module 7: Repositories, Archiving & Preservation (Lowe, White & Novak Gustainis 2015)

Le module 7 conseille les participants sur les différents types de dépôts de données existants. Dans le cadre des cas de démonstration, l'exercice pratique demande aux participants de choisir quelles données ils comptent conserver à court terme ainsi qu'à long terme.

5.4.2 L'exemple canadien

L'Université de Montréal (UdeM) a, de son côté, repris et adapté le premier module de la formation NECDMC proposée en Open Access. À la suite d'un sondage réalisé en décembre 2013 auprès des chercheurs en sciences de la santé de l'UdeM, il s'est avéré que la gestion des données de recherche était l'un des éléments qui leur importait le plus dans l'offre possiblement offerte par la bibliothèque universitaire.

Le premier module du NECDMC a été traduit en français[79] et a été adapté ainsi qu'enrichi par des problématiques locales.

La formation fut ainsi constituée des éléments supplémentaires suivants (Clairoux 2016) :

- Le contexte canadien de la gestion des données de recherche

[79] http://www.bib.umontreal.ca/sa/gestion-donnees-recherche.pdf

- Informations spécifiques à l'institution, services disponibles pour les chercheurs, les politiques de l'institution en termes d'Open Access et de copyright
- Une vidéo sur l'importance du *Data Sharing* réalisée par la NYU Health Sciences Library[80]
- Simplification et traduction des présentations Powerpoint du NECDMC

À noter qu'aucun workshop n'a été réalisé avant ou après la formation. Ainsi, selon les organisateurs, la formation devait se suffire à elle-même.

5.4.2.1 Évaluation

Selon la dernière étude sur leur formation réalisée à l'Université de Montréal en 2015 (Clairoux 2015), les participants ont été satisfaits de la formation proposée et ont apprécié les exercices pratiques liés à leur branche de recherche. Néanmoins, les participants ont regretté le temps serré imparti à cette formation. En effet, ils auraient souhaité que celle-ci soit plus longue, permettant d'approfondir la thématique. Ainsi, les formateurs évoquent la possibilité pour 2016 de traduire les modules suivants afin de compléter la formation selon les besoins.

5.5 Checklists

Diverses bibliothèques et autres centres de documentation ont créé des checklists à l'attention des chercheurs afin de les guider dans la réalisation de leur DMP. Ces checklists se réalisent généralement sous la forme d'une série de questions auxquelles le chercheur devra répondre afin d'avoir la garantie d'une gestion adéquate de ses données de recherche. Dans le contexte de notre travail, une checklist serait intéressante dans le cadre de la première formation sur le DMP. Elle pourrait être utilisée soit en amont de cette formation, où l'on demanderait aux chercheurs de la remplir avant même le rendez-vous en présentiel, soit comme exercice pratique durant le cours. En effet, il pourrait être intéressant, après une introduction, de demander aux participants de remplir un questionnaire inspiré d'une checklist, permettant ainsi d'illustrer les propos théoriques.

5.5.1 Le Data Management Plan en 20 questions du OXFORD DMP online Project

Afin de soutenir les chercheurs dans la réalisation d'un Data Management Plan, l'Université d'Oxford a réalisé une plateforme online d'aide en collaboration avec le *Digital Curation Centre.* Dans cette optique, David Shotton, responsable du projet, a

[80] https://www.youtube.com/watch?v=66oNv_DJuPc&feature=youtu.be

réalisé un document comprenant 20 questions[81] auxquelles les chercheurs se devront de répondre s'ils désirent réaliser un DMP de qualité. Ces questions, concrètes, représentent une aide non négligeable et pourraient tout à fait être reprises en tant qu'exercice pratique après adaptation dans une formation sur le Data Management Plan. Voici ces 20 questions :

- **The nature of your data**
    - 1 What is the subject discipline (domain, field) to which your research data relates?
    - 2 What is the exact nature (range, scope) of your research data?
    - 3 In what format(s), will you store your data in the short term after acquisition?
    - 4 Who owns the data arising from your research, and the intellectual property rights relating to them?
- **Data descriptions (metadata, "data about data")**
    - 5 How will your research datasets be described?
    - 6 How will these descriptive metadata be created or captured?
- **Data sharing**
    - 7 With whom will you share your research data in the short term, before publication of any papers arising from their interpretation?
- **Data storage and backup**
    - 8 Where will you store your data in the short term, after acquisition?
    - 9 Who is responsible for the immediate day-to-day management, storage and backup of the data arising from your research?
    - 10 How frequently will your research data be backed up for short-term data security?
- **Data archiving**
    - 11 Where will your research data be archived for long-term preservation?
    - 12 When will your research data be moved to a secure archive for long-term preservation and publication?
    - 13 Who will decide which of your research data are worth preserving?
    - 4 How (i.e. by what physical or electronic method) will you transfer your research datasets to their long-term archive, under the curatorial care of a separate third-party, e.g. a data repository?
- **Data publication**
    - 15 For how long will you embargo your research data before it is published for others to see and use?
    - 16 Why is public access to your research data to be restricted (if indeed it is)?

[81] http://blogs.it.ox.ac.uk/acit-rs-team/files/2014/02/Twenty_Questions_for_Research_Data_Management.pdf

    - 17 Under what data-sharing license will you publish your research data?
    - 18 What persistent identifiers will be used to permit correct citation of your datasets?
    - 19 What metadata will be published with the data to make them interpretable and reusable?
- **Future data management**
    - 20 Who will be responsible for your data, once you have left your present research group?

(Shotton 2012)

Ces questions ne sont pas réellement *explicatives* dans les sens où le chercheur se doit déjà de comprendre la problématique de la RDM avant d'y répondre. Néanmoins, elles peuvent représenter un support utile comme exercice pratique.

5.5.2 ETH - EPFL

Les bibliothèques des deux écoles polytechniques ont réalisé en mars 2016 un document commun d'aide aux chercheurs comprenant une synthèse sur la problématique de la gestion des données de recherche, expliquant le cycle de vie, etc. mais aussi une checklist de 66 questions sur les points suivants (ETH Bibliothek, EPFL Libray 2016) :

- Planning
- Data collection and creation
- Appraisal and selection
- Documentation and metadata
- File Formats
- Storage
- Ethics
- Copyright and intellectual property
- Sharing
- Long term management

5.6 Formations ludiques et questionnaires

L'autoformation de la part des chercheurs, pour une partie de la thématique tout du moins, représente une éventualité qu'il est nécessaire de prendre en compte. C'est d'ailleurs dans cette optique qu'a été développé le portail MANTRA. En effet, dans notre cas, une autoformation permettrait de dégager du temps dans les modules d'enseignement, notamment dans l'introduction à la problématique.

Or, le problème est qu'il paraît difficile de susciter la motivation des chercheurs à s'auto-former dans le sens où leur travail est déjà suffisamment prenant. Il en

résulterait des formations en présentiel où les participants auraient des lacunes dans la matière dès le commencement, lacunes qui ne seraient pas forcément comblées en cours puisque la thématique n'y serait plus forcément abordée.

Afin de contourner ce problème, une solution existante serait de proposer une activité d'apprentissage courte, efficace et ludique afin que les futurs participants n'aient pas le sentiment de réaliser un devoir trop scolaire. Dans cette optique, il est utile d'analyser ces activités.

5.6.1 Educaplay

Educaplay est une plateforme éducative qui propose la création et le partage d'enseignements ludiques sous forme de jeux. Il est possible de créer ses propres jeux, sous forme de quizz ou autre, dans la thématique souhaitée notamment dans celle des données de recherche[82][83].

5.6.2 Questionnaires

Il est possible de réaliser un questionnaire ou un texte lacunaire que les futurs participants devront remplir afin qu'ils puissent juger eux-mêmes de leur connaissance sur la thématique. Ce document peut servir en même temps d'introduction aux cours tout en permettant une forme d'auto-formation. En complément de celui-ci, il serait tout à fait possible de les renvoyer à une formation en ligne.

5.6.3 Vidéos

Les formations réalisées sur Internet proposent régulièrement des vidéos, notamment hébergées sur *Youtube*[84], qui permettent d'illustrer et synthétiser des problématiques inhérentes au RDM. Ces vidéos peuvent aussi bien représenter des schémas dynamiques, des didacticiels couplés à des exemples ou de simples commentaires filmés d'intervenants. Néanmoins, il est plus difficile de vérifier, dans ce contexte, l'acquis auprès des futurs participants.

5.7 Conclusion

Bien que les formations étudiées ci-dessus peuvent varier dans la spécialisation de la thématique, les types d'exercices proposés ou le nombre de modules enseignés, nous pouvons tout de même dégager plusieurs tendances. La première est que l'entièreté du cycle de vie des données de recherche y est présentée afin de fournir une formation complète au chercheur. Deuxièmement, leur formation comporte un nombre divers de

[82] http://fr.educaplay.com/fr/activiteeducatives/2224638/donnees__lieu_de_publication.htm
[83] http://fr.educaplay.com/fr/activiteeducatives/2244810/donnees__formes_de_publication.htm
[84] Ex: https://www.youtube.com/watch?v=gYDb-GP1CA4

modules qui représentent une durée de formation relativement conséquente. La formation MANTRA représente environ 9 heures de cours au total, la DATUM 7h30 en ne comptant que les trois premiers modules encore enseignés et, enfin, le NECDMC 7 heures en ne comptant qu'une heure de cours par module. Or, une telle durée n'est actuellement pas acceptable dans le cadre de ce projet car, comme dit dans l'introduction, seule deux formations de deux heures chacune environ seront proposées aux chercheurs.

Dès lors, trois possibilités s'offrent à nous. La première est que nous réaliserions deux formations qui soient complètes mais condensées. La seconde option serait de réaliser des formations lacunaires, ne s'intéressant qu'à certains points développés dans les formations ci-dessus. La troisième possibilité serait de « gagner du temps » en demandant aux participants de réaliser une autoformation sur internet avant la réalisation de la formation proprement dite. Dans un premier temps, il pourrait être intéressant, pour nous, de tester la première option afin de se rendre compte si celle-ci est réalisable ou non. Si cela n'est pas le cas, il sera alors intéressant de proposer une part d'autoformation ludique afin de gagner du temps ou de prolonger la durée des modules de formation.

De plus, la part d'interactivité entre les intervenants et les participants a globalement toujours été mise en avant. En effet, celle-ci permet de déjouer les idées reçues que les chercheurs pourraient avoir envers cette thématique sous la forme d'une discussion et non pas d'une simple théorie qui leur est enseignée. De plus, l'interactivité permet de créer un espace d'échange où le professionnel ID peut tout aussi bien apprendre du chercheur que l'inverse.

De même, il semble important de trouver le bon équilibre entre cours théorique et pratique afin de contextualiser la formation pour chaque participant mais aussi afin de rendre tout simplement les modules plus dynamiques et plus agréables à suivre.

Concernant la contextualisation toujours, il semble aussi nécessaire de trouver un compromis entre un contenu généraliste et un contenu spécifique. En effet, ce dernier semble plus intéresser les participants mais il demande aussi un plus grand travail en amont.

6. Formation proposée

La formation prévue est constituée de deux modules distincts mais complémentaires. Nous sommes partis du principe que les participants suivront les deux modules, dans l'ordre si possible, même si aucun n'est obligatoire dans le programme de recherche. La complémentarité induit qu'il ne sera pas nécessaire aux modules de se répéter, chacun proposant des éléments différents.

Idéalement, les cours devraient être donnés à environ huit participants à la fois, dans une salle de cours où des ordinateurs pourront être utilisés individuellement. Ainsi, nous ne serons pas limités technologiquement dans les exercices pratiques, les participants pouvant alors aller sur internet et traiter des sets de données en direct.

La fréquence des cours serait environ d'une fois par mois, donc un module toutes les deux semaines. Bien sûr, comme cette formation ne sera pas obligatoire, l'offre devra s'adapter à une demande variable, autrement dit le nombre d'inscriptions.

Le public cible sera tous les chercheurs appartenant à la faculté de biologie et médecine (FBM), la section cliniques (SSC) du CHUV et la section des sciences fondamentales (SSF) de l'Université de Lausanne. Il en résulte ainsi une grande diversité des attentes et des besoins des chercheurs de même qu'une grande diversité dans les données de recherche créées.

Concernant les connaissances théoriques, les formateurs n'exigeront aucun prérequis à l'encontre des participants. Le premier cours se devra donc de prendre en compte cet élément, servant alors entre autres d'introduction à la gestion des données de recherche. Néanmoins, les formateurs proposeront en amont de la formation de suivre un cours d'introduction en ligne, par exemple le MANTRA ou le site d'UNIRIS[85], afin que les participants aient la possibilité d'acquérir des connaissances préalables. Le seul élément que les participants se devront de connaître concrètement avant le début du module, sera le type de données qu'ils généreront ou qu'ils ont déjà générés, afin de pouvoir constamment faire des parallèles entre la théorie du cours et leur pratique.

6.1 1er module : introduction à la gestion des données de recherche et création d'un Data Management Plan

Ce premier module serait donné aux chercheurs idéalement au début du cycle de vie de leurs données, lorsqu'ils débutent un nouveau cycle de recherche. Il s'agit du moment où ils doivent réaliser un DMP afin de clarifier le périmètre de leur étude ainsi

[85] https://uniris.unil.ch/researchdata/

que leurs outputs. Ce module étant d'une durée courte pour un sujet relativement vaste, le but sera de donner une idée générale sur la problématique ainsi que sur la réalisation d'un DMP. Ainsi, nous ne comptons pas sur le fait que les participants puissent compléter exhaustivement leur DMP durant la séance mais acquérir une base théorique et pratique afin qu'ils puissent le terminer seuls ou, dans le cas contraire, de prendre l'initiative de contacter les spécialistes en information documentaire[86] de la BIUM ou de l'IUMSP afin de les aider à le compléter.

6.1.1 1ère séquence : introduction à la gestion des données de recherche

Comme il a été dit ci-dessus, ce module n'exige aucun prérequis de la part des participants. Ainsi, il sera important de débuter la formation par une introduction sur la problématique de la gestion des données de recherche comprenant ces points entre autres :

- La nature des données de recherche → définition, types existants (données d'observation, données expérimentales…), les formats existants
- Pourquoi la nécessité d'une gestion → forte augmentation de la quantité des données, éviter la duplication des efforts, accélérer la coopération, etc.
- Intégrité scientifique et reproductibilité de la recherche
- Exigences des institutions, des organes financeurs et des éditeurs scientifiques
- Qu'est-ce qu'un DMP et ses enjeux

Cette introduction expliquera aux participants la raison de leur venue à cette formation et les enjeux qui découlent de cette thématique.

6.1.2 2e séquence : le cycle de vie des données

Durant cette deuxième séquence, nous présenterons le cycle de vie des données de recherche, à partir du schéma du site *data-archive.ac.uk*[87]*.* Ce dernier est, en effet, fréquemment réutilisé dans les bibliothèques afin d'expliquer ce cycle. Une courte partie théorique expliquera le principe de chaque étape du schéma.

L'exercice pratique allant avec cette séquence consistera au remplissage du schéma par les participants. Nous leur demanderons alors quelles sont les activités qu'ils pensent réaliser à chaque stade du cycle à la lumière de l'apport théorique précédent. Le but, ici, est de leur faire prendre conscience que la gestion des données ne s'arrête pas à leur analyse et que d'autres activités seront nécessaires par la suite (voir Annexe 1).

[86] Pablo Iriarte: pablo.iriarte@chuv.ch et Cécile Lebrand: cecile.lebrand@chuv.ch
[87] http://www.data-archive.ac.uk/create-manage/life-cycle

6.1.3 3e séquence : questionnaire en vue de la réalisation du DMP

La troisième séquence de la formation sera consacrée à l'aide à la création du DMP pour les participants. Pour cela, une série de questions inspirées de checklists sera donnée aux participants. Ce questionnaire, fragmenté en six parties, reprend globalement les sections du DMP qui leur sera en même temps fourni afin de faire un parallèle entre les questions que l'on pose et les sections du DMP.

Pour chaque partie, le formateur débutera par une brève explication théorique afin d'illustrer la problématique rencontrée. Par exemple, « qu'est-ce qu'une métadonnée et quelle est son utilité » ou encore, « le stockage au jour le jour et l'importance du versioning ». Ces explications se devront être concises, le but étant que les participants aient une vision globale du sujet afin qu'ils puissent cerner les questions qui leur sont posées.

Cet exercice se réaliserait idéalement par groupe de deux ou trois afin que les participants puissent enrichir leur vision des éléments à prendre en compte. Le questionnaire accompagné par le travail en groupe permettront de créer une certaine synergie et encourageront les participants à poser des questions, offrant une plus-value à une simple explication des différents points du DMP.

Le DMP pris en exemple à l'IUMSP est celui créé par les Universités Paris Descartes, Sorbonne et Paris Diderot (Cartier, Moysan & Reymonet 2015). Dans le cadre de cette partie pratique du module, nous ne prendrons en compte que les sections concernant concrètement les données. Ainsi, les premières sections du DMP, relevant de l'administratif, ne feront pas l'objet de ce cours.

Le questionnaire, présent dans sa totalité en tant qu'Annexe 2, comprend ces six parties :

1) **La nature des données**

Éléments théoriques : sautés car déjà vus dans l'introduction

Partie pratique : 4 questions

2) **Stockage, accès et sécurité des données**

Éléments théoriques : quel type de stockage pour la gestion quotidienne, pourquoi faire du versioning ?

Partie pratique : 2 questions

3) **Métadonnées – les données sur les données**

Éléments théoriques : que sont les métadonnées, à quoi servent-elles ?

Partie pratique : 1 question

Remarque : il ne sera pas nécessaire de trop développer la théorie de ce point car il sera mis en avant dans le second module.

4) **Partage et réutilisation des données**

Éléments théoriques : pourquoi partager ces données, principe de l'Open Data

Partie pratique : 3 questions

Remarque : idem que pour la partie 3

5) **Anonymisation des données**

Éléments théoriques : lois régulant cette thématique, les types de données à anonymiser, les types d'anonymisation pratiqués au CHUV.

Partie pratique : 1 question

Remarque : les éléments théoriques de cette partie seront plus développés que pour les autres.

6) **Archivage des données**

Éléments théoriques : différence entre dépôts privés et dépôts institutionnels, parler du dépôt *data@iumsp*

Partie pratique : 2 questions

Remarque : idem que pour les parties 3 et 4

6.1.4 Séquence pédagogique du 1er module

Tableau 2 : Séquence pédagogique du 1er module de formation

Heure	Minutage	Thème	Contenu/message	Méthode	Matériel
12h	3’	Présentation des intervenants	Connaître les intervenants et leur service	Présentation orale	Powerpoint
12h03	4’	Présentation des objectifs	Comprendre la problématique générale de la gestion des données de recherche et ses enjeux. Avoir une vision globale de leur DMP.	Présentation orale	Powerpoint
12h07	5’	Évaluation des connaissances	Demander aux participants s’ils ont déjà suivi une formation sur le sujet ou s’ils ont déjà réalisé un DMP. Ne pas forcément rentrer dans les détails.	Questions orales et témoignages concis	
12h12	20’	Introduction à la gestion des données de recherche	La nature des données Pourquoi une telle nécessité Importance de la reproductibilité des expériences Exigences des institutions, des organes financeurs et des éditeurs scientifiques Qu’est-ce qu’un DMP et ses enjeux	Présentation orale	Powerpoint
12h32	20’	Cycle de vie des données	Montrer et expliquer le schéma de www.data-archive.ac.uk Demander aux participants quelles seront leurs activités à chaque étape du cycle de vie Correction de l’exercice	Présentation orale Exercice pratique	Powerpoint Feuille avec schéma à remplir

Heure	Minutage	Thème	Contenu/message	Méthode	Matériel
12h52	3’	Explication sur l’utilité du questionnaire	Donner les questionnaires + les exemplaires de DMP. Montrer les parallèles entre les deux documents. Comparaison entre une question et une des sections du DMP	Présentation orale	Powerpoint Exemplaire DMP Questionnaire
12h55	8’	Nature des données	4 questions sur la nature des données Répondre aux éventuelles questions	Exercice pratique en groupe Interactions intervenants - participants	Questionnaire
13h03	10’	Stockage, accès et sécurité	Principe du versioning, stockage au quotidien, etc. 2 questions sur le stockage, l’accès et la sécurité Répondre aux éventuelles questions	Présentation orale Exercice pratique en groupe Interactions intervenants - participants	Powerpoint Questionnaire
13h13	8’	Métadonnées	Qu’est-ce que les métadonnées, à quoi servent-elles ? Préciser que le second module développe cette thématique 1 question sur les métadonnées Répondre aux éventuelles questions	Présentation orale Exercice pratique en groupe Interactions intervenants - participants	Powerpoint Questionnaire
13h21	12’	Partage et réutilisation des données	Pourquoi partager ses données & principes de l’Open Data. Préciser que le second module développe cette thématique 3 questions sur le partage des données Répondre aux éventuelles questions	Présentation orale Exercice pratique en groupe Interactions intervenants - participants	Powerpoint Questionnaire

Heure	Minutage	Thème	Contenu/message	Méthode	Matériel
13h33	14’	Anonymisation des données	Lois régulant cette thématique, les types de données à anonymiser, les types d’anonymisation pratiqués au CHUV. 1 question sur l’anonymisation Répondre aux éventuelles questions	Présentation orale Exercice pratique en groupe Interactions intervenants - participants	Powerpoint Questionnaire
13h47	12’	Archivage des données	Différence entre dépôts privés et dépôts institutionnels, parler du dépôt data@iumsp si nécessaire. Préciser que le second module développe cette thématique 2 questions sur l’archivage des données Répondre aux éventuelles questions	Présentation orale Exercice pratique en groupe Interactions intervenants - participants	Powerpoint Questionnaire
13h59	5’	Conclusion	Expliquer que les services de la BIUM et de l’uDDSP sont présents pour leur offrir un soutien	Présentation orale	Powerpoint avec contacts

Le minutage de cette séquence est relativement serré mais il est difficile de savoir, sans l’avoir testée, quels seront les questionnements des participants après chaque exercice pratique et quelle durée sera nécessaire à leur répondre. La durée totale du questionnaire avec explication est d’environ 67 minutes. Étant donné que deux minutes environ sont prévues à cet effet à chaque point du questionnaire, cela permet d’avoir une légère marge de manœuvre concernant le timing. Si d’aventure le minutage était trop serré, nous les utiliserions pour avancer dans l’exercice pratique ainsi que dans la théorie et nous proposerions de répondre aux questions en dehors de la formation.

6.2 2nd module : Open Data, formats, métadonnées et dépôt en ligne

Ce second module se donnerait idéalement après le premier, lorsque les chercheurs seraient déjà avancés dans leur processus de recherche. Bien que ne demandant pas de prérequis, il est admis que les chercheurs auront déjà généré des sets de données et seront plus au clair avec les thématiques liées aux données de la recherche. Ainsi, ce module ne sera pas un cours introductif généraliste comme le premier mais un module plus spécialisé concentré sur la notion de l'Open Data définit ici comme la mise en Open Access de sets de données accompagnant la publication. Cette séance nous amènera à définir le choix de formats de fichiers adaptés, la création des métadonnées et le choix du lieu de dépôt à longue durée. Néanmoins, il sera recommandé aux participants de se préparer au module en consultant le site web de la BIUM qui comporte un dossier complet en anglais sur les thématiques abordées[88].

De plus, il sera demandé, si cela est possible, que les participants amènent avec eux un set de leurs données générées au cours de leur recherche. En effet, comme nous réaliserons des exercices pratiques sur ordinateur, il serait plus intéressant que les chercheurs s'exercent directement avec leurs données.

6.2.1 1ère séquence : nécessités et bénéfices de l'Open Data

Cette première séquence expliquera aux participants quelles sont les causes majeures de non reproductibilité des études publiées en science et, dans ce cadre, quels sont les bénéfices à l'Open Data. Nous demanderons le point de vue des participants concernant cette question et utiliserons une enquête réalisée récemment auprès de 1500 chercheurs et publiée dans Nature[89]. Nous parlerons non seulement des bénéfices globaux de l'Open Data, pour le monde scientifique, mais aussi les bénéfices procurés aux chercheurs. De plus, il sera important de mentionner les directives que donnent les agences de financement et les revues scientifiques spécialisées concernant le partage des données de recherche accompagnant la publication.

Inspirés notamment par le sixième module de la formation de l'University of East London ainsi que du kit MANTRA, nous commencerons par un exercice pratique où nous demanderons aux chercheurs quels sont, d'après eux, les contraintes et les inquiétudes qui découragent au partage des données de recherche et inversement, quels sont les avantages qui encouragent ce partage. À cet effet, nous leur donnerons une feuille avec deux colonnes Pour/Contre où ils pourront résumer leurs idées. Cet

[88] http://www.bium.ch/publication-open-access/data-management/
[89] http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

exercice se fera par groupe de 2 à 4 personnes, les participants échangeant ainsi leurs idées et avis sur le sujet qui peuvent varier en fonction de la discipline de recherche (Voir Annexe 3).

Une discussion suivra cet exercice où les intervenants reprendront les points exposés par les participants, afin de relativiser les points négatifs et mettre en avant les points positifs. Cette séance de discussion servira de partie théorique mais interactive où les intervenants complèteront les réponses si d'aventure il manquait certains points qu'ils jugeraient nécessaires. L'University of East London ainsi que le cours MANTRA[90] ont d'ailleurs chacun pré-rempli une grille avec les réponses négatives fréquemment données par leurs participants et les réponses des intervenants qui relativiseront ces points.

Cette partie réponse/théorie sur les bénéfices du *data sharing* devra comprendre ces points entre autres :

- Bénéfices pour l'auteur
  - Meilleure diffusion et visibilité de son travail
  - Taux de citation amélioré
  - Méthodologies améliorées
  - Meilleure collaboration avec les pairs, rendant ses propres recherches de meilleure qualité
- Bénéfices pour la communauté scientifique
  - Accélère l'innovation dans les domaines de la technologie médicale
  - Augmentation de la reproductibilité des expériences, prévenant ainsi des surcoûts dans la recherche
  - Évitement de la duplication des expériences
  - Encouragement à la collaboration
  - Promotion de l'intégrité scientifique auprès du grand public

Sans oublier les directives des agences de financement, des revues spécialisées et les exigences/encouragements des institutions.

6.2.2 2e séquence : la création de métadonnées dans le but de décrire le set de données

Une première partie théorique consistera à expliquer ce que sont les métadonnées et quelle est leur utilité.

Dans un deuxième temps, dans le cadre d'un exercice pratique, nous demanderons aux participants de définir les métadonnées qu'ils désireraient avoir s'ils devaient

[90] http://datalib.edina.ac.uk/mantra/libtraining/Session5GroupExerciseAnswers3.pdf

réutiliser des données de recherche publiées. À l'inverse, nous leur demanderons aussi de se mettre à la place de leurs collègues et d'imaginer ce qu'ils désireraient avoir à disposition comme descriptions en accompagnement des données de recherche qu'eux-mêmes produisent et publient actuellement afin de pouvoir les comprendre et les reproduire. Cet exercice se ferait en groupe de 2 à 4 afin de que les participants puissent échanger et trouver de nouvelles idées associées à des données avec lesquelles ils sont peu familiers (Voir Annexe 4).

Dans un troisième temps, nous allons expliquer que certaines de ces métadonnées sont obligatoires, d'autres recommandées et, enfin, d'autres uniquement optionnelles. À partir de là, nous donnerons un document inspiré de celui du *Datacite*[91] et du site de la BIUM (BIUM 2016) qui listera les métadonnées selon leur importance. Finalement, nous engagerons une conversation afin de souligner les métadonnées qui n'auront pas été encore relevées et qui nous semblent importantes (Voir Annexe 5).

6.2.3 3e séquence : gestion des formats de fichiers pérennes

La première partie de cette séquence aura une dimension théorique où l'on expliquera la problématique des formats, qui peuvent être propriétaires ou « ouverts » et ce que cela engendre comme problématique pour la sauvegarde et le partage à long terme.

Nous donnerons alors aux participants une feuille récapitulant les formats principaux existants ainsi que les formats conseillés pour chaque type de données basé sur le site UK Data Archive (UK Data Archive 2002c) et l'Université d'Edinburgh (University of Edinburgh 2015c) (Voir Annexe 6).

À la lumière de ces nouvelles informations, nous récapitulerons avec les participants les types de données qu'ils créent lors de leur recherche et à partir de là, leur demander quels sont les formats qu'ils utiliseront lors de leur processus de recherche.

Nous travaillerons cet aspect de manière pratique avec les données de recherche des participants qu'ils auront amenés et pourront dès lors changer le format en celui qui est conseillé afin de voir comment procéder.

6.2.4 4e séquence : les licences CC pour protéger les copyrights des chercheurs

Pour débuter cette séquence, nous allons expliquer ce que sont les licences *Creative Commons* et différencier les possibilités existantes. Cette partie théorique s'attachera donc à définir les licences CC BY, CC BY-SA, CC BY-ND, CC BY-NC, CC BY-NC-SA

[91] https://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf

et enfin CC BY-NC-ND. Il serait utile de leur fournir un petit tableau récapitulatif de ses différentes licences (voir Annexe 7).

Après ces éléments théoriques nous laisserons les participants discuter entre eux afin qu'ils choisissent quelle licence CC ils désireraient utiliser pour leurs sets de données publiées. En poursuivant la conversation, les intervenants devront préciser que certaines directives peuvent être données par plusieurs parties prenantes dans le processus de recherche, notamment les institutions ou encore les revues spécialisées.

6.2.5 5e séquence : choix du dépôt en ligne pour les sets de données

Le choix du dépôt étant primordial, nous commencerons par expliquer pourquoi un dépôt online et/ou institutionnel est nécessaire pour la pérennisation et un accès facilité aux données de recherche. Il faudra expliquer les différences entre les dépôts institutionnels, qui ne sont pas conseillés pour l'UNIL pour l'instant exceptés pour l'IUMSP (*data@iumsp*), et les dépôts en ligne et publique. De même, il sera nécessaire d'expliquer l'existence des dépôts transitoires pour la gestion des données de recherche au quotidien. Toujours à propos des dépôts institutionnels, il faudra préciser qu'une solution nationale est actuellement en projet sous le nom de DLCM[92]. Concernant les dépôts en ligne et publique, il sera important de décrire ceux qui sont conseillés par la BIUM, donner leurs qualités mais aussi leurs limites. Cette description s'intéressera aux dépôts qui sont généralement mis en évidence dans la communauté scientifique dans le domaine médical, notamment *Zenodo, Figshare, Open Science Framework et Dryad* pour les données non-structurées et diversifiées. Concernant ce dernier, il sera important de préciser que bien que souvent conseillé par les éditeurs, *Dryad* comprend un problème d'importance dans le sens où le site impose aux auteurs de déposer leurs sets sous licence CC-0, c'est-à-dire sans la nécessité de citation lors de réutilisation. Cet élément, qui contrecarre plusieurs bénéfices mis en évidence dans la première séquence du module, ne nous semble pas approprié. *Open Science Framework* est, de son côté, intéressant pour les chercheurs car ce dépôt propose de *versionner* les sets de données. Concernant *Zenodo,* nous préciserons qu'une communauté, c'est-à-dire une forme de portail, a été créé pour la FBM par les services de la BIUM. Cela leur permet de contrôler la qualité des sets de données à propos des métadonnées, des licences et de les aider en cas de problème. *Figshare*, de son côté, semble être le dépôt le plus connu des chercheurs et donc celui où le choix des données à disposition est le plus intéressant.

[92] http://dlcm.ch/

Enfin, certaines revues scientifiques de renom proposent des dépôts officiels pour les données spécialisées lorsque l'on soumet un article à la publication. Il serait ainsi utile de préciser certains de ces dépôts notamment ceux spécialisés dans la génomique. Par exemple, les revues PLOS ont édité des directives sur les méthodes de dépôt et proposent divers dépôts en ligne selon les besoins des chercheurs[93].

L'exercice pratique de cette séquence consistera en l'inscription des participants aux dépôts *Zenodo* et *Figshare*, à partir de leur compte ORCID si possible afin d'unifier leurs profils. Ces inscriptions étant très rapides, cela ne sera pas un élément trop lourd dans la séquence. À partir de là, nous demanderons aux chercheurs de déposer un set de leurs données sur les différents sites avec les métadonnées et sous la licence CC appropriée. Il sera important de préciser aux participants que le dépôt sur les deux sites sera alors effectif et qu'il ne sera plus possible d'effacer le set. Si d'aventure ils ne désireraient pas déposer définitivement le set, il sera nécessaire de stopper l'exercice avant le dernier point.

Dans un second temps, nous demanderons aux participants de réaliser une recherche sur chaque dépôt afin d'essayer de trouver des sets de données qui pourraient être pertinents pour leur travail (voir Annexe 8).

[93] http://journals.plos.org/plosone/s/data-availability

6.2.6 Séquence pédagogique du 2nd module

Tableau 3 : Séquence pédagogique du 2nd module de formation

Heure	Minutage	Thème	Contenu/message	Méthode	Matériel
12h	3’	Présentation des intervenants	Connaître les intervenants et leur service	Présentation orale	Powerpoint
12h03	4’	Présentation des objectifs	Découvrir les enjeux et les bénéfices du *Data sharing* Savoir ce que sont les métadonnées et avoir connaissance de celles qui sont obligatoires. Savoir que certains formats sont plus conseillés que d’autres et avoir une idée sur les formats qu’ils utiliseront Avoir connaissance des différentes licences CC Découvrir Zenodo et Figshare, déposer et rechercher des sets de données	Présentation orale	Powerpoint
12h07	5’	Évaluation des connaissances	Demander aux participants s’ils ont suivi le premier module de formation ou suivi une autre formation sur le sujet. Demander aux participants s’ils ont déjà mis des sets de données sur des dépôts en ligne et si oui, lesquels.	Questions orales et témoignages concis	
12h12	20’	Open Data	Demander aux participants quels sont d’après eux les contraintes et les bénéfices au *Data sharing* Engager la discussion en demandant ce qu’ils ont noté. Répondre aux inquiétudes et	Exercice pratique Présentation orale	Questionnaire Powerpoint

Heure	Minutage	Thème	Contenu/message	Méthode	Matériel
			compléter les bénéfices s'ils n'ont pas tous été donnés. Explication des raisons du manque de reproductibilité des études scientifiques, les enjeux et bénéfices du Data sharing dans ce contexte.		
12h32	20'	Métadonnées	Explication théorique sur les métadonnées et leur utilité Demander aux participants de définir les métadonnées qu'ils désireraient avoir et lesquelles ils pensaient proposer aux autres Explication sur les standards de métadonnées avec celles qui sont obligatoires, conseillées et optionnelles	Présentation orale Exercice pratique	Powerpoint Questionnaire Tableau récapitulatif des métadonnées
12h52	15'	Gestion des formats pérennes	Explication théorique sur les formats propriétaires, ouverts, sur la pérennisation ainsi que les formats conseillés par type de données. Engager une courte réflexion avec les participants sur les formats qu'ils utiliseront à partir des types de données qu'ils créent. Possibilité de changer les formats de leurs données	Présentation orale Discussion Exercice pratique	Powerpoint Tableau récapitulatif des formats Ordinateur
13h07	15'	Licences CC et citations	Explication théorique sur le principe des licences *Creative Common* et décrire les licences les plus fréquentes. Discussion entre petits groupes de	Présentation orale Discussion	Powerpoint Tableau récapitulatif des licences

Heure	Minutage	Thème	Contenu/message	Méthode	Matériel
			participants afin qu'ils définissent quelle est la licence CC qu'ils utiliseraient		
13h23	35'	Dépôts de données en ligne	Explication théorique sur l'importance d'un dépôt en ligne, la différence entre les dépôts institutionnels et les dépôts privés. Parler de *data@iumsp* si l'origine des participants le requiert. Décrire les dépôts privés qui sont connus, leurs avantages et leurs défauts. Notamment Zenodo, Figshare, Open Science Framework et Dryad. Dire enfin que certaines institutions conseillent des dépôts particuliers, tout comme les revues spécialisées qui en recommandent afin d'accompagner les articles. Demander aux participants de s'inscrire, déposer un set de données, rechercher et télécharger un set de données sur Zenodo et Fighare	Présentation orale Exercice pratique sur ordinateur	Powerpoint Feuille d'exercice Ordinateur
13h58	5'	Conclusion	Expliquer que les services de la BIUM et de l'uDDSP sont présents pour leur offrir un soutien	Présentation orale	Powerpoint avec contacts précisés

7. Conclusion

La gestion des données de la recherche représente définitivement un enjeu à la fois scientifique et économique. Cette problématique est plus que jamais d'actualité lorsque l'on se rend compte de l'explosion de la masse de données qui sont créées dans les diverses branches scientifiques. Il semblerait que les différents acteurs se soient rendu compte de cette importance et aient entamé les actions allant dans ce sens.

Sous cet angle, la responsabilité des spécialistes en information documentaire se doit d'être engagée. Cette responsabilité doit se traduire par des actions concrètes, comprenant un soutien quotidien auprès des chercheurs, la création de plateformes et de documents adaptés à leurs besoins et, bien sûr, la réalisation de formations à leur égard reprenant les points essentiels du RDM.

Cette démarche proactive est d'autant plus souhaitable que la demande de la part des chercheurs est bien présente même si elle n'est pas toujours formulée ou, si elle l'est, de manière trop vague. De plus, les actions entreprises par les professionnels ID reçoivent, dans la grande majorité des cas, un accueil tout à fait favorable de la part des chercheurs.

Le contexte de la recherche biomédicale, en ce sens, fait figure de cas d'école. En effet, elle représente de manière concrète, voire accentuée due à son hyperspécialisation, tous les enjeux et les difficultés de la problématique de la gestion des données de la recherche. Ainsi, nous avons pu voir que les données générées par cette discipline étaient relativement complexes à appréhender. La diversité des types y est très présente, citons notamment les Omics, les données de laboratoires « long tail » et les données médicales. La difficulté (matérielle et financière) à les produire mais aussi et surtout à les reproduire est l'un des points cruciaux de cette branche scientifique rendant, de ce fait, le RDM et la création de métadonnées au centre des préoccupations. Le *Data sharing* est, de plus, d'une totale nécessité afin de rendre la recherche plus efficiente et plus crédible auprès du grand public.

De même, le volume des données peut y être très conséquent, dû notamment au séquençage des gènes nécessitant de ce fait des dépôts et du matériel informatiques puissants dans l'optique de les traiter. Enfin, la problématique des données médicales ne va pas sans un volet légal et éthique entourant le respect de la personne et le secret médical. À cette fin, l'anonymisation des données est une étape certes complexe, mais obligatoire dans cette branche scientifique.

Dans ce contexte, les institutions de la BIUM et de l'uDDSP s'inscrivent totalement dans les besoins cités ci-dessus. De nombreuses initiatives et services ont déjà été réalisés et une formation sur la gestion des données de la recherche semble être tout indiquée pour prolonger cette politique.

Après avoir étudié plusieurs formations existantes, nous pouvons en tirer quelques conclusions sur la forme ainsi que sur le fond qu'il sera intéressant d'appliquer.

Premièrement, il sera pertinent de briser les idées reçues et de mettre en avant les points positifs du RDM. Il faudra notamment évoquer les bénéfices pour les chercheurs mais aussi les bénéfices dont la science dans son ensemble pourra profiter. Deuxièmement, un équilibre entre théorie et pratique semble être requis afin d'intéresser les participants tout en leur apportant les connaissances nécessaires. Troisièmement, les formations étudiées comportent à la fois du contenu généraliste, qui peut s'appliquer à toutes les gestions de données de recherche mais aussi du contenu spécifique qui permet de mettre les participants dans leur contexte de travail.

De plus, il sera important de ne pas figer la formation mais de la faire évoluer avec son contexte afin de constamment répondre au mieux aux besoins des chercheurs.

Enfin, les institutions ont aussi leur part de responsabilité. Au-delà de l'infrastructure qu'elles peuvent offrir et des directives qui sont mises en place, elles se devront d'encourager et promouvoir ce type d'initiatives auprès des chercheurs afin qu'elles rencontrent un succès digne de l'investissement en temps.

Lorsque l'on désire réaliser une formation, le danger est de vouloir être trop « généreux », c'est-à-dire de prévoir des modules trop complets qui s'avèrent être irréalistes à donner durant le temps imparti. Il a donc été nécessaire de définir au mieux la problématique afin de sélectionner les chapitres nécessaires à l'enseignement. De même, il a été important de garder constamment à l'esprit les réels besoins des mandants afin de ne pas s'écarter du périmètre de la formation. En ce qui concerne ce mandat, le but a constamment été de coller au mieux à la réalité. C'est d'ailleurs pour cette raison que des rendez-vous ont été réalisés afin d'appréhender au mieux les besoins qui découlaient de cette formation. En ce sens, nous espérons avoir réalisé un projet réaliste qui puisse être utilisé par la BIUM et l'uDDSP.

Bibliographie

Les données de la recherche :

BEGLEY, C. Glenn et IOANNIDIS, John P. A., 2015. Reproducibility in Science Improving the Standard for Basic and Preclinical Research. Circulation Research. 2 janvier 2015. Vol. 116, n° 1, pp. 116-126.

BOSTON UNIVERSITY, [sans date]. What Is « Research Data »? Site Internet de l'Université de Boston [en ligne]. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.bu.edu/datamanagement/background/whatisdata/

BURNHAM, Andrew, 2012. Research Data - Definitions [en ligne]. University of Leicester. 2012. [Consulté le 2 juillet 2016]. Disponible à l'adresse : https://www2.le.ac.uk/services/research-data/documents/UoL_ReserchDataDefinitions_20120904.pdf

BURNHAM, Andrew, 2013. An Introduction to Managing Research Data For Researchers and Students [en ligne]. University of Leicester. 2013. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www2.le.ac.uk/services/research-data/documents/an-introduction-to-managing-research-data

CARTIER, Aurore, MOYSAN, Magalie et REYMONET, Nathalie, 2015. Réaliser un plan de gestion des données [en ligne]. Université Paris Diderot. 2015. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.univ-paris-diderot.fr/DocumentsFCK/recherche/Realiser_un_DMP_V1.pdf

CLAIROUX, Natalie, 2016. Introduction à la gestion des données de recherche [en ligne]. Université de Montréal. 2016. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.bib.umontreal.ca/sa/gestion-donnees-recherche.pdf

CODATA-ICSTI, 2013. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal. 2013. Vol. 12, pp. CIDCR1-CIDCR75.

COMITÉ D'ETHIQUE DU CNRS, 2015. Les enjeux éthiques du partage des données scientifiques [en ligne]. CNRS. 7 mai 2015. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.cnrs.fr/comets/IMG/pdf/2015-05_avis-comets-partage-donnees-scientifiques-2.pdf

COOPIST, 2015. Pourquoi gérer les données de la recherche ? Site Internet de Coopérer en information scientifique et technique [en ligne]. 2012-2015. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://coop-ist.cirad.fr/gestion-de-l-information/gestion-des-donnees-de-la-recherche/decouvrir-des-plans-de-gestion-de-donnees-de-la-recherche/1-pourquoi-gerer-les-donnees-de-la-recherche

CORTI, Louise (éd.), 2014. Managing and sharing research data: a guide to good practice. Los Angeles, Calif : SAGE. ISBN 9781446267257

DEBOIN, Marie-Claude, 2014. Découvrir des plans de gestion de données de la recherche en 4 points [en ligne]. CIRAD. 2014. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://coop-ist.cirad.fr/content/download/5435/40362/version/4/file/CoopIST-plan-gestion-donnees-recherche-20140717.pdf

DIGITAL CURATION CENTRE, 2004. DCC Curation Lifecycle Model. Site Internet du DCC [en ligne]. 2004-2016. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.dcc.ac.uk/resources/curation-lifecycle-model

EPFL, 2016. Open Research Data Pilot. Site Internet de l'EPFL [en ligne]. 2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://research-office.epfl.ch/financements/international/horizon-2020/open-research-data-pilot

ETHZ, 2016. Digital Curation at ETH Zurich. Site Internet de l'ETH Zurich [en ligne]. 2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.library.ethz.ch/en/ms/Digital-Curation-at-ETH-Zurich

FACHINOTTI, Elena, GOZZELINO, Eva, LONATI, Sara et SCHNEIDER, René (Dir), 2016. Les bibliothèques scientifiques et les données de la recherche [en ligne]. Genève : Haute école de gestion de Genève. Mémoire d'étude. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://doc.rero.ch/record/258991

GAILLARD, Rémi, 2014. De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ? [en ligne]. Lyon : Enssib. Mémoire d'étude. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://eprints.rclis.org/22746/

GAILLARD, Rémi, 2015. L'ouverture des données de la recherche en 2015 : définitions, enjeux [en ligne]. Sorbonne Universités. 2015. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://fr.slideshare.net/invisu/louverture-des-donnes-de-la-recherche-en-2015-dfinitions-enjeux-dynamiques

HORIZON 2020, 2013a. Modèle général de convention de subvention multibénéficiaire, version 1.0, 11 décembre 2013 [en ligne]. Programme-cadre européen pour la recherche et l'innovation. 11 décembre 2013. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_v1.0_fr.pdf

HORIZON 2020, 2013b. Lignes directrices pour la gestion des données dans Horizon 2020, Version 1.0, 11 décembre 2013 [en ligne]. Programme-cadre européen pour la recherche et l'innovation. 16 décembre 2013. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://openaccess.inist.fr/IMG/pdf/14081_lignes_directrices_pgd_horizon_2020_tr_fr_versionavril2015-2.pdf

HORIZON 2020, 2013c. Lignes directrices pour le libre accès aux publications scientifiques et aux données de recherche dans Horizon 2020, Version 1.0, 11 décembre 2013 [en ligne]. Programme-cadre européen pour la recherche et l'innovation. 16 décembre 2013. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://openaccess.inist.fr/IMG/pdf/14086_lignes_directrices_la_horizon_2020_tr_fr_version-oct2014.pdf

HORIZON 2020, 2016a. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, Version 2.1, 15 February 2016 [en ligne]. Programme-cadre européen pour la recherche et l'innovation. 2016. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

HORIZON 2020, 2016b. Guidelines on Data Management in Horizon 2020, Version 2.1, 15 February 2016 [en ligne]. Programme-cadre européen pour la recherche et l'innovation. 15 février 2016. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

INSTITUT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE, 2004. Initiative de Budapest pour l'Accès Ouvert - Libre accès à l'information scientifique et technique. Site Internet de l'INIST pour l'Open Access [en ligne]. 2004. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://openaccess.inist.fr/?Initiative-de-Budapest-pour-l

INSTITUT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE, 2014. Une introduction à la gestion et au partage des données de la recherche [en ligne]. INIST. 16 septembre 2014. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.inist.fr/donnees/co/Donnees_recherche_web.html

INSTITUT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE, [sans date]. Données de la recherche. Site Internet de l'INIST [en ligne]. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.inist.fr/?Donnees-de-la-recherche&lang=fr

INTERNATIONAL VIRTUAL OBSERVATORY ALLIANCE, [sans date]. IVOA.net. Site Internet de l'IVOA [en ligne]. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.ivoa.net/

JACQUEMOT-PERBAL, M.-C, COSSERAT, F., 2015. Gestion et diffusion des données de la recherche [en ligne]. Formation URFIST Rennes. 11 juin 2015. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.inist.fr/IMG/pdf/urfistrennes_20150616.pdf

JAMBÉ, Carmen, 2015. La gestion des données de recherche à l'Université de Lausanne [en ligne]. Genève : Haute école de gestion de Genève. Travail de Bachelor. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://doc.rero.ch/record/258023

MILHIT, Igor, 2012. Enjeux de l'archivage à long terme des données primaires de la recherche scientifique [en ligne]. Genève : Haute école de gestion de Genève. Travail de Bachelor. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://doc.rero.ch/record/30352

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE, 2011. Données de la Recherche. Site Internet d'information sur les données de la recherche [en ligne]. 2011-2016. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.donneesdelarecherche.fr/

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE, 2013. Le libre accès aux publications et aux données de recherche - Horizon 2020. Portail français d'Horizon 2020 [en ligne]. 2013. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html

NATIONAL INSTITUTES OF HEALTH, 2003. Data Sharing Policy and Implementation Guidance. Site Internet de la NIH [en ligne]. 09.02.2012. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#ex

NATIONAL SCIENCE BOARD, 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century [en ligne]. National Science Foundation. Septembre 2005. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

OCDE, 2007. Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics [en ligne]. OCDE. 2007. [Consulté le 2 juillet 2016]. Disponible à l'adresse : https://www.oecd.org/fr/sti/sci-tech/38500823.pdf

PRYOR, Graham (éd.), 2014. Delivering research data management services: fundamentals of good practice. London : Facet. ISBN 9781856049337.

REYMONET, Nathalie, 2015. L'open access (OA) dans la production des connaissances [en ligne]. Université Paris Diderot. 2015. [Consulté le 2 juillet 2016]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-01115565

SCHNEIDER, René, 2013. Research Data Literacy. In : Worldwide Commonalities and Challenges IN Information Literacy Research and Practice. Springer International Publishing. pp. 134-140. ISBN 978-3-319-03918-3

SCIENCE OPEN, 2015. ScienceOpen. Site Internet de Science Open, Research and Publishing Network [en ligne]. 2015-2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.scienceopen.com/

SHEARER, Kathleen, 2014. Mise en oeuvre de politiques de gestion des données de recherche [en ligne]. Sous-comité des politiques de Données de recherche Canada. Mai 2014. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.rdc-drc.ca/wp-content/uploads/fKM-1780J_RDC-Implementing-RDM-Policies-Backgrounder.pdf

SWISS UNIVERSITIES, [sans date]. Programme CUS 2013-2016 P-2. Site Internet de Swissuniversities [en ligne]. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.swissuniversities.ch/fr/organisation/projets-et-programmes/programme-cus-2013-2016-p-2-information-scientifique-acces-traitement-et-sauvegarde/

THESSEN, Anne et PATTERSON, David, 2011. Data issues in the life sciences. ZooKeys. 28 novembre 2011. Vol. 150, pp. 15-51

UK DATA ARCHIVE, 2002a. Data management planning. Site Internet de l'UK Data Archive [en ligne]. 2002-2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.ukdataservice.ac.uk/manage-data/plan/planning

UK DATA ARCHIVE, 2002b. Research data lifecycle. Site Internet de l'UK Data Archive [en ligne]. 2002-2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.data-archive.ac.uk/create-manage/life-cycle

UNIVERSITÉ BRETAGNE-LOIRE, [sans date]. Les données de la recherche, introduction. Site Internet de l'Université Bretagne-Loire [en ligne]. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://guides-formadoct.ueb.eu/donnees_recherche

UNIL, UNIRIS, 2014. Archives des savoirs : De la gestion des données de recherche vers une gestion des données pour la recherche [en ligne]. Université de Lausanne. 16 octobre 2014. [Consulté le 2 juillet 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

UNIVERSITY OF BRISTOL, 2002. Research Data Service. Site Internet de l'Université de Bristol [en ligne]. 2002-2012. [Consulté le 1 juillet 2016]. Disponible à l'adresse : https://data.bris.ac.uk/

UNIVERSITY OF BRISTOL, 2013a. Managing research data [en ligne]. Université de Bristol. 2013. [Consulté le 1 juillet 2016]. Disponible à l'adresse : https://data.bris.ac.uk/files/2013/03/Grant-writing-Med-databris.pdf

UNIVERSITY OF BRISTOL, 2013b. data.bris Benefits Report [en ligne]. Université de Bristol. 6 mars 2013. [Consulté le 2 juillet 2016]. Disponible à l'adresse : https://data.bris.ac.uk/files/2013/06/data-bris-benefits-report-V2.pdf

UNIVERSITY OF EDINBURGH, [sans date]. Edinburgh Data Management Plan Template [en ligne]. Université d'Edinburgh. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.ed.ac.uk/files/imports/fileManager/Edinburgh_DMP_template_web.pdf

UNIVERSITY OF EDINBURGH, 2015a. Research Data Management (RDM) Roadmap August 2012 – July 2016 [en ligne]. Université d'Edinburgh. Septembre 2015. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.ed.ac.uk/files/atoms/files/uoe-rdm-roadmap_-_v2_0_0.pdf

UNIVERSITY OF EDINBURGH, 2015b. Research Data Management Policy. Site Internet de l'Université d'Edinburgh [en ligne]. 2015-2016. [Consulté le 2 juillet 2016]. Disponible à l'adresse : http://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy

Les données biomédicales et contexte CHUV-IUMSP :

ABOU EL KALAM, Anas, DESWARTE, Yves, TROUESSIN, Gilles et CORDONNIER, Emmanuel, 2004a. Gestion des données médicales anonymisées: problèmes et solutions. ResearchGate [en ligne]. 2004. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.researchgate.net/publication/228422166_Gestion_des_donnees_medicales_anonymisees_problemes_et_solutions

ABOU EL KALAM, Anas, DESWARTE, Yves, TROUESSIN, Gilles et CORDONNIER, Emmanuel, 2004b. Une démarche méthodologique pour l'anonymisation de données personnelles sensibles. ResearchGate [en ligne]. 2004. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.researchgate.net/publication/229000683_Une_demarche_methodologique_pour_l'anonymisation_de_donnees_personnelles_sensibles

ANDERSON, Nicholas et al., 2007. Issues in Biomedical Research Data Management and Analysis: Needs and Barriers. Journal of the American Medical Informatics Association. 2007. Vol. 14, n° 4, pp. 478-488. DOI 10.1197/jamia.M2114.

AROFAN, Gregory, 2011. The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes [en ligne]. Open Data Foundation. 2011. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://odaf.org/papers/DDI_Intro_forNSIs.pdf

BENCHEIKH, Soulaymani, [sans date]. La vie d'un médicament [en ligne]. Centre Anti Poison du Maroc. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.who.int/medicines/areas/quality_safety/safety_efficacy/trainingcourses/1vie_medicaments.pdf

BESSE, Camille, 2015. Guide des bonnes pratiques relatives à la mise en valeur des e-books [en ligne]. Haute école de gestion de Genève. Travail de Bachelor. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://doc.rero.ch/record/257989/

BIUM, 2016. Data management & Open Data. Site Internet de la BIUM [en ligne]. 2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.bium.ch/en/publication-open-access/data-management/

BRUEL, J-M et GALLIX, B, 2007. Imagerie Médicale : Bases techniques Indications, Risques, Bénéfices [en ligne]. CHU Montpellier. 2007. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.med.univ-montp1.fr/enseignement/cycle_1/PCEM2/mod-base/MB3_biophysique/Ressources_locale/Imagerie_radio/LIPCOM_MB3_Biophysique_Rayonnement_Imagerie_diagnostique.pdf

BULL, Susan et al., 2015. Best Practices for Ethical Sharing of Individual-Level Health Research Data From Low- and Middle-Income Settings. Journal of Empirical Research on Human Research Ethics. 1 juillet 2015. Vol. 10, n° 3, pp. 302-313.

BULL, Susan, ROBERTS, Nia et PARKER, Michael, 2015. Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public Health Research: A Systematic Scoping Review. Journal of Empirical Research on Human Research Ethics. juillet 2015. Vol. 10, n° 3, pp. 225-238.

CAMBRIDGE HEALTHTECH INSTITUTE, 2015. Omes and -omics glossary & taxonomy. Site Internet du Cambridge Healthtech Institute [en ligne]. 12.01.2015. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.genomicglossaries.com/content/omes.asp

CARR, David et LITTLER, Katherine, 2015. Sharing Research Data to Improve Public Health A Funder Perspective. Journal of Empirical Research on Human Research Ethics. 1 juillet 2015. Vol. 10, n° 3, pp. 314-316.

CHUV, 2016a. Biobanque Institutionnelle de Lausanne - CHUV. Site Internet de la biobanque du CHUV [en ligne]. 20.06.2016. [Consulté le 6 juillet 2016]. Disponible à l'adresse : http://www.chuv.ch/biobanque/bil_home.htm

CHUV, 2016b. Cellular Imaging Facility. Site Internet du Cellular Imaging Facility [en ligne]. 30.05.2016. [Consulté le 6 juillet 2016]. Disponible à l'adresse : http://cifweb.unil.ch/

CONFÉDÉRATION SUISSE, 2002. Guide relatif au traitement des données personnelles dans le domaine médical : Traitement des données personnelles par des personnes privées et des organes fédéraux [en ligne]. Confédération Suisse. 2002. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.edoeb.admin.ch/datenschutz/00628/00629/index.html?lang=fr

CONFÉDÉRATION SUISSE, 2014. RS 810.305 Ordonnance du 20 septembre 2013 sur les essais cliniques dans le cadre de la recherche sur l'être humain (Ordonnance sur les essais cliniques, OClin). Site Internet relatif au droit fédéral [en ligne]. 02.07.2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.admin.ch/opc/fr/classified-compilation/20121176/index.html

CUGGIA, Marc, 2016. Exploitation des données massives en santé pour la recherche médicale : méthodes, outils et cas d'utilisation [en ligne]. IUMSP. 2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.iumsp.ch/sites/default/files/Colloque_IUMSP_20160412.pdf

DDI, 2015. Data Documentation Initiative. Site Internet de la DDI [en ligne]. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://www.ddialliance.org/

DIRECTION DE LA PROSPECTIVE ET DU DIALOGUE PUBLIC, 2013. La recherche clinique [en ligne]. Direction de la Prospective et du Dialogue Public. 2013. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.cengeps.fr/sites/default/files/fichiers-pages/Recherche_clinique_agglo_lyonnaise_2013.pdf

FACULTÉ DE BIOLOGIE ET DE MÉDECINE, [sans date]. Center for Integrative Genomics. Site Internet du Center for Integrative Genomics [en ligne]. [Consulté le 6 juillet 2016 a]. Disponible à l'adresse : https://www.unil.ch/cig/en/home/menuinst/research/core-facilities/dr-harshman---gtf.html

FACULTÉ DE BIOLOGIE ET DE MÉDECINE, [sans date]. Mouse Metabolic Evaluation Facility. Site Internet du Center for Integrative Genomics [en ligne]. [Consulté le 6 juillet 2016 b]. Disponible à l'adresse : https://www.unil.ch/cig/en/home/menuinst/research/core-facilities/mef.html

FACULTÉ DE BIOLOGIE ET DE MÉDECINE, [sans date]. Protein Analysis Facility. Site Internet du Center for Integrative Genomics [en ligne]. [Consulté le 6 juillet 2016 c]. Disponible à l'adresse : https://www.unil.ch/cig/en/home/menuinst/research/core-facilities/dr-quadroni---paf.html

FACULTÉ DE BIOLOGIE ET DE MÉDECINE, [sans date]. Vital-IT. Site Internet de Vital-IT [en ligne]. [Consulté le 6 juillet 2016 d]. Disponible à l'adresse : https://www.unil.ch/cig/en/home/menuinst/research/core-facilities/vital-it.html

FERGUSON, Adam et al., 2014. Big data from small data: data-sharing in the 'long tail' of neuroscience. Nature neuroscience. novembre 2014. Vol. 17, n° 11, pp. 1442-1447.

GAANS, Deborah van et al., 2015. The Development of the Public Health Research Data Management System. Electronic Journal of Health Informatics. 8 mai 2015. Vol. 9, n° 1, pp. 10.

HAREL, Arye et al., 2011. Omics Data Management and Annotation. In : Bioinformatics for Omics Data [en ligne]. Humana Press. pp. 71-96. Methods in Molecular Biology, 719. [Consulté le 3 juillet 2016]. ISBN 978-1-61779-026-3. Disponible à l'adresse : http://dx.doi.org/10.1007/978-1-61779-027-0_3

HATE, Ketaki et al., 2015. Sweat, Skepticism, and Uncharted Territory: A Qualitative Study of Opinions on Data Sharing Among Public Health Researchers and Research Participants in Mumbai, India. Journal of Empirical Research on Human Research Ethics. juillet 2015. Vol. 10, n° 3, pp. 239-250.

HAUT CONSEIL DE LA SANTÉ PUBLIQUE, 2009. Les systèmes d'information pour la santé publique [en ligne]. Paris : Haut Conseil de la Santé Publique. 2009. [Consulté le 9 juillet 2016]. Disponible à l'adresse : http://www.hcsp.fr/explore.cgi/avisrapportsdomaine?clefr=175

IRIARTE, Pablo, 2016. data@iumsp: IUMSP Research Data Repository. IUMSP. 2016.

IUMSP, [sans date]. A propos de l'IUMSP. Site Internet de l'IUMSP [en ligne]. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.iumsp.ch/fr/a-propos

IUMSP, 2015a. Au service de la santé publique [en ligne]. IUMSP. 2015. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.iumsp.ch/sites/default/files/pdf/depliant_IUMSP_fr.pdf

IUMSP, 2015b. Brochure d'accueil de l'IUMSP [en ligne]. IUMSP. 2015. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.iumsp.ch/sites/default/files/pdf/brochure_accueil_2015_31.03.2015.pdf

IUMSP, 2016a. DATA@IUMSP : IUMSP Research Data Repository Home. Dépôt institutionnel de l'IUMSP [en ligne]. 2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://data.iumsp.ch/home

IUMSP, 2016b. Rapport annuel 2015 [en ligne]. IUMSP. 2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.iumsp.ch/sites/default/files/pdf/IUMSP_rapport_annuel_2015.pdf

JACQUEMONT, Nathalie, 2016. Gestion des données cliniques des patients pour la Recherche Clinique [en ligne]. CHUV. 21 mars 2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://uniris.unil.ch/files/researchdata/document/05_Biobanque_BIL_Nathalie-JACQUEMONT.pdf

JOUIS, Véronique, GUERY, Laurence et VICAUT, Eric, 2011. Les différentes phases en recherche clinique [en ligne]. 2011. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.rechercheclinique pariscentre.fr/wp-content/uploads/2012/01/PhasesRC-COURS-IRC-TEC-4-nov-2011.pdf

LES ENTREPRISES DU MÉDICAMENT, 2013. Le développement préclinique ou la première évaluation. Site Internet des entreprsies du médicament [en ligne]. 2013. [Consulté le 8 juillet 2016]. Disponible à l'adresse : http://www.leem.org/article/developpement-preclinique-premiere-evaluation-0

LIBRARY OF CONGRESS, [sans date]. Format Description Categories - Sustainability of Digital Formats. Site Internet de la préservation digitale [en ligne]. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml

MICHEEL, Christine M., NASS, Sharly J. et OMENN, Gilbert S. (éd.), 2012. Evolution of Translational Omics: Lessons Learned and the Path Forward [en ligne]. Washington, D.C. : National Academies Press. [Consulté le 9 juillet 2016]. ISBN 978-0-309-22418-5. Disponible à l'adresse : http://www.nap.edu/catalog/13297

NATIONAL INSTITUTES OF HEALTH, 2015. Principles and Guidelines for Reporting Preclinical Research. Site Internet de la NIH [en ligne]. 05.01.2015. [Consulté le 3 juillet 2016]. Disponible à l'adresse : https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

NELSON, Gregory S., 2015. Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification [en ligne]. ThotWave Technologies. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf

OMICS.ORG, 2013. Omes and Omics. Site Internet d'Omics.org [en ligne]. 18.08.2013. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://omics.org/index.php/Omes_and_Omics

OMS, 2003. Méthodologie de la recherche dans le domaine de la santé : guide de formation aux méthodes de la recherche scientifique [en ligne]. OMS. 2003. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.wpro.who.int/health_research/documents/dhs_hr_health_research_methodology_a_guide_for_training_in_research_methods_second_edition_fr.pdf

PARKER, Michael et BULL, Susan, 2015. Sharing Public Health Research Data: Toward the Development of Ethical Data-Sharing Practice in Low- and Middle-Income Settings. Journal of Empirical Research on Human Research Ethics. juillet 2015. Vol. 10, n° 3, pp. 217-224.

PISANI, Elizabeth et ABOUZAHR, Carla, 2010. Sharing health data: good intentions are not enough. Bulletin of the World Health Organization. juin 2010. Vol. 88, n° 6, pp. 462-466.

POCHON, Mireille, 2000. L'enseignement de la médecine à l'ère numérique : le développement des technologies de l'information et de la communication au sein de la Médiathèque de la Faculté de médecine de Lausanne.. Genève : Haute école de gestion de Genève. Travail de Bachelor.

Q-CROC, 2010. Qu'est-ce que la recherche translationnelle? Site Internet du Consortium de recherche en oncologie clinique [en ligne]. 2010. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.qcroc.ca/informations-aux-patients/quest-ce-que-la-recherche-translationnelle

SANTÉ PUBLIQUE SUISSE, 2013. Des données de meilleure qualité pour augmenter l'efficacité du système de santé [en ligne]. Santé Publique Suisse. 2013. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.hausarztmedizin.uzh.ch/dam/jcr:ffffffff-b788-cf1c-ffff-ffff9dc32b6d/130816-ManifestGesundheitsdaten-F-def.pdf

SANTÉ PUBLIQUE SUISSE, [sans date]. Santé publique Suisse. Site Internet de Santé Publique Suisse [en ligne]. [Consulté le 9 juillet 2016]. Disponible à l'adresse : http://www.public-health.ch/logicio/pmws/publichealth__publichealth__fr.html

TRANSCELERATE BIOPHARMA, 2013. Data De-identification and Anonymization of Individual Patient Data in Clinical Studies: A Model Approach [en ligne]. Transcelerate Biopharma. 2013. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-FINAL.pdf

UK DATA ARCHIVE, 2002c. File Formats Table. Site Internet de l'UK Data Archive [en ligne]. 2002-2016. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.data-archive.ac.uk/create-manage/format/formats-table

UNIVERSITÉ PARIS DESCARTES, 2010. Essais cliniques. Université Paris Descartes. 2010.

UNIVERSITY OF EDINBURGH, 2015c. Edinburgh DataShare: Recommended File Formats [en ligne]. Université d'Edinburgh. 2015. [Consulté le 3 juillet 2016]. Disponible à l'adresse : http://www.ed.ac.uk/files/atoms/files/recommended_file_formats-apr2015.pdf

WADE, Ted D., 2014. Traits and types of health data repositories. Health Information Science and Systems. 2014. Vol. 2, pp. 4.

YOO, Illhoi et al., 2012. Data mining in healthcare and biomedicine: a survey of the literature. Journal of Medical Systems. août 2012. Vol. 36, n° 4, pp. 2431-2448.

Formations existantes sur les données de la recherche :

CANAVAN, MJ, MCGINTY, Steve et REZNIK-ZELLEN, Rebecca, 2015. New England Collaborative Data Management Curriculum - Module 4: Data Storage, Backup, and Security [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/modules

CLAIROUX, Natalie, 2015. En Français S'il Vous Plaît: Translation and Adaptation of the New England Collaborative Data Management Curriculum's Introductory Module. Journal of eScience Librarianship [en ligne]. 14 août 2015. Vol. 4, n° 1. DOI 10.7191/jeslib.2015.1079. Disponible à l'adresse : http://escholarship.umassmed.edu/jeslib/vol4/iss1/7

CLAIROUX, Natalie, 2016. Introduction à la gestion des données de recherche [en ligne]. Université de Montréal. 2016. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://www.bib.umontreal.ca/sa/gestion-donnees-recherche.pdf

COBURN, Elizabeth, FURFEY, John et WALTON, Jen, 2015. New England Collaborative Data Management Curriculum - Module 3: Contextual Details Needed to Make Data Meaningful to Others [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/modules

CREAMER, Andrew et al., 2015. New England Collaborative Data Management Curriculum - Module 1: Overview of Research Data Management [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/modules

DIGITAL CURATION CENTRE, 2004. University of East London | Digital Curation Centre. Site Internet du DCC [en ligne]. 2004-2016. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://www.dcc.ac.uk/tailored-support/institutional-engagements/east-london

DUKE, Monica, 2014a. SupportDM (UEL) Module 5: Cataloguing Data [en ligne]. University of East London. 2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28035

DUKE, Monica, 2014b. SupportDM (UEL) Module 6: Sharing Data [en ligne]. University of East London. 2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28320

ETH BIBLIOTHEK et EPFL LIBRARY, 2016. Data Management Checklist. ETHZ. 2016

FERGUSON, Jen, 2015. New England Collaborative Data Management Curriculum - Module 2: Types, Formats, and Stages of Data [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/modules

GRACE, Stephen et MCELROY, David, 2014. Sharing and Archiving Your Research Data Workshop Materials (UEL) [en ligne]. University of East London. 2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28323

GRACE, Stephen et MCELROY, David, 2015. Writing a Data Management Plan - Workshop Materials (UEL) [en ligne]. University of East London. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28324

GRACE, Stephen, 2014a. SupportDM (UEL) Module 2: guidance and support for researchers [en ligne]. University of East London. 2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28032

GRACE, Stephen, 2014b. SupportDM (UEL) Module 4: What Data to Keep and Why [en ligne]. University of East London. 2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28034

JONES, Sarah, MURTAGH, John et GRACE, Stephen, 2014. SupportDM (UEL) Module 1: Introducing RDM [en ligne]. University of East London. 2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : https://zenodo.org/record/28031

JONES, Sarah, 2014. SupportDM (UEL) Module 3: Data Management Planning [en ligne]. University of East London. 2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28033

KAFEL, Donna, PALMER, Lisa et PIORUN, Mary, 2015. New England Collaborative Data Management Curriculum - Module 5: Legal and Ethical Considerations for Research Data [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/modules

LOWE, David, WHITE, Darla et NOVAK GUSTAINIS, Emily, 2015. New England Collaborative Data Management Curriculum - Module 7: Repositories, Archiving & Preservation [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/modules

MANTRA, 2014. MANTRA - Library Training. Site Internet du cours Mantra [en ligne]. 10.09.2014. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://datalib.edina.ac.uk/mantra/libtraining.html

MANTRA, 2015. MANTRA. Site Internet du cours Mantra [en ligne]. 18.08.2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://datalib.edina.ac.uk/mantra/

MCELROY, David, JONES, Sarah et GRACE, Stephen, 2015. Managing Your Research Data Workshop Materials (UEL) - Slides and Workbook [en ligne]. University of East London. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://zenodo.org/record/28322

MCLEOD, Julie, CHILDS, S. et LOMAS, E., 2011. Practical Exercise 1. Research scenarios [en ligne]. Northumbria University. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : https://www.northumbria.ac.uk/static/5007/ceispdf/scenarios.pdf

MCLEOD, Julie, 2011. DATUM for Health Research data management training for health studies: JISC Final Report [en ligne]. Northumbria University. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : https://www.northumbria.ac.uk/static/5007/ceispdf/report.pdf

NORTHUMBRIA UNIVERSITY, 2011a. Session 1 : Introduction to Research Data Management [en ligne]. Northumbria University. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://find.jorum.ac.uk/resources/18276

NORTHUMBRIA UNIVERSITY, 2011b. Session 2 : Data Curation Lifecycle [en ligne]. Northumbria University. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://find.jorum.ac.uk/resources/18276

NORTHUMBRIA UNIVERSITY, 2011c. Session 3 : Problems and Practical Strategies and Solutions [en ligne]. Northumbria University. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://find.jorum.ac.uk/resources/18276

NORTHUMBRIA UNIVERSITY, 2011d. Session 4 : Data For Life - Digital Preservation for Health Sciences [en ligne]. Northumbria University. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://find.jorum.ac.uk/resources/18276

NORTHUMBRIA UNIVERSITY, 2011e. Overview of the Training Programme [en ligne]. Northumbria University. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://find.jorum.ac.uk/resources/18276

NORTHUMBRIA UNIVERSITY, 2013. DATUM for Health: Research data management training for health studies. [en ligne]. Northumbria University. 2013. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://www.webarchive.org.uk/wayback/archive/20140614071433/http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain/datum.aspx

PRONGUÉ, Nicolas, 2015. Données : lieu de publication. Site Internet d'Educaplay [en ligne]. 22.12.2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://fr.educaplay.com/fr/activiteeducatives/2224638/donnees__lieu_de_publication.htm

PRONGUÉ, Nicolas, 2016. Données: formes de publication. Site Internet d'Educaplay [en ligne]. 26.01.2016. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://fr.educaplay.com/fr/activiteeducatives/2244810/donnees__formes_de_publication.htm

SHERIDAN, Matt, et al., 2015. New England Collaborative Data Management Curriculum - Module 6 Data Sharing & Reuse Policies [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/modules

SHOTTON, David, 2012. Twenty Questions for Research Data Management [en ligne]. University of Oxford. 2012. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://blogs.it.ox.ac.uk/acit-rs-team/files/2014/02/Twenty_Questions_for_Research_Data_Management.pdf

UNIVERSITY OF EAST LONDON, [sans date]. Research Data Management Policy for UEL [en ligne]. University of East London. [Consulté le 4 juillet 2016]. Disponible à l'adresse : https://www.uel.ac.uk/wwwmedia/services/library/lls/resources/rspresearchtools/Research-Data-Management-policy-for-UEL-FINAL.pdf

UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL, 2011. New England Collaborative Data Management Curriculum [en ligne]. University of Massachusetts Medical School. 2011. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/index

UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL, 2015. New England Collaborative Data Management Curriculum - How to Teach RDM using Cases [en ligne]. University of Massachusetts Medical School. 2015. [Consulté le 4 juillet 2016]. Disponible à l'adresse : http://library.umassmed.edu/necdmc/how_to_teach

Formation proposée :

BAKER, Monya, 2016. 1,500 scientists lift the lid on reproducibility. Nature. 25 mai 2016. Vol. 533, n° 7604, pp. 452-454.

CREATIVE COMMONS FRANCE, [sans date]. 6 LICENCES gratuites. Site Internet de Creative Commons France [en ligne]. [Consulté le 5 juillet 2016]. Disponible à l'adresse : http://creativecommons.fr/licences/

DATACITE, 2015. DataCite Metadata Schema for the Publication and Citation of Research Data [en ligne]. DataCite - International Data Citation. 2015. [Consulté le 5 juillet 2016]. Disponible à l'adresse : https://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf

DRYAD, 2016. Dryad. Site Internet du dépôt Dryad [en ligne]. 07.07.2016. [Consulté le 7 juillet 2016]. Disponible à l'adresse : http://datadryad.org/

FIGSHARE, [sans date]. Figshare. Site Internet du dépôt Figshare [en ligne]. [Consulté le 5 juillet 2016]. Disponible à l'adresse : https://figshare.com/

OPEN SCIENCE FRAMEWORK, 2011. Open Science Framework. Site Internet du Dépôt Open Science Framework [en ligne]. 2011-2016. [Consulté le 5 juillet 2016]. Disponible à l'adresse : https://osf.io/

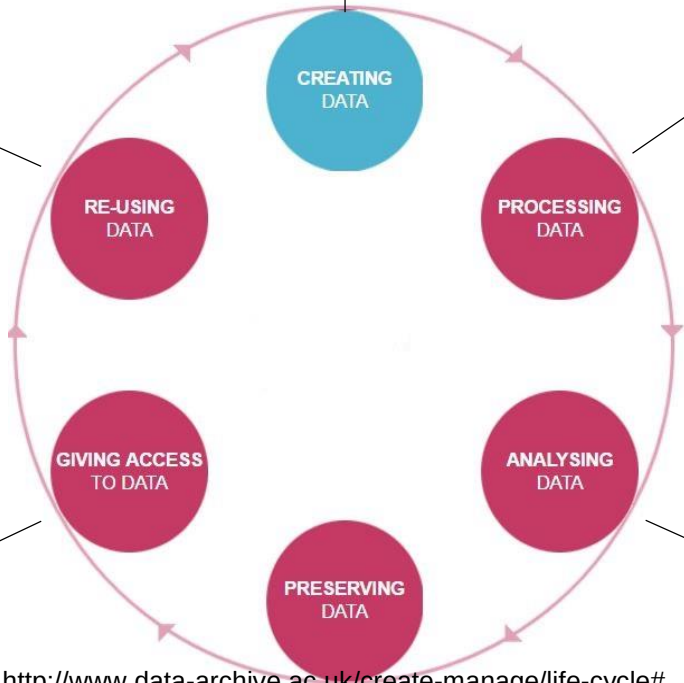
PLOS, [sans date]. PLOS ONE. Site Internet de PLOS ONE [en ligne]. [Consulté le 6 juillet 2016]. Disponible à l'adresse : http://journals.plos.org/plosone/s/data-availability

SWISS UNIVERSITIES, 2015. Data life-cycle management Project. Site Internet du DLCM [en ligne]. 01.09.2015. [Consulté le 5 juillet 2016]. Disponible à l'adresse : http://dlcm.ch/

ZENODO, [sans date]. Zenodo. Site Internet du dépôt Zenodo [en ligne]. [Consulté le 5 juillet 2016]. Disponible à l'adresse : https://zenodo.org

Annexe 1 : Exercice Cycle de vie

Citez des actions que vous réaliserez à chaque étape du cycle de vie



http://www.data-archive.ac.uk/create-manage/life-cycle#

Annexe 2 : Questionnaire concernant la gestion des données de recherche[94]

Ce questionnaire comprend une série de 13 questions sur le déroulement de vos recherches et la gestion des données que vous créerez. Son but est de vous aider à remplir les différentes sections du Data Management Plan qui accompagne ce document.

1) La nature de vos données

À qui appartiennent les données créées dans le cadre de vos recherches ?

Par quel(s) domaine(s) de recherche (sujet, discipline) vos données seront-elles concernées ?

Quels sont les types de données (observationnelles, statistiques, images) que vous pensez créer ?

Quelles seront les méthodes de création de ces données ? allez-vous réutiliser d'autres sets de données ?

2) Stockage, accès et sécurité des données

À quelle fréquence pensez-vous sauvegarder (versioning) vos données ?

[94] Ce questionnaire a été réalisé à l'aide de la checklist créée par David Shotton pour l'Université d'Oxford, accessible à l'adresse http://blogs.it.ox.ac.uk/acit-rs-team/files/2014/02/Twenty_Questions_for_Research_Data_Management.pdf ainsi que la checklist créée conjointement par l'EPFL et l'ETH accessible à l'adresse http://library.epfl.ch/files/content/sites/library3/files/research-data/dmp/Data_management_plan_checklist_EPFL_2016.pdf

Qui est responsable de la gestion immédiate, le stockage au jour le jour et de la sauvegarde des données résultant de vos recherches ?

3) Métadonnées – les données sur les données

Quelles informations accompagnant vos données seront d'après vous nécessaires à leur compréhension ?

4) Partage et réutilisation des données

Y aura-t-il un embargo (restriction de diffusion durant une période donnée) sur vos données ? et si oui, de quelle durée ?

Pensez-vous partager la totalité de vos données ou y aura-t-il une quelconque restriction à l'Open Access ?

Quel est la politique de votre institution concernant le partage des données ?

5) Anonymisation des données

Aurez-vous des données sensibles et/ou confidentielles à traiter ?

6) Archivage des données

Connaissez-vous la politique de votre institution en termes de dépôt à long terme ?

Où pensez-vous déposer vos données une fois que les résultats auront été publiés ?

Annexe 3 : Contraintes et bénéfices au partage des données de recherche[95]

Listez les contraintes et vos inquiétudes vis-à-vis du partage des données de recherche	Listez les bénéfices résultants selon vous du partage des données de recherche

[95] Cet exercice est basé sur le cours *DIY Research Data MANTRA Training Kit for Librarians* de *EDINA and Data Library, University of Edinburgh* accessible à l'adresse http://datalib.edina.ac.uk/mantra/libtraining.html ainsi que le cours *Data sharing* de l'*University of East London*

Annexe 4 : Questionnaire sur les métadonnées

Imaginez-vous devant un set de données que vous ne connaissez pas, mais que vous voudriez réutiliser afin de réaliser une expérience. Quelles sont, d'après vous, les métadonnées que vous aimeriez avoir et pourquoi ?

-

-

-

-

-

-

-

-

-

-

Inversement, si d'aventure un collègue voulait réutiliser les données que vous êtes en train de créer, quelles seraient les informations complémentaires qui lui seraient nécessaires afin de reproduire votre expérience ?

-

-

-

-

-

-

-

-

-

-

-

Annexe 5 : Schéma listant les métadonnées à incorporer

Ce schéma, présent sur le site de la BIUM[96], est repris de *DataCite Metadata Schema*[97]. Il y est listé 18 types de métadonnées à incorporer selon trois niveaux d'obligation de présence :

- Mandatory (M) → Obligatoire
- Recommended (R) → Recommandé
- Optional (O) → Optionnel

ID	DataCite-Property	Obligation	Definition
1	Identifier	M	The Identifier is a unique string that identifies a resource (DOI, ...).
2	Creator (=Author) -name identifier -affiliation	M	The main researchers involved in producing the data, or the authors of the publication, in priority order. - EX: Smith, John; Miller, Elizabeth.... - ORCID - Affiliation
3	Title	M	A name or title by which a resource is known.
4	"Publisher" (=data repository)	M	The name of the repository where the data are archived. This property will be used to formulate the citation, so consider the prominence of the role. EX: Figshare, Zenodo or Dryad
5	"Publication Year" for the dataset	M	The year when the data was or will be made publicly available on the repository (embargo period).
ID	**DataCite-Property**	**Obligation**	**Definition**
6	Subject	R	Key words or phrase describing the resource.
7	Contributor (with type, name identifier, and affiliation sub-properties)	R	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource. Ex: contact person, funder, hosting institution, data collector= the person collectin the data under the PI's guideline,...)
8	Date	R	Different dates relevant to the work (ex: data of the article publication).
9	Language	O	The primary language of the resource.
10	ResourceType	R	A description of the resource type (ex: audiovisual, image, text, software,...)
11	Alternate Identifier	O	An identifier or identifiers other than the primary Identifier applied to the resource being registered.
12	RelatedIdentifier	R	Identifiers of related resources. These must be globally unique identifiers. Ex: article realted to the dataset DOI, PMID, Wos number,...)
13	Size	O	Unstructured size information about the resource (ex: page number, Mega Bytes....).
14	Format	O	Technical format of the resource (ex, PDF/A, Tiff, MPEG4,...).
15	Version	O	The version number of the resource.
16	Rights	O	Any rights information for this resource. License: CC BY creative commons and link to the license
17	Description	R	All additional information that does not fit in any of the other categories. May be used for technical information.
18	GeoLocation	R	Spatial region or named place where the data was gathered or about which the data is focused.

[96] http://www.bium.ch/en/publication-open-access/data-management/
[97] http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf

Annexe 6 : Formats des données de recherche[98]

Type de données	Formats recommandés	Formats acceptés mais non conseillés
Tabular data with extensive metadata	• SPSS portable format (.por) • Structured text or mark-up file of metadata information, e.g. DDI XML file	• SPSS (.sav) • Stata (.dta) • MS Access (.mdb/.accdb)
Tabular data with minimal metadata	• Comma-separated values (.csv) • Tab-delimited file (.tab)	• Delimited text (.txt) • Ms Excel (.xls/.xlsx) • Ms Access (.mdb/.accdb) • Dbase (.dbf), • Opendocument spreadsheet (.ods)
Geospatial data Vector and raster data	• ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) • Geo-referenced TIFF (.tif, .tfw) • CAD data (.dwg) • Tabular GIS attribute data • Geography Markup Language (.gml)	• ESRI Geodatabase format (.mdb) • Mapinfo Interchange Format (.mif) for vector data • Keyhole Mark-up Language (.kml) • Adobe Illustrator (.ai), CAD data (.dxf or .svg) • Binary formats of GIS and CAD packages
Textual data	• Rich Text Format (.rtf) • Plain text, ASCII (.txt) • Extensible Mark-up Language (.xml)	• Hypertext Mark-up Language (.html) • MS Word (.doc/.docx) • OpenDocument text (.odt)
Image data	• TIFF (.tif)	• JPEG (.jpeg, .jpg) • GIF (.gif) • RAW image format (.raw) • Photoshop files (.psd) • BMP (.bmp) • PNG (.png) • Adobe Portable Document

[98] Ce tableau a été réalisé à partir du tableau du site *ukdataservice* accessible à la page https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats , le sixième module de la formation de l'University of East of London ainsi que de l'Université d'Edinburgh http://www.ed.ac.uk/files/atoms/files/recommended_file_formats-apr2015.pdf

		Format (PDF/A, PDF) (.pdf)
Audio data	• Free Lossless Audio Codec (FLAC) (.flac)	• MPEG-1 Audio Layer 3 (.mp3) • Audio Interchange File Format (.aif) • Waveform Audio Format (.wav)
Video data	• MPEG-4 (.mp4) • OGG video (.ogv, .ogg) • Motion JPEG 2000 (.mj2)	• AVI (.avi)
Documentation and scripts	• Rich Text Format (.rtf) • PDF/UA, PDF/A or PDF (.pdf) • XHTML or HTML (.xhtml, .htm) • Opendocument Text (.odt)	• Plain text (.txt) • Ms word (.doc/.docx), • Ms excel (.xls/.xlsx) • Xml marked-up text (.xml)

Annexe 7 : Tableau récapitulatif des licences CC[99]

Terme abrégé	Symboles	Droits et obligations
CC-Zero		Le créateur renonce à ses droits. Aucune restriction d'exploitation et aucune nécessité à citer le nom de l'auteur
CC-BY Attribution		Aucune restriction quant à l'exploitation ou la création d'œuvres dérivées à condition de citer le nom de l'auteur
CC-BY-SA Attribution + Partage dans les mêmes conditions		Exploitation de l'œuvre uniquement sous une licence identique. Nécessité de citer le nom de l'auteur
CC-BY-ND Attribution + Pas de Modification		Exploitation autorisée mais pas de création d'œuvres dérivées. Nécessité de citer le nom de l'auteur
CC-BY-NC Attribution + Pas d'Utilisation Commerciale		Exploitation autorisée tout comme la création d'œuvres dérivées à condition qu'il ne s'agisse pas d'œuvres commerciales. Nécessité de citer le nom de l'auteur
CC-BY-NC-SA Attribution + Pas d'Utilisation Commerciale + Partage dans les mêmes conditions		Exploitation pour des œuvres non commerciales autorisée tout comme la création d'œuvres dérivées mais uniquement sous une licence identique. Nécessité de citer le nom de l'auteur
CC-BY-NC-ND Attribution + Pas d'Utilisation Commerciale + Pas de Modification		Exploitation pour des œuvres non commerciales autorisée mais pas de création d'œuvres dérivées. Nécessité de citer le nom de l'auteur

[99] Ce tableau a été réalisé à partir des informations du site http://creativecommons.fr/licences/ ainsi que la page https://fr.wikipedia.org/wiki/Licence_Creative_Commons pour les symboles

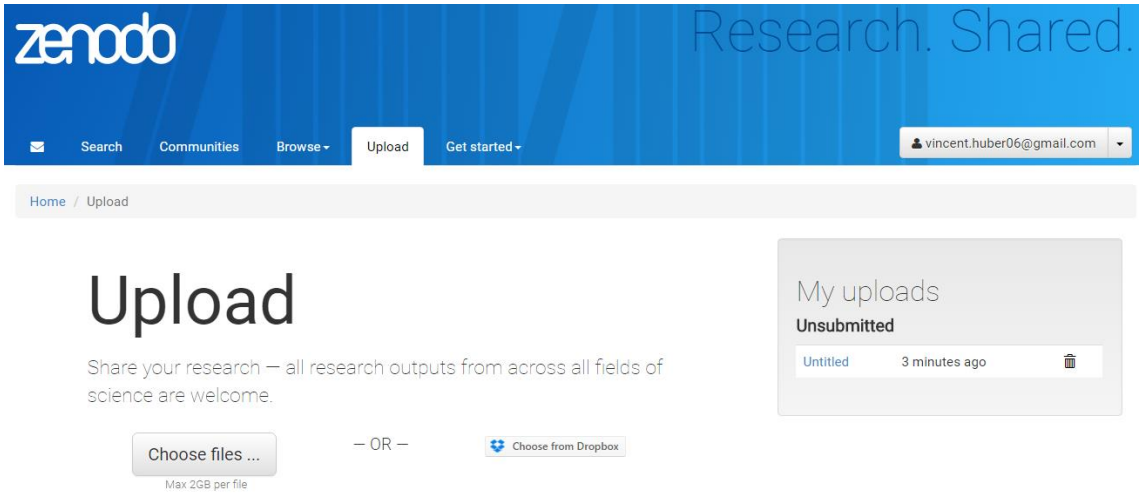
Annexe 8 : Exercice dépôts en ligne

Durant cet exercice, nous allons nous entraîner à utiliser divers dépôts en ligne qui vous seront utiles dans l'optique de sauvegarder vos données à long terme. Pour cela, nous allons déposer vos sets de données, les traiter et gérer les droits mais aussi réaliser des recherches et télécharger des sets de données qui hypothétiquement pourraient vous intéresser.

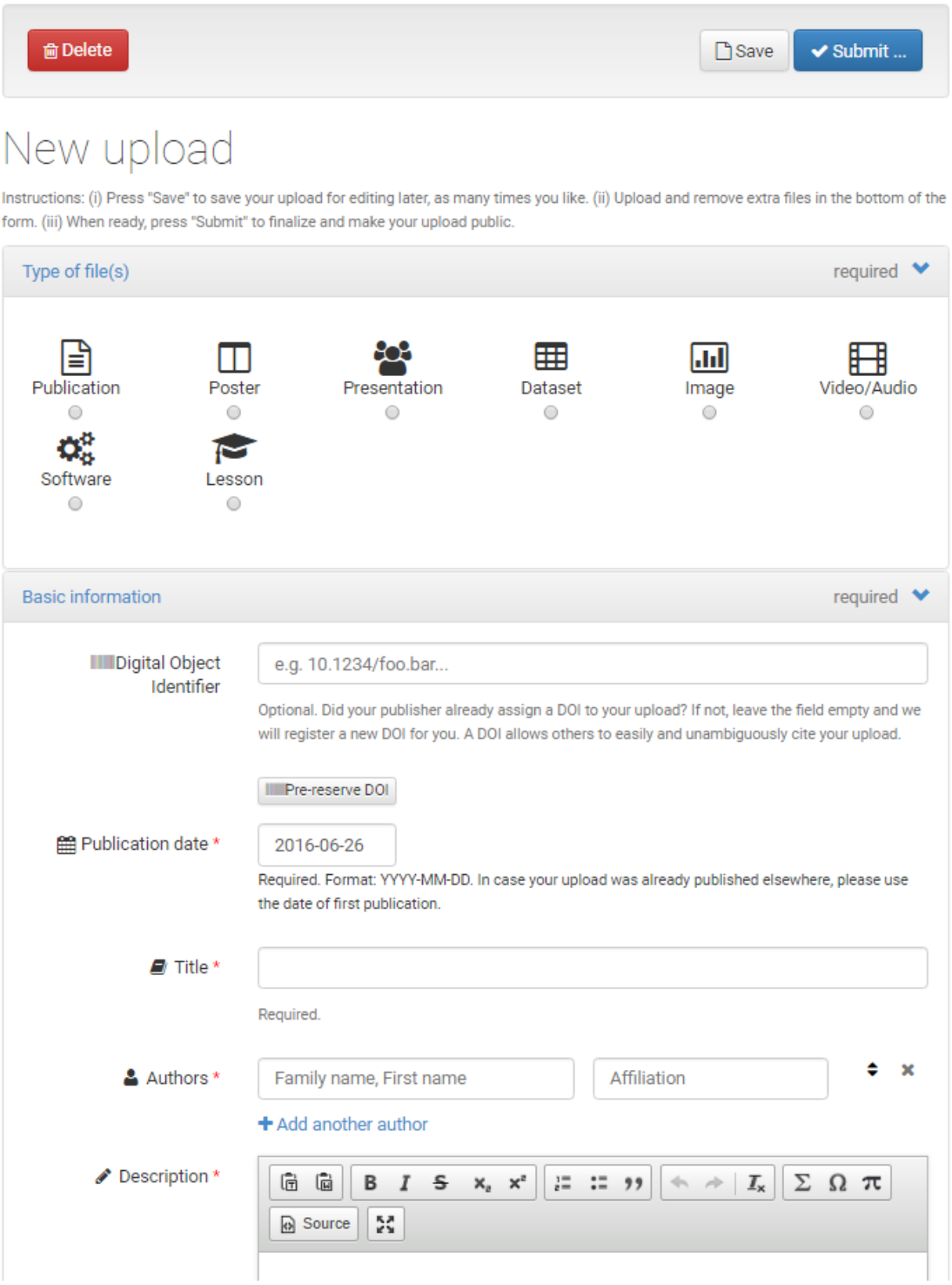
Zenodo

1) Allez sur https://zenodo.org/ et enregistrez-vous comme utilisateur. Vous pouvez, à cet effet, vous inscrire à l'aide de votre identifiant ORCID permettant ainsi de jumeler vos profils.

2) À partir de la page d'accueil, cliquez sur l'onglet **Upload** afin de déposer un set de vos données. Vous pouvez choisir les fichiers individuellement et pouvez en prendre plusieurs à la fois.



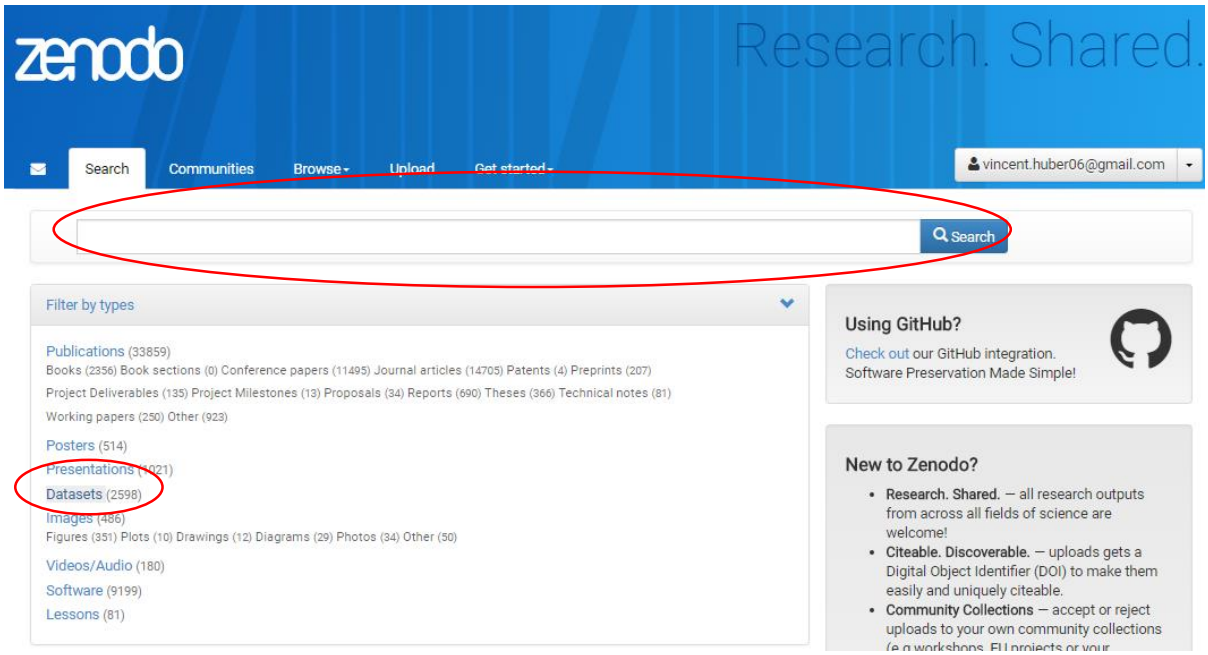
3) Une fois que vos fichiers ont été uploadés, vous arrivez sur une page **d'édition de votre set de données.** Il est donc important que vous remplissiez les différents champs le plus précisément possible afin que vous-mêmes mais aussi les autres utilisateurs puissent en comprendre le contenu. Vous pouvez choisir les mots-clés accompagnant votre set mais aussi le type de licence de manière précise.



À noter que le bouton *save* sauvegarde vos données mais il vous sera redemandé plus tard d'éditer le set. Si d'aventure vous désirez uploader définitivement votre set, cliquez sur Submit **MAIS ATTENTION, Zenodo publiera alors votre set et vous ne pourrez plus l'effacer !**

4) À présent, nous allons réaliser une recherche dans l'optique de trouver un set de données à télécharger. Cliquez sur l'onglet **Search**. Vous pouvez ainsi réaliser une **recherche dans la barre à cet effet** en inscrivant un terme médical qui

décrirait efficacement ce que vous cherchez. Néanmoins, vous pouvez aussi trier le contenu du site par type de document souhaité. Par exemple, cliquez sur **Datasets.**



Vous avez changé d'onglet et êtes sous l'onglet **Browse.** Dorénavant, lorsque vous tapez un mot-clé dans la barre de recherche, vous ne tomberez alors que sur les sets de données relatifs à ce mot. Si d'aventure vous cherchez un autre type de fichiers, cliquer sur **Browse** et choisissez le type de votre choix.

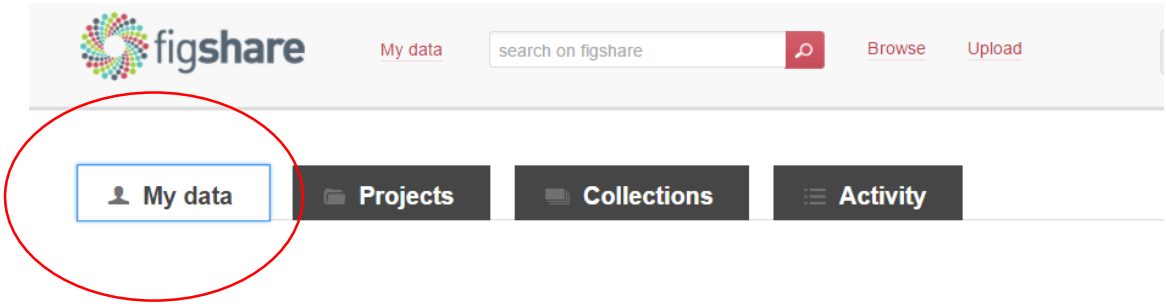
5) Recherchez un terme qui vous semble pertinent puis, choisissez un set de données proposé qui est en Open Access en cliquant dessus. À noter que vous pouvez trier vos résultats par date de publication, par auteur, etc.

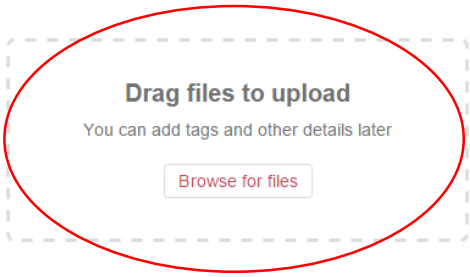
Vous arrivez alors sur la page du set de données avec son descriptif et parfois la publication qui l'accompagne. **Les fichiers se trouvent généralement au bas de la page.** Pour télécharger le fichier qui vous intéresse, cliquez sur **download.**

Figshare

1) Allez sur https://figshare.com/ et enregistrez-vous comme utilisateur avec votre adresse email.

2) Une fois connecté vous arrivez sur la page de votre profil, comprenant 4 onglets que sont **My data, Projects, Collections et Activity.** Pour uploader votre set de données, mettez-vous dans l'onglet **My data** et cliquez sur **Browse for files** au milieu de votre écran.



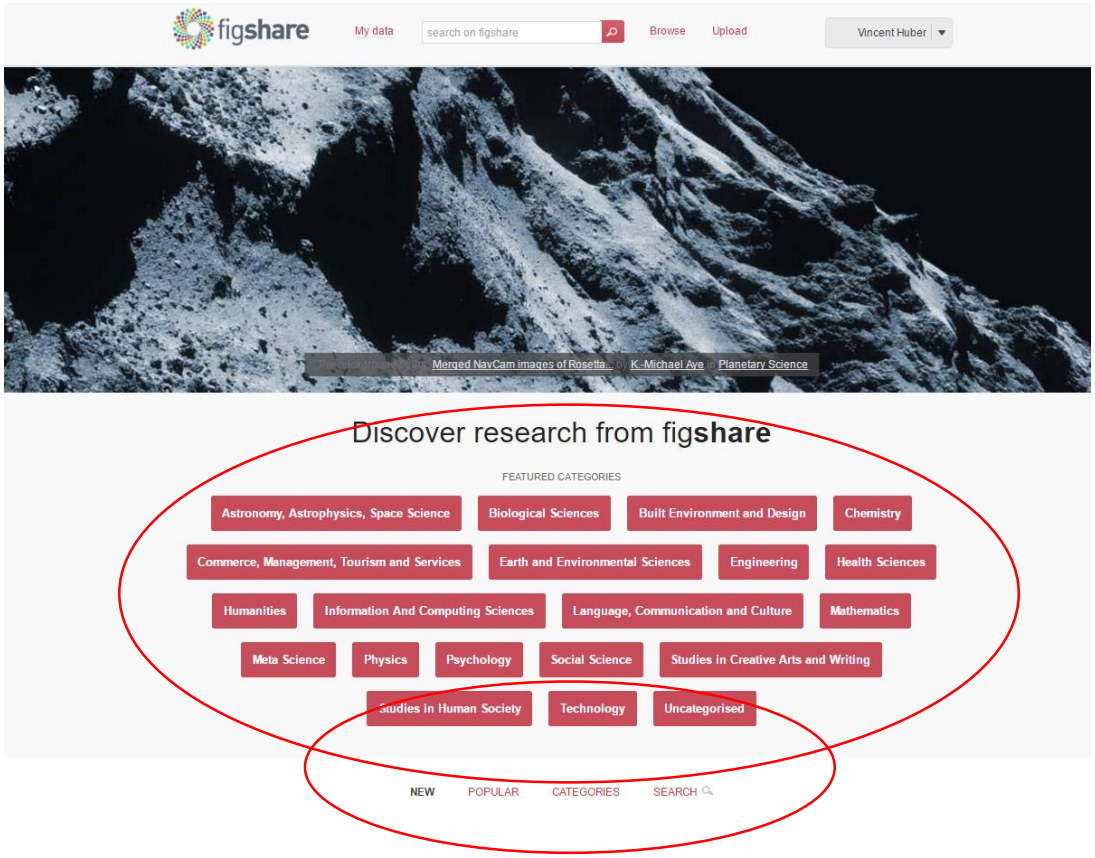


3) Une fois que vos fichiers sont uploadés, une fenêtre d'édition s'ouvre comme pour Zenodo. Remplissez alors les divers champs comme précédemment afin de réunir le plus informations importantes possibles : catégories, mots-clés (tags), description, etc.

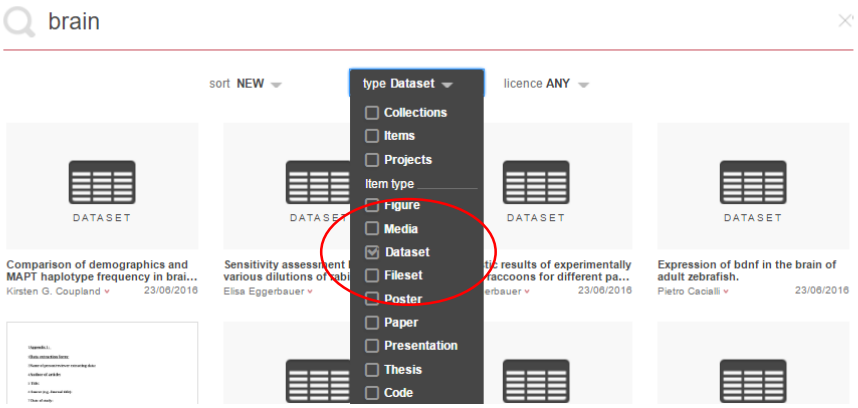
À noter que le nombre de licences CC proposées est ici moins nombreux. Vos options sont donc plus limitées.

ATTENTION, comme pour Zenodo, ne soumettez pas votre set si d'aventure il n'est pas complet car il serait publié !

4) Pour la recherche, vous avez ici aussi, en haut de la page, une barre dans laquelle vous pouvez rechercher le terme de votre choix. Mais si vous cliquez sur l'onglet **Browse**, à côté de la barre de recherche, le site vous propose des thématiques déjà présélectionnées ainsi qu'une recherche plus précise, notamment par **catégorie** mais aussi par **type de fichiers.**



5) Cliquez sur **Search** en dessous des catégories et tapez le terme de recherche de votre choix. Afin de limiter votre recherche aux sets de données, **choisissez *Datasets*** dans les types de données, juste en dessous de votre barre de recherche.



6) Choisissez alors l’un des sets de données qui semble vous convenir en cliquant dessus. Une fois arrivez sur la page du set de données, vous pourrez lire la description de celui-ci et le télécharger en cliquant sur **Download.**