h e g

Haute école de gestion
Genève

Préservation des données de recherche : proposer des services de soutien aux chercheurs du site Uni Arve de l'Université de Genève

Travail de master réalisé par :

Manuela BEZZI

Sous la direction de :

Alain Dubois, Archiviste cantonal du Valais

Nyon, le 17 août 2020

Information Documentaire

Haute École de Gestion de Genève (HEG-GE)

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Remerciements

Je tiens à remercier chaleureusement toutes les personnes qui m'ont apporté leur aide pour la réalisation de ce travail :

Alain Dubois, mon directeur de travail de master, pour sa disponibilité, son soutien et son implication ;

Annabel Chanteraud, responsable du site Uni Arve de la Bibliothèque de l'Université de Genève, pour m'avoir confié ce mandat et pour son soutien et ses conseils ;

Audrey Bellier, bibliothécaire « Spécialiste Recherche » et experte des questions données de recherche et Open Access, pour avoir proposé cette problématique et avoir suivi ce travail en tant que répondante, ainsi que pour nos échanges fructueux et son implication ;

Lydie Echernier, coordinatrice du projet DLCM, pour son expertise sur Yareta et pour son implication ;

Les personnes rencontrées pour les entretiens pour leur disponibilité et la qualité des informations transmises ;

Marielle Guirlet pour tous les bons moments partagés durant ce master et pour nos discussions, échanges et encouragement durant cette dernière ligne droite ;

Mes ami(e)s, collègues et famille pour leur soutien et leurs encouragements durant ces années d'étude.

Résumé

Ce travail porte sur les pratiques des chercheurs du site Uni Arve (faculté des sciences) de l'Université de Genève concernant la préservation et la réutilisation des données de recherche, et son objectif est d'évaluer les besoins des chercheurs afin de leur proposer des services de soutien appropriés.

La préservation des données de recherche s'inscrit dans le mouvement de l'Open Data dont l'objectif est de rendre les données de recherche publiquement accessibles, intelligibles et réutilisables, en particulier lorsque ces données ont été produites grâce à des recherches financées par des fonds publics. Pour ce faire, le FNS demande aux chercheurs de déposer leurs données dans des archives publiques répondant aux principes FAIR. Or, depuis juin 2019, l'Université de Genève met à disposition de ses chercheurs une archive institutionnelle, Yareta, répondant aux critères du FNS.

Afin de répondre aux mieux aux besoins des chercheurs, une approche en deux temps a été adoptée : (1) une analyse des jeux de données déposés sur Yareta a permis d'identifier les problématiques faisant obstacle à la réutilisation des données. (2) Puis, des entretiens menés avec des chercheurs ont permis d'analyser leurs pratiques de préservation et leurs besoins.

Les informations récoltées par ces deux approches ont permis de faire les propositions suivantes:

- un guide d'archivage portant sur quatre activités permettant de garantir une bonne préservation : format, contexte, métadonnées, licence
- la mise en place de ressources additionnelles (page web ou formation) couvrant des notions peu comprises par les chercheurs
- la modification de pages web existantes pour des raisons de cohérence
- l'ajout d'information dans l'outil Yareta

Ces propositions sont des solutions concrètes, basées sur les ressources existantes de l'Université de Genève afin de pouvoir être complémentaires aux services de soutien et aux ressources déjà proposés par l'Université de Genève. De plus, ces propositions pourront bénéficier à toute la communauté de l'Université de Genève et pas uniquement aux chercheurs du site Uni Arve.

Mots-clés : Open Access, Open Data, préservation, réutilisation, donnée de recherche, dépôt institutionnel, soutien à la recherche, Yareta, Université de Genève.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ANDS	Australian National Data Service
CC	Creative Commons
DCC	Digital Curation Center
DCMES	Dublin Core Metadata Element Set
DLCM	Data Life Cycle Management
DMP	Data Management Plan
DOI	Digital Object Identifier
ELN	Electronic Laboratory NotebooK
EPFL	Ecole Polytechnique Fédérale de Lausanne
ETHZ	Eidgenössische Technische Hochschule Zürich ou Ecole Polytechnique Fédérale de Zurich
FAIR	Findable, Accessible, Interoperable, Reusable
FITS	File Information Tool Set
FNS	Fonds National Suisse de la Recherche Scientifique
GB	Gigabyte
GDR	Gestion des Données de Recherche
METS	Metadata Encoding and Transmission Standard
OAI-ORE	Open Archives Initatives Object Reuse and Exchange
OAIS	Open Archival Information System
OCDE	Organisation de Coopération et de Développement Economiques
ODS	Open Data Commons

1. Présentation du mandat

Ce travail de master, mandaté par la Bibliothèque de l'Université de Genève, porte sur les pratiques des chercheurs concernant la préservation des données de recherche et se base sur les constats suivants :

- En remplissant un plan de gestion des données (Data Management Plan ou DMP), le chercheur s'engage à préserver et à rendre accessibles les données sur lesquelles se base sa publication (FNS [sans date]b)
- Ces DMP étant obligatoires depuis octobre 2017 (FNS 2017), il est probable que certains chercheurs termineront prochainement leurs projets de recherche
- Les services de soutien proposés par l'Université de Genève sont pour l'instant principalement axés sur la rédaction du DMP et la diffusion des données, mais il existe peu de soutien concernant les étapes précédents le dépôt des données

L'Université de Genève envisage donc de développer des services de soutien en lien avec la préservation des données de recherche, et veut cibler au mieux les besoins des chercheurs afin de leur proposer une offre de soutien la plus adéquate possible.

Pour des raisons de faisabilité et bien qu'au départ le mandat concernait l'Université de Genève, ce travail a été restreint au site de Uni Arve (faculté des sciences). En outre, plus que la notion de préservation et d'accessibilité, la problématique s'est orientée sur les possibilités de réutilisation des données déposées dans le dépôt institutionnel de l'Université de Genève, et ce afin de vérifier le postulat de base, à savoir que les données déposées dans un dépôt de préservation sont réutilisables.

D'entente avec la mandante, l'objectif de ce travail consiste à :

- Définir ce qui permet à une donnée d'être réutilisée
- Evaluer les jeux de données déposés dans l'archive institutionnelle Yareta d'un point de vue de la réutilisation
    - Élaborer une grille d'analyse
    - Analyser les jeux de données en fonction de cette grille d'analyse
- S'entretenir avec des chercheurs ayant utilisé le dépôt institutionnel de l'Université de Genève afin de comprendre leurs pratiques
- En fonction de l'analyse des dépôts et des entretiens, évaluer les besoins des chercheurs et leur proposer des services de soutien appropriés

Dans ce document, nous présentons les différentes étapes adoptées pour permettre la réalisation de ce mandat ainsi que les résultats obtenus.

2. Problématique

Afin de pouvoir élaborer des livrables adaptés aux besoins des chercheurs de l'Université de Genève, nous nous sommes appuyés sur l'existant, à savoir les jeux de données déposés dans Yareta et les pratiques des chercheurs pour déposer ces jeux de données.

Ce travail s'est fait en trois étapes principales :

1. Une revue de la littérature
2. Une analyse de l'existant – 49 jeux de données déposés sur Yareta
3. Des entretiens avec les chercheurs – sept entretiens mené du 20 mai au 18 juin 2020

Les propositions de services de soutien se présentent sous la forme de deux livrables :

- Un guide d'archivage sur la qualité des données
- Des propositions de formation à l'intention du chercheur

Ce travail se divise en cinq parties et présente tout d'abord la gestion des données de recherche et le contexte dans lequel elle s'inscrit (Open Access / Open Data), et nous nous concentrerons ensuite sur les notions de préservation et de réutilisation des données de recherche (section 3). Puis, nous nous intéresserons aux dépôts de préservation, et nous présenterons en particulier Yareta, le dépôt de préservation institutionnel mis en place par l'Université de Genève. Dans cette partie, nous regarderons également le type de soutien proposé aux chercheurs et nous présenterons plus en détails la structure d'accompagnement mis en place par l'université de Delft (section 4). Grâce aux informations présentées dans les parties « Préservation » et « Réutilisation » de la section 3, nous élaborerons une grille d'analyse, nous analyserons les jeux de données déposés sur Yareta, et nous ferons une synthèse des problématiques rencontrées dans les jeux de données, complétée pour chaque problématique d'une proposition de soutien pour y remédier (section 5). Des entretiens seront menés avec des groupes de recherche affiliés à la faculté de sciences (Uni Arve) et portant sur leurs pratiques de préservation. De ces entretiens, nous ferons une synthèse des pratiques des répondants et de leurs problématiques, accompagnée de proposition de soutien (section 6). Finalement, basé sur les synthèses présentées dans les sections 5 et 6, nous produirons les deux livrable : le guide d'archivage et les propositions de formation (section 7).

2.1 Revue de littérature

La revue de littérature s'est faite principalement dans le catalogue de bibliothèque RERO, les bases de données LISA, LISTA et Emerald Insight, ainsi que dans les moteurs de recherche Google et Google Scholar.

Les sites web d'institutions suisses et internationales ont été consultés pour la revue des ressources mises à disposition des chercheurs.

2.2 Périmètre de ce travail

Tous les dépôts de Yareta ont été analysés indépendamment de leur unité organisationnelle, et nous avions aussi bien des dépôts provenant de la faculté des sciences que des dépôts provenant de la faculté de médecine ou de la haute école de santé. Cependant, pour les entretiens, nous nous sommes limités à la faculté des sciences pour des raisons de faisabilité.

Ce travail ne s'intéresse qu'aux activités concernant la préservation à long terme des données de recherche et la gestion des données actives de recherche (*active data management*) est hors du périmètre de ce travail.

2.3 Limitation de ce travail

Dans la liste des chercheurs retenus pour participer aux entretiens, nous n'avons eu que des cas de dépôts liés à une publication. Les pratiques de sélection des données de recherche n'ont de ce fait pas pu être analysées, puisque la sélection se faisait automatiquement par le biais de la publication : le chercheur déposait les données présentées dans sa publication.

En raison du confinement, je n'ai pu utiliser que mon ordinateur personnel pour analyser les jeux de données, et cette analyse a été limitée par la performance de mon matériel informatique : impossibilité de télécharger des jeux de données dont la taille dépassait 1GB, impossibilité d'analyser certains documents de métadonnées.

3. Gestion des données de recherche

Une bonne gestion des données de recherche (GDR) est nécessaire pour pouvoir faire de l'Open Access ou de l'Open Data (Wilkinson et al. 2016). Dans cette section, nous allons définir ce qu'est l'Open Access et l'Open Data, puis nous allons nous intéresser aux données de recherche.

3.1 Open Access / Open Data

La préservation des données de recherche s'inscrit dans le mouvement de l'Open Access et de l'Open Data. Ce mouvement, né suite à l'émergence d'internet (digitalisation) et l'augmentation des coûts des éditeurs, promeut aussi bien l'accès libre aux publications (Open Access) qu'aux résultats de la recherche (Open Data) (Delft University of Technology 2020b ; Budapest Open Access Initiative 2002 ; Berlin declaration 2003 ; OCDE 2007). Le but de cette ouverture des données est de les rendre publiquement accessibles, intelligibles, et réutilisables, en particulier lorsque ces données ont été produites grâce à des recherches financées par des fonds publics (OCDE 2007 ; Gaillard 2014 ; FNS [sans date]e).

En Suisse, l'Open Access est soutenu aussi bien par le FNS (principal bailleur de fonds des chercheurs) que par les hautes écoles suisses, représentées par Swissuniversities, et il existe un plan national sur l'Open Access dont l'objectif est d'avoir, d'ici 2024, toutes les publications scientifiques bénéficiant de fonds publics accessibles librement et gratuitement (FNS [sans date]c ; Swissuniversities [sans date]a).

L'Open Data est également soutenu par le FNS et par Swissuniversities, et le FNS s'engage afin que toute les recherches financées par des fonds publics soient accessibles publiquement et gratuitement, et le chercheur bénéficiant d'un fond du FNS a l'obligation de publier le résultat de ses recherches en Open Access (FNS [sans date]e ; FNS [sans date]d ; Swissuniversities [sans date]b). Cet engagement du FNS s'est concrétisé depuis 2017 par l'obligation pour le chercheur de fournir un DMP dans sa demande de financement, ainsi que l'obligation d'archiver les données dans des dépôts en libre accès (FNS [sans date]c) :

« Par conséquent, le FNS demande à tous les chercheuses et chercheurs qu'il finance :

- *d'archiver les données de recherche sur lesquelles ils ont travaillé et qu'ils ont produites durant leurs travaux,*
- *de partager ces données avec d'autres chercheuses et chercheurs, à moins qu'ils/elles soient lié-e-s par des clauses légales, éthiques, de copyright, de confidentialité ou autres, et*
- *de déposer leurs données et métadonnées dans des archives publiques existantes, dans des formats accessibles et réutilisables sans restriction par tout un chacun*. »

Plusieurs freins sont évoqués par les chercheurs pour expliquer leurs réticences à partager leurs données. Parmi ceux-ci, on peut mentionner des inquiétudes quant aux futures possibilités de publication et à une mauvaise utilisation des données, trop d'efforts nécessaires pour partager les données, une inquiétude concernant les aspects légaux ou juridiques, un manque de temps et de financement, un manque de formation, un manque d'encouragement et de récompense (Tenopir et al. 2011 ; Tenopir et al. 2015 ; Van den Eynden et al. 2016 ; Plomp et al. 2019 ; Académie des sciences naturelles 2018).

Cependant, il y a également des avantages à pratiquer l'ouverture des données: l'augmentation de la visibilité et de l'impact du travail du chercheur, l'augmentation de son taux de citation, l'opportunité de nouvelles collaborations, la reproductibilité ainsi que la transparence et l'intégrité de la recherche scientifique, la réutilisation des données, la conformité aux exigences des bailleurs de fonds et des éditeurs (Van den Eynden 2011 ; Swissuniversities [sans date]b ; Gaillard 2014 ; Delft University of Technology 2020c ; McKiernan et al. 2016).

3.2 Données de recherche

Il existe plusieurs définitions des données de recherche, mais celle que l'on retrouve le plus souvent est celle de l'OCDE : « *Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche*» (OCDE 2007). C'est également cette définition que l'on retrouve sur les pages web de l'Université de Genève dédiées aux données de recherche ainsi que dans sa politique institutionnelle sur la GDR (Université de Genève [sans date]a ; Université de Genève 2018).

l'Australian National Data Service (ANDS) propose une autre définition qui élargit les données de recherche à tout données produites lors d'une recherche: « *Research data means data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media."* (Australian National Data Service (ANDS), 2017b).

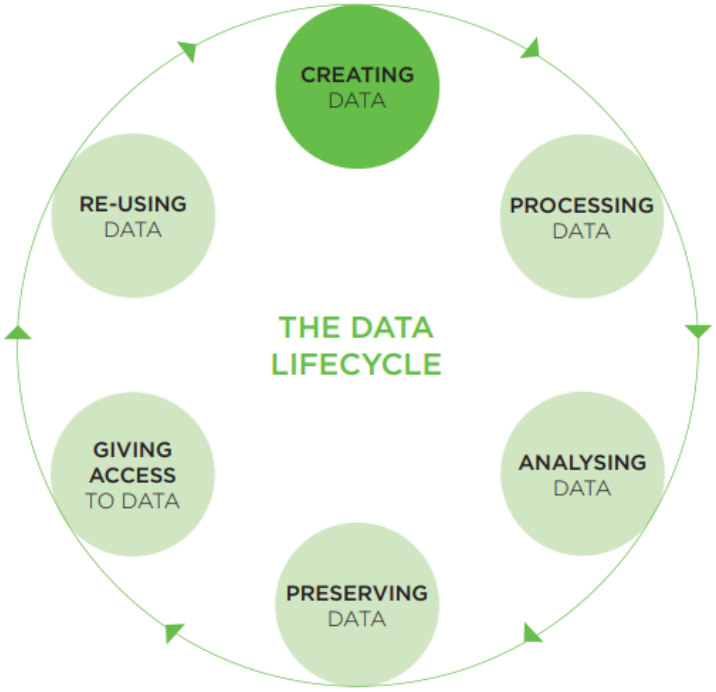
La définition de l'ANDS concerne donc toutes les données, y compris les données actives. Or, les données actives étant hors du périmètre de ce travail, la définition de l'OCDE convient tout à fait à ce travail.

3.2.1 Le cycle de vie d'une donnée de recherche

Les données de recherche peuvent avoir une durée de vie plus grande que le projet de recherche en lui-même (UK Data Service [sans date]b), et cette durée de vie est schématisée dans le modèle proposé par le UK Data Archive (Figure 1). Ce modèle est représenté sous la forme d'un cycle composé de six étapes distinctes à travers lesquelles la donnée transite :

- La création des données
- Le traitement des données
- L'analyse des données
- La préservation des données
- L'accès aux données
- La réutilisation des données

Figure 1 : Cycle de vie d'une donnée de recherche



Source : UK Data Archive, Training Materials, septembre 2011, p.19

Une notion importante qui transparait dans ce modèle est la notion de continuité. Une donnée peut être réutilisée, pour une deuxième analyse ou dans un nouveau projet, et recommencer ainsi son cycle de vie. Ainsi, une donnée doit pouvoir être réutilisée à (très) long terme, et donc être préservée à (très) long terme.

Une autre notion importante à différencier est celle de la gestion de données actives de recherche « active data management » qui concerne les données produites durant le processus de recherche et qui sont en cours d'utilisation par le chercheur (EduTech Wiki 2020 ; Gaillard 2014 ; Université de Lausanne [sans date]a). Dans le cycle de vie, ces données actives de recherche se retrouvent dans les étapes « Creating data », « Processing data », « Analysing data », donc avant l'étape « Preserving » (Gaillard 2014 ; Université de Lausanne [sans date]a). Comme mentionné dans la section 2.2, les données actives sont hors du périmètre de ce travail.

3.2.2 La gestion des données de recherche

Il devient de plus en plus important de gérer ses données de recherche afin de pouvoir faire face à l'explosion du volume des données digitales (Pryor 2012 ; Pinfield et al. 2014 ; Blumer et Burgi 2015 ; Bari, Bezzi et Guirlet 2020).

Cox et Verbaan (2018, p.4) donnent de la GDR la définition suivante : « *is about creating, finding, organising, storing, sharing and preserving data within any research process*". Dans cette definition, on remarque d'une part que la GDR est constituée de plusieurs activités, celles que l'on retrouve dans le cycle de vie (Figure 1), et d'autre part qu'elle couvre toutes les étapes du cycle de vie d'une donnée, de la génération de cette donnée à sa réutilisation (Blumer et Burgi 2015).

Une bonne GDR nécessite de nouvelles compétences de la part du chercheur et la mise en place de service de soutien pour l'accompagner (Pryor 2012 ; Bari, Bezzi et Guirlet 2020; Blumer et Burgi 2015). Bien que ses services de soutien soient souvent sous la responsabilité des bibliothèques académiques, la collaboration avec d'autres acteurs, tels que le chercheur ou le service informatique, est essentielle (Pinfield et al. 2014 ; Blumer et Burgi 2015 ; Bagnoud 2016 ; Bari, Bezzi et Guirlet 2020).

Nous allons maintenant nous intéresser aux activités de préservation et de réutilisation des données.

3.3 Préserver une donnée de recherche

Préserver une donnée signifie garantir l'accès et l'intelligibilité à son contenu et cela suppose au minimum un format adéquat (non-propriétaire et ouvert), une documentation de préservation (métadonnée) et un environnement sécurisé (dépôt de préservation) (Bagnoud 2016 ; France Archives [sans date] ; Corti et al. 2011).

Si on s'intéresse à la préservation du point de vue du chercheur, plusieurs activités sont requises de sa part : il doit la sélectionner, la préparer (format et documentation de préservation), puis la verser dans un dépôt de préservation.

3.3.1 Sélectionner une donnée

Pour des contraintes de coût, on ne peut préserver toutes les données et une sélection doit être faite (Université de Genève [sans date]i). Cette sélection est sous la responsabilité du chercheur car lui seul peut véritablement connaitre la valeur de la donnée qu'il a générée (Université de Genève 2018 ; Gaillard 2014 ; EPFL [sans date] ; Digital Curation Center (DCC) [sans date]a). Pour effectuer cette sélection, le chercheur peut s'aider de plusieurs questions (Université de Genève [sans date]i ; EPFL [sans date] ; Delft University of Technology 2020a) :

- Il y a-t-il une obligation à préserver la donnée (de la part du bailleur de fond ou de l'institution par exemple) ? Par exemple, le FNS demande que toutes données supportant une publication doit être préservées et partagées (FNS [sans date]d)
- Est-ce que les données sont en accord avec les lois de protection des données ? Si ce n'est pas le cas, les données devront être rendues anonymes ou l'accès aux données devra être restreint
- Est-ce que la donnée est « unique » : est-il possible de la répliquer ? Est-il moins cher de répliquer la donnée plutôt que de la préserver ? Par exemple, les données d'une étude de marché sont uniques car il est impossible d'obtenir des données identiques en répliquant l'étude
- Est-ce que la donnée a de la valeur : quelquefois il vaut mieux conserver l'algorithme qui a permis de générer les données que les données elles-mêmes
- Est-il possible de réutiliser la donnée ? Est-ce que vos données sont suffisamment documentées ?

Alors que la notion de réutilisation n'apparait pas de manière explicite dans la plupart des définitions de la préservation, on se rend compte que tout le monde s'accorde à dire qu'il est inutile de préserver une donnée si on ne peut la réutiliser. D'ailleurs, l'Université de Genève le mentionne clairement dans sa politique (Université de Genève 2018): « *Toute décision*

de préservation à long terme des données de recherche se fondera sur leur intérêt et leur qualité, ainsi que sur les possibilités de ***réutilisation***. »

3.3.2 Préparer une donnée

Par format adéquat, on entend l'utilisation d'un format non-propriétaire (n'appartenant pas à une entreprise) et ouvert (la description du format est publiquement accessible). Certains formats, bien que propriétaires, sont néanmoins ouverts et peuvent être utilisés pour la préservation des données (c'est le cas par exemple d'Adobe PDF ou de Microsoft OOXML) (Australian National Data Service (ANDS) 2017a). A l'opposé, un format propriétaire et fermé nécessitera l'utilisation d'un logiciel spécifique pour pouvoir accéder à la donnée, avec le risque qu'un jour ce logiciel ne soit plus pris en charge (Université de Genève [sans date] e ; Australian National Data Service (ANDS) 2017a). Quelquefois, il est indispensable de conserver les données dans le format d'origine pour ne pas perdre de l'information lors de la conversion du format. Dans ce cas, il est recommandé de sauver les données dans le format d'origine et dans un format non-propriétaire (Université de Genève [sans date] e ; Stanford University [sans date]).

La question de la durée de préservation se pose également. En effet, cette durée dépend de la valeur de la donnés et, pour certaines données, cette valeur peut diminuer avec le temps (avancement des connaissances, technologie plus sensible) (EPFL [sans date] ; Académie des sciences naturelles 2018 ; Burgi, Blumer et Makhlouf-Shabou 2017). Ainsi, pour une donnée de recherche, une durée de préservation de cinq à vingt ans semble raisonnable, et le FNS recommande une préservation de dix ans (EPFL [sans date] ; FNS [sans date]a). Si une donnée doit être préservée plus de dix ans, il est essentiel d'utiliser un format ouvert pour en garantir l'accès. Cependant, on estime que pour une préservation inférieure à dix ans, l'utilisation de formats propriétaires usuels, tel que Microsoft excel, est acceptable (ETHZ [sans date]a ; UK Data Service [sans date]a).

Le concept de documentation de préservation (métadonnée) sera abordé du point de la réutilisation d'une donnée (section 3.4.2.3).

3.3.3 Déposer une donnée

Il existe trois types de dépôts pour déposer des données de recherche en libre accès (Université de Genève 2019b):

- le dépôt disciplinaire, spécialisé dans un domaine de recherche
- le dépôt générique, capable d'accepter des données de tous les domaines de recherche (et donc de tout type)
- le dépôt institutionnel, appartenant à une institution

Ces dépôts sont répertoriés dans le registre re3data[1] qui permet de rechercher un dépôt par le biais de plusieurs critères, entre autre la discipline, le type de donnée, ou la localisation géographique. On en dénombre plus de 2500 et la grande majorité de ces dépôts sont des dépôts disciplinaires (plus de 2000).

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

Le Tableau 1 liste certains avantages et inconvénients de ces différents types de dépôts (Université de Genève 2019c ; Gaillard 2014 ; EPFL [sans date] ; Whyte 2015 ; Schneider 2018) :

Tableau 1 : Avantages et inconvénients des différents types de dépôts

Type de dépôt	Avantage	Inconvénient	Exemple de dépôt
Dépôt disciplinaire	Spécialisé dans un domaine	Pas de garanti de pérennité	FORS[2], DASH[3]
Dépôt générique	Convient à tous les types de données	Peut ne pas être conforme aux exigences de l'institution	Zenodo[4], Dryad[5], Dataverse[6]
Dépôt institutionnel	Conforme aux exigences de l'institution, garantie de pérennité	Peu visible	Yareta[7] (université de Genève), 4tu.ResearchData[8] (université de Delft)

En Suisse, le FNS accepte que les données soient déposées dans les trois types de dépôts pour autant que le dépôt réponde aux principes FAIR et soit non-commercial (FNS [sans date]b).

Le dépôt disciplinaire étant spécialisé dans la gestion et le traitement d'un type de donnée, on recommande en général aux chercheurs de déposer ses données en priorité dans un dépôt disciplinaire, et de n'utiliser un dépôt générique ou institutionnel qu'en l'absence d'un dépôt disciplinaire répondant aux exigences du bailleur de fonds, de l'institution ou de l'éditeur (Whyte 2015 ; Schneider 2018). La même recommandation est faite aux chercheurs de l'Université de Genève et il n'y a donc aucune obligation à utiliser Yareta (Université de Genève 2020).

3.3.3.1 Le modèle OAIS

Le modèle de référence pour l'archivage numérique est le modèle OAIS (Open Archival Information System), également enregistré comme norme ISO sous la référence ISO :14721 :2012 (France Archives 2018 ; Iacconi 2018). Ce modèle met l'accent sur la pérennisation du contenu informatif et il est suffisamment abstrait et généraliste pour pouvoir s'appliquer à toute organisation devant conserver de l'information (CCSDS 2012 ; France Archives 2018 ; CCSDS 2017, p. 2-5).

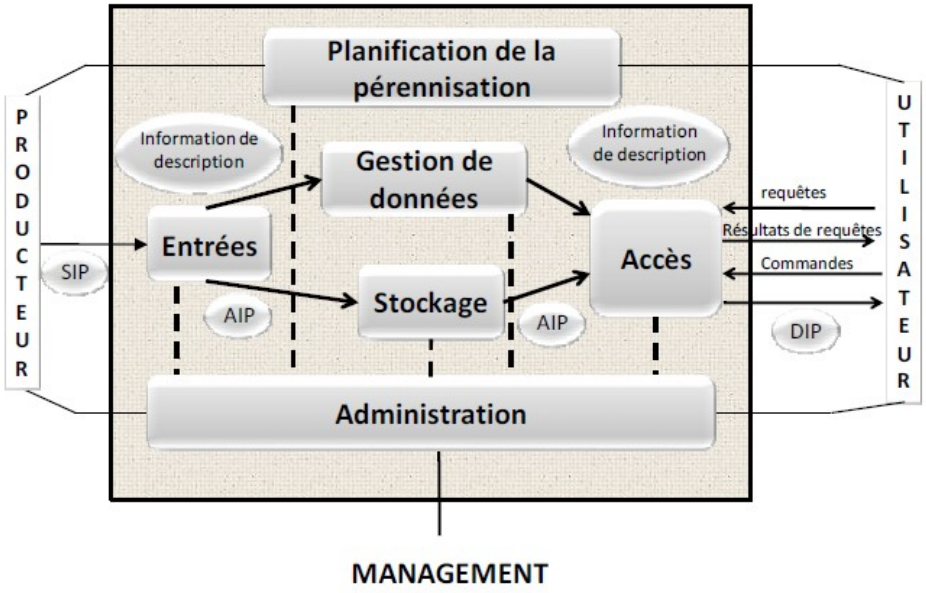
On retrouve trois acteurs principaux dans l'environnement d'une archive OAIS, ayant chacun un rôle défini (CCSDS 2012 ; Iacconi 2018 ; Tièche et Dubois, 2015 ; CCSDS 2017) :

[2] https://forscenter.ch/
[3] https://dash.nichd.nih.gov/
[4] https://zenodo.org/
[5] https://datadryad.org/stash
[6] https://dataverse.org/
[7] https://yareta.unige.ch/
[8] https://data.4tu.nl/

- Le producteur fournit l'information à pérenniser
- L'utilisateur cherche et récupère l'information qui l'intéresse
- Le management définit le mandat et les priorités de l'archive OAIS, en fonction de la politique globale de l'institution

Le modèle OAIS est constitué de six entités fonctionnelles, dont quatre entités de base : « Entrée » ; « Gestion de données », « Stockage », « Accès ». Les deux autres entités fonctionnelles sont l'entité « Administration » et l'entité « Planification de la pérennisation » (Figure 2) (CCSDS 2017 ; CCSDS 2012).

Figure 2 : modèle OAIS



(Source : CCSDS 20017, schéma 4-1)

Trois types de paquets d'information vont circuler dans ces quatre entités de base (Figure 2) (CCSDS 2012 ; CCSDS 2017, p. 1-15) :

- Le paquet d'information à verser, ou SIP (*submitted Information package*) qui est livré par le producteur
- Le paquet d'information archivé, ou AIP (*archived Information package*), qui est pérennisé dans l'OAIS
- Le paquet d'information diffusé, ou DIP (*diffused Information package*), qui est envoyé à l'utilisateur

Chaque paquet d'information est composé d'une information de contenu et d'une information de pérennisation, toutes deux identifiées et encapsulées par une information d'empaquetage, et résultant en un paquet qui peut être retrouvé grâce à l'information de description (CCSDS 2012 ; Tièche et Dubois 2015 ; CCSDS 2017, p. 2-6).

Nous avons vu que le format est une composante essentielle de la préservation. Dans le modèle OAIS, les migrations numériques des formats sont gérées au niveau de l'entité fonctionnelle « Planification de préservation ». En effet, c'est elle qui « *assure les fonctions et services relatifs à la surveillance de l'environnement de l'OAIS et à la production de recommandations visant à ce que les informations stockées dans l'OAIS restent accessibles,*

compréhensibles et suffisamment utilisables pour la Communauté d'utilisateurs cible sur le Long terme, même si l'environnement informatique d'origine devient obsolète » (CCSDS 2017, p. 1-11).

3.4 Réutiliser une donnée de recherche

Cette section répond au premier objectif de ce mandat, à savoir quelles sont les qualités qu'une donnée de recherche doit avoir pour être réutilisable.

3.4.1 Une donnée de qualité

L'OCDE (p. 24) mentionne que la valeur et l'utilité d'une donnée de recherche dépend de sa qualité et qu'il y a des normes de qualité à respecter, ou à élaborer si ces normes n'existent pas encore. Cependant, ces normes dépendent du domaine de recherche et il n'est pas envisageable d'avoir une norme de qualité universelle (OCDE 2007) :

« *la valeur et l'utilité des données de recherche dépendent pour une large part de la qualité des données elles-mêmes* »

« *Il n'est cependant pas réaliste d'envisager des normes universelles de qualité des données car certains domaines de recherche exigent des normes plus rigoureuses* »

Le FNS parle de données suffisamment précises pour permettre leur réutilisation (FNS [sans date]b). Dans sa politique institutionnelle sur la gestion des données de recherche (Université de Genève 2018), l'Université de Genève, mentionne que « *toute décision de préservation à long terme des données de recherche se fondera sur leur intérêt et leur qualité, ainsi que sur les possibilités de réutilisation* », et qu'il est de la responsabilité du chercheur de veiller à la bonne qualité des données ainsi qu'à leur description complète, mais sans définir la notion de « qualité ».

Il apparait donc qu'il n'existe pas de norme universelle de qualité des données de recherche, que la qualité d'une donnée de recherche dépend du domaine de recherche, et qu'il est de la responsabilité du producteur de la donnée (en l'occurrence le chercheur) d'évaluer la qualité de sa donnée.

3.4.2 Une donnée FAIR

La notion de réutilisation se retrouve dans les principes FAIR (Wilkinson et al. 2016). Ces principes définissent des caractéristiques permettant aux données d'être trouvées (*Findable*), accessibles (*Accessible*), interopérables (*Interoperable*), et réutilisables (*Reusable*) aussi bien par les machines que par les hommes (Wilkinson et al. 2016). Dans ce travail, nous nous intéressons surtout au R de *Reusable*, mais la définition complète des principes FAIR se trouve en Annexe 1.

Pour la réutilisation, on trouve en autre les notions de description, de licence et de standard :

- une description riche : « *meta(data) are richly described with a plurality of accurate and relevant attributes* », « *(meta)data are associated with detailed provenance* »
- une licence adéquate : « *(méta)data are released with a clear and accessible data usage licence* »
- un standard : « *(meta)data meet domain-relevant community standards* »

3.4.2.1 Une description riche

Concernant la richesse de la description, le UK Data Archive a élaboré un document de bonnes pratiques sur la gestion et le partage des données, qui liste les informations à fournir pour bien documenter une donnée (Van den Eynden 2011). Cette liste, suffisamment générale pour pouvoir s'appliquer à tout type de donnée indépendamment du domaine de recherche, contient des informations sur :

- Le contexte de la collecte des données : objectifs et hypothèses du projet de recherche
- Les méthodes de collecte des données : échantillonnage, procédure et protocole appliqués, appareillage et logiciel utilisés, date et lieu de la collecte
- Les liens entre les fichiers et la structure du jeu de données
- L'intitulé et la description des variables
- La définition des codes et des acronymes ou abréviations utilisés
- La description des modifications apportées aux données brutes
- Les informations concernant l'accès et les conditions d'utilisation des données

Ces informations peuvent s'ajouter au niveau des métadonnées, ou dans un fichier de type README mais en sachant qu'à la différence des métadonnées, le fichier README ne sera pas lisible par une machine (EPFL [sans date]).

3.4.2.2 Une licence adéquate

Nous avons vu que pour être réutilisable, une donnée doit avoir une licence permettant le partage (Wilkinson et al. 2016). Les Creative Commons[9] (CC) sont les sources les plus connues pour les licences ouvertes (Université de Genève [sans date]c). Fondées sur les droits d'auteur, elles s'appliquent dans le monde entier et permettent aux titulaires des droits d'attribuer certaines conditions à ses données : l'autorisation de l'usage commercial et de la modification de ses données, ainsi que l'obligation de partager les données modifiées sous la licence initiale (Creative Commons [sans date]a). Les différentes conditions possibles sont résumées dans le Tableau 2.

Tableau 2 : Conditions des différentes licences CC

Licence	Acronyme	Description de la licence
Attribution	BY	Toutes les licences Creative Commons obligent ceux qui utilisent vos œuvres à vous créditer de la manière dont vous le demandez, sans pour autant suggérer que vous approuvez leur utilisation ou leur donner votre aval ou votre soutien.
Attribution + partage dans les mêmes conditions	BY-SA	Le titulaire des droits autorise toute utilisation de l'œuvre originale (y compris à des fins commerciales) ainsi que la création d'œuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'œuvre originale.

[9] https://creativecommons.org/

Licence	Acronyme	Description de la licence
Attribution + pas de modification	BY ND	Le titulaire des droits autorise toute utilisation de l'œuvre originale (y compris à des fins commerciales), mais n'autorise pas la création d'œuvres dérivées.
Attribution + pas d'utilisation commerciale + pas de modification	BY NC-ND	Le titulaire des droits autorise l'utilisation de l'œuvre originale à des fins non commerciales, mais n'autorise pas la création d'œuvres dérivés.
Attribution + pas d'utilisation commerciale	BY NC	le titulaire des droits autorise l'exploitation de l'œuvre, ainsi que la création d'œuvres dérivées, à condition qu'il ne s'agisse pas d'une utilisation commerciale (les utilisations commerciales restant soumises à son autorisation).
Attribution + pas d'utilisation commerciale + partage dans les mêmes conditions	BY NC-SA	Le titulaire des droits autorise l'exploitation de l'œuvre originale à des fins non commerciales, ainsi que la création d'œuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'œuvre originale.

Source : adapté de Creative Commons France, [sans date]

3.4.2.3 Des métadonnées

Les métadonnées sont des données sur une donnée (data about data) (Joudrey et Taylor 2017 ; Riley 2017), ou d'après Baykoucheva (2015, p. 72) « *a specific kind of information structured to describe, explain, locate, or make it possible to retrieve, use , or manage data* ».

Les caractéristiques principales d'une métadonnée sont de suivre un schéma, d'être lisible par une machine, et de permettre le moissonnage et l'interopérabilité entre les machines (Joudrey et Taylor 2017 ; Riley 2017 ; MANTRA 2020 ; Van den Eynden 2011).

Elles sont généralement divisées en trois catégories (Rice et Southall 2016, p. 26 ; MANTRA 2020 ; Riley 2017) :

- Les métadonnées descriptives qui permettent l'indexation, la découverte et la récupération (par exemple le titre, le nom de l'auteur ou le résumé)
- Les métadonnées techniques qui décrivent comment un ensemble de données a été produit, structuré ou comment il devrait être utilisé (par exemple la taille du fichier ou la date de création)
- Les métadonnées administratives qui permettent un accès adéquat et une gestion approprié du matériel

Il n'existe pas un standard universel de métadonnées mais des standards de métadonnées (Digital Curation Center (DCC) [sans date]b ; Université de Lausanne [sans date]c ; University of North Caroline [sans date] ; Baykoucheva 2015). Certains standards sont génériques (par exemple le Dublin Core[10]), alors que d'autres standards sont spécifiques à une discipline de

[10] https://dublincore.org/

recherche (par exemple Ecological Metadata Language[11]). On trouve sur plusieurs sites internet des liens vers les standards de métadonnées par discipline, mais on peut citer le site du Digital Curation Center (DCC) qui liste les standards par ordre alphabétique[12] ou par discipline[13]. Les métadonnées spécifiques à une discipline étant plus riches que les métadonnées génériques, il est recommandé de les utiliser en priorité lorsqu'elles existent (Université de Genève 2019a).

En conclusion, pour qu'une donnée soit réutilisable il faut pouvoir y accéder (format adéquat et licence adéquate) et la comprendre (description et métadonnée suffisantes).

3.5 Politique institutionnelle sur la gestion des données de recherche de l'Université de Genève

L'Université de Genève est la première haute école de Suisse à s'être dotée d'une politique institutionnelle sur la gestion des données de recherche en juin 2018 (Université de Genève 2018 ; Université de Genève 2019a). Cette politique définit les responsabilités de l'université et des chercheurs et mentionne clairement dans son préambule non seulement la diffusion mais également « *l'accessibilité des résultats de recherche et des données générées* » (Université de Genève 2018).

Concernant spécifiquement les données de recherche, cette politique définit les responsabilités suivantes (Université de Genève 2018) :

- L'université est responsable de fournir une infrastructure, des services et de l'assistance
- Le chercheur est responsable de la gestion de ses données et de veiller entres autres à leur description complète, à leur préservation et à leur éventuelle diffusion
- Le chercheur est responsable de mettre à disposition ses données de recherche, et ceci le plus largement possible et notamment lors de financement public

Ainsi donc il est de la responsabilité du chercheur de décrire ses données et de les rendre accessibles à un large public.

[11] https://eml.ecoinformatics.org/
[12] https://www.dcc.ac.uk/guidance/standards/metadata/list
[13] https://www.dcc.ac.uk/guidance/standards/metadata

4. Dépôts de préservation et services associés

4.1 Dépôts de préservation dans les institutions suisses

Au sein des institutions, on peut voir trois situations concernant les dépôts en accès public des données de recherche (Gaillard 2014) :

- Les institutions qui n'ont pas de dépôt institutionnel pour les données de recherche
- Les institutions qui ont créé un dépôt institutionnel pour les données de recherche
- Les institutions qui utilisent leur dépôt de publication pour les données de recherche

En classifiant les principales institutions suisses d'après ces trois situations, on peut remarquer qu'à l'heure actuelle, l'Université de Genève est la seule institution à disposer d'un dépôt institutionnel en libre accès et spécifique aux données de recherche. Toutefois, même si elle possède un dépôt institutionnel, sa recommandation est d'utiliser un dépôt spécifique à la discipline si un tel dépôt existe (Université de Genève 2020). Les autres institutions n'ont soit pas de dépôt institutionnel, soit utilisent le même dépôt que pour les publications (Tableau 3). A noter que ce tableau ne liste pas les dépôts en accès privé proposés par les institutions pour le stockage ou la préservation des données de recherche de ses chercheurs.

Tableau 3 : Dépôts en accès public proposés pour les données de recherche

Institution	Dépôt institutionnel de données de recherche	Nom du dépôt institutionnel ou dépôt recommandé
Université de Genève	Dépôt institutionnel	Yareta (Université de Genève [sans date]k)
EPFL	Sans dépôt institutionnel	Recommande l'utilisation de Zenodo, Dryad et Figshare (EPFL [sans date])
Université de Bâle	Sans dépôt institutionnel	Renvoie sur Zenodo, Dryad et re3data.org (University of Basel [sans date]b)
Université de Zurich	Sans dépôt institutionnel	Renvoie sur re3data.org et sur la liste des dépôts utilisés par les chercheurs de l'institution (University of Zurich [sans date])
Université de Lausanne	Sans dépôt institutionnel	Recommande l'utilisation de Zenodo (Université de Lausanne [sans date]b)
ETHZ	Même dépôt que pour les publications	Research Collection (ETHZ [sans date]b)
Université de Berne	Même dépôt que pour les publications	BORIS (University of Bern [sans date]a), mais recommande plutôt Zenodo, Dryad, B2SHARE et Dataverse car BORIS n'est pas adapté pour les fichiers lourds (University of Bern [sans date]b)

Mais cette situation est en train de changer puisque que l'université de Bâle planifie d'avoir un dépôt en partenariat avec d'autres institutions suisses et en lien avec le projet DLCM[14] (University of Basel [sans date]a), alors que l'université de Lausanne et l'université de Zurich sont partenaires pour un autre dépôt national, SWISSUbase, disponible à partir de janvier 2021(SWISSUbase [sans date] ; Swissuniversities [sans date]c). Quant à l'université de Berne, elle est en train de créer une solution institutionnelle adaptée aux données de recherche et nommée BORIS Research Data (University of Bern [sans date]b).

4.1.1 Le dépôt institutionnel de l'Université de Genève - Yareta

Comme nous l'avons vu dans la section précédente, l'Université de Genève possède sa propre solution d'archivage et de partage des données de recherche, nommée Yareta (Université de Genève [sans date]k ; Université de Genève [sans date]j). Yareta est un dépôt générique dont le développement s'est fait dans le cadre du projet national « Data Life-Cycle Management » (DLCM[15]) et du projet de loi cantonal « Infrastructure et service numérique pour la recherche » (Burgi 2019 ; Université de Genève 2019d).

Débuté en 2015, l'objectif du projet DLCM est de fournir de nouvelles ressources aux chercheurs afin de leur permettre d'implémenter une gestion des données de recherche (Burgi 2015 ; Burgi et Blumer 2018 ; Blumer et Burgi 2015 ; Bari, Bezzi et Guirlet 2020).

Ce projet est constitué de deux phases, et si l'on s'intéresse spécifiquement à Yareta (Université de Genève [sans date]f):

- 1ère phase de DLCM (2015-2018) : phase de conception et première étapes de développement de Yareta
- 2ème phase de DLCM (2018-2020) : passage du prototype au service (Burgi, Cazeaux, 2019)

Yareta a été mis en service en juin 2019 (Université de Genève [sans date]f) et comptabilisait au 12 août 2020 plus de 170 jeux de données[16]. Yareta est un dépôt cantonal de données de recherche disponible uniquement pour les hautes écoles genevoises, car son financement provient du projet de loi 12146 (PL 12146), adopté par le Grand Conseil de la République et canton de Genève le 24 novembre 2017, cependant les données qui y sont déposées sont librement accessibles (Université de Genève [sans date]g). Un dépôt national d'archivage des données de recherche, nommé OLOS, est actuellement en cours de réalisation à l'Université de Genève et devrait être disponible en 2020 (DLCM [sans date]).

Yareta, mené selon une méthode agile, continue d'être amélioré et on peut trouver la liste de ses dernières fonctionnalités sur le site web de l'Université de Genève[17] (Université de Genève [sans date]f).

Le nombre de métadonnées à saisir pour déposer un jeu de données est minimal et correspond aux informations nécessaires pour générer un DOI : le titre, la description, la date de publication et le contributeur (Burgi et Cazeaux 2019 ; Université de Genève 2020).

[14] Ce dépôt correspond probablement au dépôt national OLOS, géré par l'Université de Genève
[15] https://www.dlcm.ch/
[16] https://yareta.unige.ch/frontend/
[17] https://www.unige.ch/eresearch/fr/services/yareta/dernieres-fonctionnalites/

Les principales caractéristiques de Yareta sont les suivantes (Université de Genève [sans date]j ; Université de Genève [sans date]f) :

- Intégration du modèle de référence pour la préservation numérique OAIS
- Respect des exigences du FNS et de H2020 (non-commercial et conformité avec les principes FAIR)
- Attribution pour chaque jeu de données d'un DOI et d'une licence CC
- Intégration du schéma de métadonnées DataCite[18]
- Compatibilité avec les formats de toutes les disciplines de recherche
- Possibilité de choisir entre trois niveaux d'accès : public, restreint, fermé
- Période de conservation (rétention) des données de cinq ou dix ans (selon le choix du chercheur), suivi d'une ré-évaluation de la rétention
- Fonctionnalités automatiques et interface web intuitive

Concernant les formats, Yareta ne propose pas de migrations numériques et cela pourrait être problématique pour des données devant être conservées plus de dix ans (CCSDS 2017, p. 5-1).

Deux freins au partage des données souvent mentionnés par les chercheurs sont l'activité chronophage et la peur de perdre le contrôle de ses données (Plomp et al. 2019 ; Tenopir et al. 2015). La possibilité d'avoir un accès restreint ou fermé, la quantité minimale d'information à fournir et la simplicité de l'interface permettent de répondre à ces deux freins (Burgi et Cazeaux 2019).

4.2 Services proposés par les institutions suisses

A l'exception de l'université de Berne, les ressources que ces institutions allouent à la GDR ont été analysées en janvier 2020 dans le cadre de notre projet de master sur les dispositifs d'e-learning en GDR du point de vue du DLCM (Bari, Bezzi et Guirlet 2020). De cet inventaire des ressources, nous avions retenus les caractéristiques suivantes :

- Les formations se font majoritairement en présentiel et sont de type atelier, cours, ou consultation individuelle
- Les ressources sont produites en interne, sans mutualisation pour les ressources génériques
- La responsabilité des formations est majoritairement sous l'égide de la bibliothèque, à l'exception de l'université de Bâle où la responsabilité est partagée entre bibliothèque/IT/recherche
- Le nombre de ressources allouées à la GDR a été augmenté

En s'intéressant plus spécifiquement aux activités de préservation et de réutilisation, on remarque que les ressources proposées par ces institutions sont comparables et couvrent les thématiques suivantes: sélection des données à préserver, choix des formats, nommage des fichiers, licence et dépôt de préservation (Bari, Bezzi et Guirlet 2020).

[18] https://schema.datacite.org/

Comparé à cet inventaire fait en janvier 2020, deux nouvelles ressources en lien avec notre thématique de préservation sont proposées : l'EPFL a rajouté un cours en présentiel d'une heure sur les métadonnées à raison de deux fois par an[19] et l'Université de Genève a rajouté un cours sur la gestion des données quantitative[20].

En regardant les ressources proposées par les institutions à l'internationale[21], on remarque qu'elles sont similaire à celles des institutions suisses (Bari, Bezzi et Guirlet 2020). On peut cependant remarquer que deux institutions offrent des formations en ligne : l'université d'Edinburgh a élaboré une formation sur la GDR (MANTRA) et l'université de Delft propose un MOOC « Open Science: Sharing Your Research with the World » (MANTRA 2020 ; Delft University of Technology 2020b).

Dans la section suivante, nous allons nous concentrer sur les services de soutien proposés par l'université de Delft.

4.3 A l'international – le cas de l'université de Delft

Reproduire et réutiliser des données nécessite une bonne GDR, cependant il peut être difficile pour le chercheur d'avoir une bonne pratique de GDR, et les enquêtes montrent que les obstacles sont plus d'ordre culturel que technique : manque d'encouragement, activité chronophage, manque de formation, préférence à partager ses données sur demande (Plomp et al. 2019 ; Tenopir et al. 2015).

L'université de Delft (TU Delft)[22] est un bon exemple de service spécialement dédiés pour les données de recherche, en particulier pour les deux profils de compétence qu'elle a mis en place :

- Un profil de généraliste : Data Steward
- Un profil de spécialiste : Data Champion

Cette université, basée aux Pays-Bas, est constituée de huit facultés comprenant au total quarante disciplines scientifiques et technologiques et possède un dépôt institutionnel pour les données certifié Data Seal of Approval depuis 2013 (Delft University of Technology [sans date]a ; Delft University of Technology [sans date]b).

Bien qu'ayant déjà des « Data Stewards » dans la majorité des facultés, une enquête ainsi que des entretiens menés en 2017-2018 ont montré que les pratiques de GDR pouvaient être améliorées (Plomp et al. 2019). En effet, près de la moitié des chercheurs n'utilisaient pas de dépôts de données alors qu'ils en connaissaient l'existence, et certains chercheurs n'étaient pas au courant des différents services et soutiens mis à leur disposition (Plomp et al. 2019). De plus, la méthodologie utilisée et le type de données générées étant très différents selon les facultés, chaque faculté avait des besoins différents et nécessitait d'avoir accès à des services adaptés à ces besoins (Plomp et al. 2019).

[19] https://www.epfl.ch/campus/library/services/services-researchers/rdm-training-events/
[20] https://www.unige.ch/researchdata/fr/actualites/formation-donnees-quantitatives/
[21] Les sites web de plusieurs bibliothèques académiques ont été consultés (Johns Hopkins Sheridan Libraries, Griffith University, Stanford Libraries, University of Oxford, The university of Edinburgh, Delft university)
[22] https://www.tudelft.nl/en/

Suite à ces constatations, l'université de Delft a initié un projet nommé «Data Stewardship » afin d'augmenter la sensibilité des chercheurs à la GDR et de leur proposer des services aptes à améliorer leurs compétences (Plomp et al. 2019). Pour ce faire, l'université a engagé un « Data Steward » par faculté, dont le rôle est de répondre aux questions des chercheurs concernant la GDR (Plomp et al. 2019). Le fait que ces « Data Stewards » soient intégrés à la faculté leur permet d'avoir un contact étroit avec les chercheurs, et de centraliser toutes les questions. De plus, le fait d'avoir un seul point de contact par faculté est également plus simple pour le chercheur. En outres, ces « Data Stewards » possèdent des PhD en lien avec la faculté dans laquelle ils sont engagés afin de faciliter la communication avec les chercheurs.

Ces « Data Stewards » sont chapeautés par un « Data Steward Coordinator », employé par la bibliothèque, et qui permet d'assurer une cohérence et une collaboration efficace entre les « Data Stewards » de chaque faculté.

Le « Data Steward » est un généraliste qui n'est pas familier avec chaque type de données présent dans sa faculté. Pour pallier à cela, l'université de Delft a également mis en place une communauté de « Data Champions ». Ces « Data Champions » sont des chercheurs ayant de bonnes pratiques en GDR et motivés à les préconiser auprès de leurs collègues. Au moment de la publication de cet article (Plomp et al. 2019), 45 « Data Champions » étaient répartis dans chaque faculté et dans presque chaque département. Le Tableau 4 détaille les différences entre un Data Steward et un Data Champion.

De prime abord, on peut avoir l'impression que la GDR est sous la responsabilité des facultés, cependant la bibliothèque garde un rôle actif et prépondérant par l'entremise des « Data Stewards Coordinators ».

Tableau 4 : Caractéristiques d'un Data Steward et d'un Data Champion

	Data Steward	**Data Champion**
Compétence	Généraliste	Spécialiste
Profile	Personne ayant un PhD en lien avec la faculté qui l'emploie	Chercheur
Rôle	Répondre aux questions concernant la GDR, et être la personne référente pour la GDR dans sa faculté	Avoir de bonnes pratiques en GDR et les préconiser auprès de ses collègues
Nombre	Un par faculté	Plus d'un par faculté, et idéalement un (ou plus) par département
Appartenance	Faculté	Faculté

L'université de Delft n'est pas la seule université à avoir une communauté de « Data Champions ». On retrouve cette communauté à l'université de Cambrigde[23], qui a d'ailleurs inspirée l'université de Delft (Plomp et al. 2019), et plus près de nous, à l'EPFL[24].

[23] https://www.data.cam.ac.uk/intro-data-champions
[24] https://www.epfl.ch/campus/library/services/services-researchers/rdm-contacts-communities/epfl-data-champions/

5. Analyse des dépôts de Yareta

Afin de proposer des livrables en adéquation avec les besoins des chercheurs, nous avons analysé l'existant, à savoir les jeux de données déposés sur Yareta jusqu'au 16 avril 2020.

A cette date, il y avait 88[25] jeux de données déposés sur Yareta, dont 78 dépôts publics. De ces 88 dépôts, 31 dépôts ont été entièrement analysés, 18 dépôts ont été partiellement analysés, et 39 dépôts n'ont pas été analysé, soit parce qu'ils avaient été créés pour les formations Yareta et contenaient sciemment des erreurs, soit parce qu'ils correspondaient à des versions antérieures d'un même dépôt (Tableau 5) (voir Annexe 2 pour le détail de dépôts).

Tableau 5 : Résumé des dépôts versés sur Yareta au 16 avril 2020

<table>
<tr><td rowspan="6">88 dépôts</td><td rowspan="3">78 dépôts publics</td><td>31 dépôts entièrement analysés (Datacite & METS[26] & fichiers téléchargés)</td></tr>
<tr><td>10 dépôts partiellement analysés (métadonnées Datacite ± METS ± fichiers téléchargés)
3 dépôts : Datacite + fichiers téléchargés
3 dépôts : Datacite + METS
4 dépôts : Datacite</td></tr>
<tr><td>37 dépôts pas du tout analysés
33 dépôts correspondant à des versions antérieures du même dépôt
4 dépôts créés pour la formation Yareta</td></tr>
<tr><td rowspan="2">7 dépôts restreints</td><td>5 dépôts partiellement analysés (Datacite ± METS)</td></tr>
<tr><td>2 dépôts pas du tout analysés (créés pour la formation Yareta)</td></tr>
<tr><td>3 dépôts fermés</td><td>3 dépôts partiellement analysés (Datacite ± METS)</td></tr>
</table>

Par analyse partielle, on entend une analyse ou un ou plusieurs critères n'ont pas pu être analysés à cause du niveau d'accès au dépôt ou de contraintes techniques dues à mon matériel informatique (Tableau 6) :

- Soit le jeu de données n'a pas pu être téléchargé (accès restreint/fermé ou dépôt trop lourd), mais les formats ont pu être analysés dans le document correspondant aux métadonnées METS
- Soit le jeu de données a pu être téléchargé, mais les formats n'ont pas pu être analysés dans le document correspondant aux métadonnées METS
- Soit le jeu de données n'a pas pu être téléchargé et les formats n'ont pas pu être analysés dans le document correspondant aux métadonnées METS, et l'analyse s'est limitée aux informations accessibles depuis le portail de Yareta (titre, description, licence/accès, métadonnées Datacite, date de collecte, taille)

[25] Pour des raisons de lisibilité, tous les nombres sont écrits en chiffre
[26] Datacite et METS sont des standards de métadonnées utilisés par Yareta

Tableau 6 : Analyse complète versus analyse partielle

	Métadonnées Datacite	Métadonnées METS (format)	Téléchargement et analyse des fichiers	Nombre de dépôts
Analyse complète	√	√	√	31
Analyse partielle	√	√	-	10
Analyse partielle	√	-	√	3
Analyse partielle	√	-	-	5
Non analysé	-	-	-	39

5.1 Choix des critères d'analyse

Le but de l'analyse des jeux de données était d'évaluer la possibilité de réutiliser les données ou non. Dans la section 3, nous avons vu qu'il n'y a pas de définition précise de ce qu'est une donnée de qualité, et qu'il n'existe pas un standard de métadonnées universel à toutes les données, mais des standards par discipline (Digital Curation Center (DCC) [sans date]b ; University of North Caroline [sans date] ; Université de Lausanne [sans date]c).

A partir des informations discutées dans les sections 3.3 et 3.4 et des informations à disposition sur le portail de Yareta (description, date de collecte, accès, licence), une liste des éléments principaux permettant la réutilisation d'une donnée a été élaborée. Pour qu'un jeu de données déposé sur Yareta soit réutilisable, il faut :

- qu'il soit accessible
  - une licence permettant la réutilisation (CC)
  - un format ouvert
  - en accès public
  - possible de le télécharger (en l'occurrence, dû à des contraintes techniques, il m'était impossible de télécharger des jeux de données dont la taille dépassait 1GB)
- qu'il soit compréhensible
  - suffisamment contextualisé pour être compréhensible : métadonnées riches, README

Des critères concernant la publication ont été inclus dans la grille d'analyse, d'une part pour pouvoir faire la distinction entre les données de recherche soutenant une publication de celles qui ne sont pas liées à une publication, et d'autre part pour pouvoir mettre en parallèle la présence (ou l'absence) de publication avec la présence (ou l'absence) de README. De plus Yareta n'étant pas encore référencé sur re3Data, le moyen le plus directe et le plus simple d'accéder aux données déposées sur Yareta est d'y accéder via la publication (i.e. mention du

DOI de Yareta dans la publication). Or, si le dépôt ou si le DOI du jeu de données ne sont pas mentionnés dans la publication, il sera difficile de trouver le jeu de données, et donc de le réutiliser.

En résumé, la grille d'analyse (Annexe 3) contient des critères couvrant trois aspects du jeu de données : son accessibilité, sa compréhension, et son lien à une publication.

Critères concernant l'accessibilité du jeu de données :

- Accès : le niveau d'accès (public, restreint ou fermé) choisi par le chercheur
- Licence : choix d'une licence à partir d'une liste
- Format des fichiers: information fournie par les métadonnées METS et par les extensions des fichiers téléchargés
- Téléchargement : possible ou non

Critère concernant la compréhension du jeu de données :

- Titre et description du jeu de données
- README : ce critère contient toutes les informations permettant de comprendre la donnée. Il a été subdivisé en plusieurs sous-sections portant sur le contexte de la récolte de données, les méthodes de collectes, la description des variables, la définition d'acronymes, …
- Date de collecte des données : début et fin de collecte
- Clarté des noms de fichiers
- Description des changements appliqués aux données (curation des données)

Critères concernant la publication :

- Données soutenant une publication
- Mention du jeu de données dans l'article
- Mention de la publication dans Yareta

Yareta étant continuellement amélioré (méthode agile), certaines fonctionnalités actuellement disponibles sur cet outil ne l'étaient pas au moment de l'élaboration de la grille d'analyse des dépôts et n'ont de ce fait pas été pris en compte (par exemple le champ « mots clés »).

Un exemple de deux grilles d'analyses complétées pour un dépôt entièrement analysé et un dépôt partiellement analysé se trouve dans l'Annexe 4[27].

5.2 Analyse des critères

Cette section résume l'analyse des jeux de données déposés sur Yareta en fonction des critères de la grille d'analyse.

[27] L'analyse des 49 dépôts entièrement ou partiellement analysés a été remise à la mandante ainsi qu'à la coordinatrice du projet DLCM.

5.2.1 Critère « Titre » et « Description »

Que les jeux de données soutiennent une publication ou pas, le titre du jeu de données est généralement suffisamment clair et précis. Pour les dépôts soutenant une publication, il correspond dans la majorité des cas au titre de la publication.

Il y a une grande disparité dans la description du jeu de données, qui va d'une description très générale (« *original data* », « *dataset* ») ou reprenant le titre de la publication (« *data shown in …* », « *original data files for the article …* »), à une description détaillée (« *This deposit contains motion capture files during walking and bi-plane x-rays of 2 patients with hip osteo-arthritis and 2 patients with total hip arthroplasty* »).

Afin de permettre l'accès à ces jeux de données via les moteurs de recherche, il est important que le titre et la description du jeu de données soient aussi précis que possible, et pour s'assurer de cette précision, une liste d'informations à mentionner dans ces deux champs pourrait être proposée aux chercheurs.

5.2.2 Critère « Publication »

Pour ce critère, l'analyse s'est faite sur les dépôts ayant pu être téléchargés, afin de pouvoir vérifier l'éventuelle mention de la publication dans le README, au cas où la publication n'était pas mentionnée dans le titre ou la description du jeu de données.

Sur les 34 dépôts téléchargés :

- 10 dépôts ne sont pas liés à une publication
- 24 dépôts sont liés à une publication :
  - 16 dépôts mentionnent la publication dans Yareta, alors que 8 dépôts n'en font pas mention
  - Seulement 3 dépôts mentionnent dans leur publication que leurs données sont disponibles sur Yareta

16 dépôts mentionnent clairement dans Yareta que leurs données sont liées à une publication, en mettant le titre ou la référence de la publication dans le titre ou la description du dépôt.

On retrouve rarement la mention du jeu de données et le DOI de Yareta dans la publication, mais cela est certainement dû au fait que Yareta n'est fonctionnel que depuis juin 2019 et que la plupart des publications ont été envoyées aux éditeurs, voir approuvées, avant cette date.

Lorsque les données soutiennent une publication, il faudrait recommander au chercheur de le mentionner dans Yareta ou rajouter un champ « publication » dans Yareta afin de faciliter l'accès à la publication et de permettre la compréhension et la réutilisation des données, et ce d'autant plus en l'absence de README.

5.2.3 Critère « Format »

Les formats de 41 dépôts ont été répertoriés depuis le document correspondant aux métadonnées administratives METS. Ce document, généré automatiquement par Yareta lors du dépôt du jeu de données, est uniquement accessible après avoir téléchargé le jeu de données. Pour les dépôts trop lourds à télécharger, ou en accès restreint ou fermé, ce document m'a été fourni en interne.

A l'exception de trois formats, tous les formats répertoriés dans le document correspondant aux métadonnées METS sont conformes aux formats listés par le UK Data Service et mentionnés sur la page web de l'Université de Genève (Université de Genève [sans date]e) (Tableau 7).

Trois formats sont mentionnés dans le document correspondant aux métadonnées METS et ne sont pas répertoriés par le UK Data Service : MATLAB (mentionné dans trois dépôts), DNG (mentionné dans deux dépôts), et DICOM (mentionné dans un dépôt).

Tableau 7 : Formats répertoriés dans les métadonnées METS

Format	**Recommandation du UK Data Service**[28]	**Nombre de dépôt ayant ce format**
PNG	Acceptable	1
PDF	Acceptable pour les images recommandé pour la documentation et les scripts	6
csv	Recommandé	8
JPEG	Image : acceptable Vidéo : recommandé	1
rtf	Recommandé	4
xml	Tabular : recommandé Text : recommandé Documentation and script : acceptable	3
OOXML	Tabular : recommandé Text : recommandé Documentation and script : acceptable	5
Plain text file	Text: recommandé Documentation and script : acceptable	12
x-wave	Audio : acceptable	3
ods	Tabular : acceptable	1
Markdown/Text	Tabular : recommandé	3

Certains formats mentionnés dans le document correspondant aux métadonnées METS ne correspondent pas aux formats des fichiers téléchargés, et plusieurs explications sont possibles :

- le jeu de données est déposé en tant que fichier zip et, Yareta le reconnaissant comme tel, les fichiers à l'intérieur du zip ne sont pas analysés
- le format est inconnu de FITS[29] (utilisation d'un format propriétaire)
- le fichier est mal formaté (l'extension ne correspond pas au type de fichier) et n'a pas été reconnu par FITS
- le fichier est correctement formaté mais n'a pas été reconnu par FITS

[28] Reprise de la page web de l'Université de Genève (Université de Genève, [sans date]e)
[29] Outil d'identification de format utilisé par Yareta

Les trois premiers cas sont des erreurs lors de la préparation du fichier ou lors du dépôt du jeu de données et seront pris en compte dans les livrables. Le dernier cas est potentiellement une erreur de l'outil Yareta et sera vérifié en interne.

Deux dépôts contiennent des fichiers en format rtf associé à word 2007. Yareta ne faisant pas de migration numérique, il y a un risque que ces documents ne soient plus lisibles à la fin de la période de préservation (10 ans). Plusieurs dépôts contiennent des fichiers générés par des instruments de mesures. Ces fichiers sont dans des formats propriétaires, impossible à ouvrir sans le logiciel adéquat. Dans ce cas, la recommandation de l'Université de Genève est de sauver les données dans un format ouvert, ou dans le format d'origine et dans un format ouvert en cas de perte d'information lors de la conversion du format (Université de Genève [sans date]e).

Il aurait été intéressant de répertorier le nombre de format non-reconnus par FITS dans le document correspondant aux métadonnées METS, mais pour des contraintes de temps, cette analyse n'a pas pu être faite.

5.2.4 Critère « Licence »

Lorsque le dépôt est en accès public, le chercheur doit obligatoirement choisir une licence CC dans la liste proposée par Yareta. Par contre, lorsque le dépôt est en accès restreint ou fermé, le chercheur a le choix de ne pas mentionner de licence.

Sur les 49 dépôts entièrement ou partiellement analysés, plus de la moitié des dépôts ont une licence CC BY 4.0 (Tableau 8).

Tableau 8 : Répartition des licences des dépôts analysés

Types de licences telles qu'elles sont mentionnées sur Yareta	Accès public (41 dépôts)	Accès restreint	Accès fermé
CC0 1.0	4	-	-
Creative Commons Attribution 4.0	23	1	1
Creative Commons Attribution 4.0 International	1	2	1
Creative Commons Attribution-NonCommercial 4.0 International	5	-	-
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International	7	1	-
Creative Commons Attribution-ShareAlike 4.0 International	1	-	-
Aucune licence mentionnée	-	1	1

Il est intéressant de remarquer que des licences CC, donc permettant le partage, ont été attribuées à des jeux de données en accès restreint ou fermé, donc non-partageables. De plus, certains groupes ayant déposé plusieurs dépôts ont utilisés des licences différentes pour leurs dépôts. Ces deux points seront abordés avec les groupes concernés lors des entretiens.

Les acronymes des licences (CC, BY, NC, …) ne sont pas mentionnés sur Yareta et cela peut rendre la compréhension plus difficile pour les chercheurs habitués aux acronymes durant les formations. De plus, deux licences CC BY 4.0 sont listées sur Yareta (Creative Commons Attribution 4.0 et Creative Commons Attribution 4.0 International) alors que ces deux licences sont à priori identiques.

5.2.5 Critère « Date de collecte »

Sur 49 dépôts entièrement ou partiellement analysés, il y a :

- 10 dépôts qui ont correctement complété les dates de collecte
- 6 dépôts qui ont complété les dates de collecte de manière incorrecte : les dates de collectes correspondent aux dates du dépôt du jeu de données et mentionnent la même date pour le début et la fin de collecte
- 30 dépôts qui n'ont pas de date de collecte
- 3 dépôts qui ont partiellement complété les données : uniquement la date de début de la collecte est mentionnée

La date de collecte des données n'est pas un champ obligatoire, et la majorité des chercheurs n'ont pas complété ce champ d'information. On peut cependant remarquer que ce champ est parfois compris comme étant la date du dépôt du jeu de données, bien qu'il soit défini sur le portail Yareta comme étant « *la date à laquelle vous avez commencé (ou terminé) à collecter/générer les données* ». Il vaudrait donc la peine de clarifier ce que l'on entend par « date de collecte » dans le guide d'archivage ou durant les formations.

5.2.6 Critère « README »

Il n'y a pas d'obligation à fournir un README, et le guide d'utilisation de Yareta[30] n'en fait pas mention. Il n'est donc pas surprenant de trouver si peu de README. En effet, sur les 34 dépôts téléchargés, la majorité des dépôts n'ont pas ce document (Tableau 9) :

Tableau 9 : Résumé des dépôts ayant un README

34 dépôts téléchargés	9 dépôts avec un README	7 dépôts liés à une publication
		2 dépôts non liés à une publication
	25 dépôts sans README	17 dépôts liés à une publication
		8 dépôts non liés à une publication

Cependant, il est extrêmement difficile de réutiliser des données s'il n'y a ni README, ni publication liée aux données. Il faudrait donc inciter le chercheur à fournir un README (ou tout autre document permettant la compréhension du jeu de données), en particulier lorsque les données ne sont pas liées à une publication.

Concernant la localisation du README, il n'y a pas vraiment de cohérence, et selon les dépôts, le README se trouve soit dans le dossier « documentation », soit dans le dossier « researchdata ». Afin de faciliter la découverte et l'accès à ce document, nous pourrions recommander aux chercheurs de déposer son README dans le dossier « documentation »,

[30] https://yareta.unige.ch/doc/Yareta-QuickStartGuide.html#metadata

ou de déposer une TOC listant la localisation des README dans le dossier « documentation » s'il est plus judicieux d'avoir ce document dans le même dossier que les données.

5.2.7 Critère « Réutilisabilité »

L'analyse de la réutilisabilité s'est faite selon trois critères :

- Le format du fichier
- La possibilité d'ouvrir ou non le fichier
- L'existence d'une publication ou d'un README

Concernant le format, et en me référant aux formats recommandés par le UK Data Service et référencés sur le site de l'Université de Genève (Université de Genève [sans date]e), tous les formats listés dans le document correspondant aux métadonnées METS sont des formats considérés comme recommandés ou acceptables.

Concernant la possibilité d'ouvrir les fichiers, et en tenant compte des contraintes techniques liées à mon matériel informatique, sur les 34 dépôts téléchargés, il y a 16 dépôts dont tous les fichiers s'ouvrent. De ces 16 dépôts:

- 9 dépôts sont liés à une publication, et en posant comme postulat que toutes les informations concernant les données sont décrites dans la publication, on peut supposer que les données de ces dépôts sont réutilisables
- 7 dépôts ne sont pas liés à une publication et aucun de ces dépôts ne possédant un README, on peut supposer que ces données ne sont pas réutilisables car nous ne possédons aucune information sur le contexte, les méthodes utilisées, etc.

Sur les 34 dépôts téléchargés, en posant comme postulat que toutes les informations nécessaires à la réutilisation des données sont mentionnées dans la publication, et en tenant compte des contraintes de mon matériel informatique, 9 jeux de données auraient pu être réutilisés.

5.3 Synthèse de l'analyse des dépôts de Yareta

A partir de cette analyse des dépôts de Yareta, plusieurs problématiques ont pu être identifiées, et des propositions d'amélioration sont proposées pour chaque problématique (voir Tableau 10). Ces propositions sont à inclure soit dans les livrables de ce travail (guide d'archivage ou formation) ou dans le guide d'utilisation de Yareta[31], soit à implémenter directement dans l'outil Yareta pour autant que cela soit techniquement possible.

Les deux problématiques les plus importantes qui ressortent de cette analyse sont l'impossibilité à ouvrir les fichiers et la difficulté à comprendre les données en tant que non-expert (ou grand public).

[31] https://yareta.unige.ch/doc/Yareta-QuickStartGuide.html#metadata

Tableau 10 : Identification des problématiques et proposition d'amélioration

Titre	
Situation / Problématique	Titre imprécis
Proposition	Proposer une règle commune pour mentionner le titre
Support	Guide d'archivage, éventuellement Yareta
Description	
Situation / Problématique	Description imprécise
Proposition	Lister les informations minimales à mentionner dans la description et proposer un exemple de description
Support	Guide d'archivage, éventuellement Yareta
Publication	
Situation / Problématique	Aucune mention de la publication dans Yareta
Proposition	Rajouter un champ « Publication » dans Yareta*
Support	Yareta
Format	
Situation / Problématique	Jeu de données déposé en tant que fichier zip
Proposition	Clarifier comment déposer des données sur Yareta
Support	Guide d'utilisation de Yareta, formation
Format	
Situation / Problématique	Fichier d'origine en format propriétaire et non-ouvert (ex : instrument de mesure)

Proposition	Recommander de sauver dans un format ouvert, ou dans le format d'origine et dans un format ouvert
Support	Guide d'archivage, formation
Format	
Situation / Problématique	Fichier sauvé avec une extension incorrecte
Proposition	Message d'erreur de Yareta*
Support	Yareta
Format	
Situation / Problématique	Fichier d'origine en format non-pérenne (ex : word 2007)
Proposition	Recommander de sauver dans le format d'origine et un format ouvert
Support	Guide d'archivage, formation
Licence	
Situation / Problématique	Licence inadéquate
Proposition	Expliquer les licences. Sur Yareta, proposer un questionnaire oui/non pour choisir la licence (e.g. usage commerciale, modification des données, …) *
Support	Guide d'archivage, formation, Yareta
Date de collecte	
Situation / Problématique	Mention de la date du dépôt, ou autre information fausse
Proposition	Clarifier l'information attendue
Support	Guide d'archivage, guide d'utilisation de Yareta
README	
Situation / Problématique	Non fourni
Proposition	Recommander de l'inclure dans le dépôt

Support	Guide d'archivage, formation
README	
Situation / Problématique	Insuffisant pour comprendre le jeu de données
Proposition	Proposer un modèle ou une liste minimale d'informations à fournir
Support	Guide d'archivage, formation
README	
Situation / Problématique	Localisé dans le dossier « researchdata »
Proposition	Recommander de le déposer dans le dossier « documentation » ou déposer une TOC dans le dossier « documentation »
Support	Guide d'archivage, guide d'utilisation de Yareta, formation
Réutilisabilité	
Situation / Problématique	Impossible d'ouvrir le fichier
Proposition	Recommandation portant sur les formats à utiliser
Support	Guide d'archivage, formation
Réutilisabilité	
Situation / Problématique	Impossible de comprendre les données
Proposition	Reprendre les recommandations portant sur le README
Support	Guide d'archivage, formation
Réutilisabilité	
Situation / Problématique	Impossible de comprendre ce que le fichier représente
Proposition	Recommandation portant sur le nommage des fichiers et/ou des dossiers
Support	Guide d'archivage, formation

* pour autant que cela soit techniquement possible

6. Entretiens

Suite à l'analyse des dépôts, des groupes de recherche ont été contactés afin de participer à des entretiens. Ces entretiens se sont déroulés du 20 mai au 18 juin 2020 en visioconférence (zoom), en français ou en anglais, et ont duré entre 15 minutes et 60 minutes.

Dans un premier temps, les répondants ont été choisis parmi les groupes ayant déposés plusieurs jeux de données sur Yareta, et étant affiliés à la faculté des sciences. Puis, afin d'augmenter la diversité des disciplines, cette sélection s'est étendue à des groupes affiliés à la faculté des sciences et ayant déposé un jeu de donné sur Yareta, ou ayant déposé leurs données sur un autre dépôt que Yareta. Au total, sept entretiens[32] ont été menés avec des répondants appartenant à trois disciplines différentes : chimie, physique et biologie.

Le guide d'entretien (voir Annexe 5) a été validé par la mandante et envoyé aux répondants avant l'entretien. Les questions portaient principalement sur la procédure pour déposer les jeux de données et sur leurs éventuels besoins :

- Procédure :
  - Qui dépose et quand
  - Convention et règle au sein du groupe de recherche
  - Préparation des données : sélection et modification des données, README, format, licence, accès
- Besoins actuels et futurs
- Utilisation d'un Electronic Lab Notebook (ELN)

Les principaux résultats de ces entretiens sont les suivants :

- Tous les répondants déposent uniquement les données soutenant une publication
- Toutes les données déposées ont été traitées et ce ne sont jamais les données brutes qui sont déposées
- Dans chaque groupe, une seule personne dans le groupe est responsable de déposer les données
- Les données sont centralisées (soit sur le serveur de l'Université de Genève, soit sur un serveur du groupe) et sont préparées parallèlement à la publication
- Tous les répondants déposent leurs données essentiellement pour répondre aux exigences du FNS
- Aucun répondant n'a exprimé de besoin spécifique en termes de formation ou de soutien, si ce n'est pour déposer des dépôts lourds

6.1 Analyse des entretiens

Cette section résume les réponses fournies par les répondants aux questions listées dans le guide d'entretien (Annexe 5).

[32] Un document contenant une analyse plus détaillée des réponses a été remis à la mandante.

6.1.1 Formation et guide d'utilisation

La majorité des répondants n'ont pas eu besoin du guide d'utilisation de Yareta pour déposer leurs jeux de données, et il semblerait donc que Yareta soit simple et intuitif à utiliser.

6.1.2 Responsabilité

Les données sont préparées en amont par la personne en charge de la publication, mais la responsabilité de déposer le jeu de données est toujours attribuée à la même personne, qui peut être soit le chef de groupe, soit une personne du groupe. Au sein d'un groupe, il n'y a donc actuellement qu'une seule et unique personne qui dépose sur Yareta tous les jeux de données de toutes les publications de son groupe de recherche. Quelquefois, cette personne vérifie aléatoirement quelques fichiers avant de déposer le jeu de données.

Dans les dépôts d'un même groupe de recherche, on remarque une grande cohérence au niveau des noms des dossiers / fichiers et de l'arborescence, certainement dû au fait que ces dépôts sont gérés par une seule personne.

6.1.3 Règles établies dans le groupe et convention pour déposer

Il n'y a pas à proprement parlé de convention ou de règle parmi les groupes interrogés mais les fichiers sont nommés la plupart du temps en fonction de la publication (Fig1, Fig2, …). De plus, on remarque une cohérence dans les jeux de données déposés par un même groupe (arborescence, noms des fichiers, titre et description du jeu de données) dû au fait que les dépôts sont faits par une seule et même personne.

6.1.4 A quelle étape du projet se fait le dépôt

Tous les groupes interrogés déposent leurs données avant ou au moment de la publication (en général, le dépôt est initié lorsque l'article est en phase finale de revue).

- Soit le dépôt est validé lorsque l'article devient public.
- Soit le dépôt est validé lorsque l'article est approuvé, et le chercheur met une période d'embargo pour que les données deviennent publiques au moment de la publication de l'article.

Deux répondants n'ont pas réussi à modifier le niveau d'accès au moment de la publication de l'article, et ceci explique pourquoi leurs dépôts, bien qu'ayant des licences CC, étaient en accès restreint ou fermé sur Yareta (section 5.2.4).

6.1.5 Activités avant le dépôt

Les données sont toujours centralisées avant d'être déposées sur Yareta, et les tâches faites avant le dépôt sont les suivantes :

- créer l'arborescence
- nommer les fichiers
- transformer en format non-propriétaire, si nécessaire

Ces activités se font parallèlement à la rédaction de l'article et ne semblent pas être chronophage ou problématique pour le chercheur.

6.1.6 Format

Dans la grande majorité des groupes, les données générées sont déjà en format non-propriétaire, ou peuvent être converties en format non-propriétaire sans perte d'information.

Deux groupes travaillent avec des données en format propriétaire qui ne peuvent être converties sans perte d'information. Un des groupes conserve les données sous deux formats (format propriétaire et format ouvert). L'autre groupe conserve un format propriétaire pour des données générées par des instruments de mesure et pour lesquelles il est essentiel de garder le format d'origine si on veut conserver et diffuser des données riches. Ces données ne sont donc pas converties en format ouvert mais l'accessibilité au grand public à ces données est possible car le logiciel permettant de lire ces données est téléchargeable et utilisable gratuitement pendant une période limitée.

6.1.7 Licence

Les répondants concernés par l'utilisation de licences différentes (voir section 5.2.4) n'ont pas pu expliquer leurs choix de licence, et une grande majorité des répondants ne comprennent pas les différences entre les licences CC et choisissent une licence au hasard.

6.1.8 README

Il est important de rappeler que tous les groupes interrogés ont déposé des données qui soutenaient une publication.

La plupart du temps, il n'existe pas à proprement parlé de README car les répondants considèrent que leurs données sont soit auto-explicatives (« *self-explaining* ») et qu'il n'y a pas besoin d'information additionnelle à la publication pour les comprendre, soit les documents possèdent un en-tête expliquant le contenu. Ils considèrent qu'un chercheur travaillant dans le même domaine n'aura aucune peine à comprendre leurs données.

Ils estiment que la structure de l'arborescence ainsi que le nom des dossiers et des fichiers permettent de comprendre le contenu du fichier et par là-même la donnée. Le cas échéant, la publication ou le « S*upplementary Information* » remplacent le README.

Quelques groupes fournissent des informations complémentaires sous la forme de :

- Document textuel mentionnant l'arborescence du jeu de données
- Document textuel donnant des détails sur les données (e.g. paramètres appliqués, procédures à suivre pour obtenir le graphe, …)
- Scans de cahier de laboratoire : ce groupe utilise des cahiers de laboratoire manuscrit car il n'existe pas d'ELN capable de répondre complètement à leurs besoins

6.1.9 Objectif de Yareta

Tous les répondants déposent leurs données principalement pour répondre aux exigences du FNS. Mais ils considèrent que Yareta est utile pour s'assurer que les données ne seront pas perdues.

La majorité des répondants n'ont ni le besoin, ni l'habitude de réutiliser des données d'autres groupes pour leurs propres recherches. Cependant, tous sont d'accord de mettre leurs données à disposition d'autres personnes, et tous le faisaient déjà avant cette exigence du FNS, en fournissant leurs données sur demande.

6.1.10 Besoins spécifiques

Les répondants n'ont pas exprimé de besoins spécifiques si ce n'est pour déposer des dépôts lourds.

6.1.11 ELN

Aucun des groupes interrogés n'utilisent d'ELN de manière systématique. L'utilisation d'ELN pour faciliter la préparation des données avant le dépôt sur Yareta semble donc peu appropriée à l'heure actuelle.

6.2 Synthèse de l'analyse des entretiens

A partir de cette analyse des entretiens, plusieurs problématiques ont pu être identifiées, et des propositions d'amélioration sont proposées pour chaque problématique (voir Tableau 11). Ces propositions sont à inclure soit dans les livrables de ce travail (guide d'archivage ou formation) ou dans le guide d'utilisation de Yareta[33], soit à implémenter directement dans l'outil Yareta pour autant que cela soit techniquement possible.

L'information la plus intéressante résultant de ces entretiens est celle concernant la responsabilité du dépôt. En effet, nous nous attendions à ce que la personne en charge de la publication soit également en charge de déposer les données sur Yareta, et qu'il y ait donc plusieurs personnes par groupe de recherche impliquées dans l'utilisation de Yareta. Le fait que la responsabilité de tous les dépôts soit donnée à une seule personne par groupe de recherche nous permettra de mieux cibler les formations : formation dédiée aux personnes responsables des dépôts et centrée sur l'outil Yareta, formation par domaine de recherche ou type de données. Une autre information intéressante est la difficulté qu'ont les chercheurs à comprendre les différentes licences CC.

Tableau 11 : Identification des problématiques et proposition d'amélioration

Responsabilité	
Situation / Problématique	Une seule personne responsable du dépôt au sein du groupe
Proposition	Nommer cette personne « personne référente » pour le groupe de recherche et s'assurer qu'elle soit formée aux bonnes pratiques de préservation Grouper ces personnes en fonction de leur domaine de recherche et leur proposer des formations spécifiques à leur type de données Proposer à cette personne une formation plus approfondie sur Yareta et les outils de gestion des données
Support	Formation
Etape du projet	
Situation / Problématique	Impossible de modifier le niveau d'accès après la publication de l'article

[33] https://yareta.unige.ch/doc/Yareta-QuickStartGuide.html#metadata

Proposition	Expliquer la procédure pour faire cette modification, ou expliciter / simplifier cette étape sur Yareta*
Support	Guide d'utilisation de Yareta, atelier Yareta, éventuellement Yareta
Format	
Situation / Problématique	Fichier d'origine en format propriétaire et non-ouvert (ex : instrument de mesure)
Proposition	Recommander de convertir dans un format ouvert, ou dans le format d'origine et dans un format ouvert
Support	Guide d'archivage, formation
Licence	
Situation / Problématique	Licence choisie au hasard
Proposition	Expliquer les licences. Sur Yareta, remplacer le choix de la licence par liste déroulante par un questionnaire oui/non sur les caractéristiques des licences (e.g. usage commerciale, modification des données, …)*
Support	Guide d'archivage, formation, éventuellement Yareta
README	
Situation / Problématique	Données insuffisamment explicites pour se passer d'un README
Proposition	Sensibiliser le chercheur aux informations nécessaires pour comprendre une donnée (définition des variables et des acronymes, …)
Support	Guide d'archivage, formation
Objectif de Yareta	
Situation / Problématique	Uniquement pour répondre aux exigences du FNS
Proposition	Sensibiliser le chercheur aux avantages qu'il peut avoir en rendant ses données accessibles (augmentation de sa visibilité et de collaboration, …)
Support	Formation

* pour autant que cela soit techniquement possible

7. Présentation et analyse des livrables

Nous avions décidé de produire deux livrables :

- Un guide d'archivage sur la qualité d'une donnée
- Des scénarios de formation

Pour ces deux livrables, je me suis inspirée de plusieurs sources, en privilégiant toujours les informations et les recommandations proposées par l'Université de Genève, et ce afin que ce guide puisse s'intégrer dans les ressources déjà proposées par cette institution.

Deux sections ont été rajoutées : une section contenant les modifications à implémenter directement dans Yareta, pour autant que cela soit techniquement possible, et une section contenant des modifications à faire sur des pages web.

7.1 Guide d'archivage sur la qualité d'une donnée

Pour qu'une donnée soit d'une qualité suffisamment bonne pour pouvoir être réutilisée, il faut qu'on puisse la trouver, y accéder (ouvrir le fichier), la comprendre, et qu'elle soit sous une licence nous permettant de la réutiliser. Ce guide a donc été divisé en quatre sections différentes portant chacune sur un de ces aspects :

- Pour trouver une donnée : lien à la publication et métadonnées
- Pour accéder à une donnée : format
- Pour comprendre une donnée : contexte
- Pour choisir la licence adéquate: licence

Lors des entretiens, nous nous sommes rendu compte que les chercheurs préparaient leurs données parallèlement à la rédaction de la publication, et les déposaient ensuite sur Yareta au moment (ou juste avant) l'approbation de la publication. Nous avons suivi cette même chronologie dans le guide d'archivage :

- Activités avant de dépôt sur Yareta
  - Choisir le format : Format
  - Contextualiser les données : Arborescence / Convention de nommage / README
- Activités durant le dépôt sur Yareta
  - Remplir les métadonnées sur Yareta : Métadonnée
  - Choisir sa licence sur Yareta : Licence

Une forme synthétique a été privilégiée afin que l'accès à l'information soit la plus rapide et la plus facile possible, tenant compte du fait que l'information mentionnée dans ce guide est déjà accessible dans les ressources mises à la disposition du chercheur, soit via des ateliers et des formations, soit via des pages web. Ce guide est donc un matériel additionnel à d'autres ressources, et non pas un matériel remplaçant les autres ressources mises à disposition par l'Université de Genève.

7.1.1 Format

Sur la page web de l'Université de Genève (Université de Genève [sans date]e), deux sources différentes sont proposées pour les formats : les formats recommandés par le UK Data

Service, et les formats recommandés par la bibliothèque de Stanford. Nous nous sommes limités aux formats recommandés par le UK Data Service, car d'une part ce sont ces formats qui sont inclus dans le texte de l'Université de Genève alors que les formats de la bibliothèque de Stanford sont plutôt mentionnés comme une ressource annexe (liste située sur le panneau latéral droit). D'autre part il y a des différences d'évaluation pour certains formats entre le UK Data Service et la bibliothèque de Stanford. Par exemple, le UK Data Service recommande le format TIFF pour les images et évalue le format GIF comme étant acceptable, alors que la bibliothèque de Stanford évalue ces deux formats de manière identique (tous deux sont des « preferred format »). Il a donc été décidé de se baser sur une seule source d'information, en l'occurrence le UK Data Service, afin de proposer une information cohérente aux chercheurs.

L'ETHZ fait clairement la distinction entre les formats recommandés pour une préservation de dix ans ou moins, et les formats recommandés pour une préservation de plus de dix ans (ETHZ [sans date]a). En effet, il est probable que les formats les plus usuels (y compris des formats propriétaires) seront encore accessibles dans dix ans (UK Data Service [sans date]a). Une donnée devant être préservée 20-30 ans doit donc être traitée différemment qu'une donnée devant être préservée dix ans. En combinant les recommandations du UK Data Service et de l'ETHZ (voir Annexe 6), les formats à privilégier selon la durée de la préservation ont été listés.

7.1.2 Contexte

Les recommandations proposées pour l'arborescence et le nommage des fichiers ont été élaborées en utilisant les informations fournies par l'Université de Genève ainsi que les ressources fournies durant le MOOC proposé par l'université de Delft (Delft University of Technology 2020c).

Le modèle proposé pour le README a été élaboré en combinant les informations fournis par l'Université de Genève (Université de Genève 2019a ; Cornell University [sans date]) et par l'université de Delft (Delft University of Technology 2017).

7.1.3 Métadonnées

La proposition au départ était de lister les métadonnées minimales à fournir par type de discipline (ou type de donnée). Cependant, au vu de la grande diversité des disciplines et de la difficulté à élaborer une liste qui soit suffisamment générale pour s'appliquer à tout type de données mais en même temps suffisamment pertinente pour être utile aux chercheurs, cette proposition a été abandonnée.

L'analyse des dépôts (section 5) a montré que certaines informations n'étaient pas suffisamment précises (titre, description du jeu de données) ou fausses (date de collecte). La section concernant les métadonnées se concentre donc sur les métadonnées à fournir pour Yareta (métadonnées obligatoires et optionnelles) afin de s'assurer que cette information soit de la meilleure qualité possible.

7.1.4 Licence

Les différences entre les licences ne sont pas clairement comprises, or c'est une notion importante à maitriser, surtout dans le cas de données sensibles ou personnelles. Afin de favoriser une meilleure compréhension de cette notion, deux schémas différents sont proposés : l'un sous forme d'un tableau synthétique et l'autre sous forme d'arbre de décision. L'arbre de décision permet au chercheur de choisir la licence en utilisant une approche

séquentielle (réponse oui/non à une question), et devrait normalement lui permettre de mieux comprendre les différences entre les licences.

7.2 Scénario de formation

La plupart des problématiques identifiées durant l'analyse des dépôts de Yareta et les entretiens font déjà l'objet d'une formation, ou ont des pages web mises à la disposition du chercheur (Tableau 12). De plus, l'Université de Genève offre la possibilité d'avoir des formations sur mesures sur la GDR en fonction des besoins des chercheurs, et fournit également des liens sur d'autres programmes de formation dédiés aux données de la recherche (e-learning ou MOOC) (Université de Genève [sans date]d).

Les ressources concernant le format et le nommage des fichiers sont suffisamment spécifiques et détaillées. Il ne me semble donc pas nécessaire d'en créer de nouvelles, mais il faudrait plutôt insister sur les problématiques identifiées durant l'analyse des dépôts et des entretiens lors les formations déjà existantes.

Les ressources concernant les licences sont également déjà existantes, mais comme c'est une notion importante à comprendre (spécialement en cas de données sensibles) et qu'elle est peu comprise par les répondants, il vaudrait la peine de proposer de nouveaux services.

Les informations concernant le README sont dispersées dans plusieurs ressources, et il me semble important de proposer une formation spécifique sur cette thématique. En effet, sans contextualisation de la donnée, il est difficile de la comprendre et sans une bonne compréhension de la donnée, il est impossible de la réutiliser. De plus, afin d'inciter les chercheurs à fournir un README, une recommandation de fournir un README pourrait être ajoutée dans la formation « Publier ses données de recherche avec Yareta » ainsi que dans le guide d'utilisation de Yareta.

Les entretiens ont montré qu'une seule personne est en charge de déposer les données sur Yareta. Il pourrait être intéressant de connaître cette personne et de s'assurer qu'elle a suivi la formation proposée par l'Université de Genève sur la GDR (Université de Genève [sans date]b), qu'elle possède de bonnes pratiques en GDR, et éventuellement de lui fournir une formation sur mesure en fonction de son domaine de recherche ou du type de données qu'elle manipule. Il pourrait également être intéressant de grouper ces personnes responsables des dépôts par domaine de recherche et de leur proposer une formation plus adaptées aux types de données qu'elles doivent gérer, ou une formation plus approfondie sur l'outil Yareta.

Tableau 12 : Propositions de nouvelles ressources

Format	
Problématique	Jeu de données déposé en tant que zip, utilisation de format propriétaire, erreur dans l'extension du fichier
Ressources déjà existantes	Formation sur mesure (Université de Genève [sans date]d) ; formation « Stockage des données » (Université de Genève [sans date]b) ; page web (Université de Genève [sans date]e)

Nouvelles ressources à mettre en place	Non
Licence	
Problématique	Licence choisie au hasard
Ressources déjà existantes	Formation sur mesure (Université de Genève [sans date]d) ; formation « Diffuser ses données » (Université de Genève 2019b) ; page web (Université de Genève [sans date]c)
Nouvelles ressources à mettre en place	Oui, et partiellement fait (une nouvelle formation de 15 minutes sera proposée dès l'automne 2020[34])
README	
Problématique	Donnée insuffisamment contextualisée pour permettre la réutilisation
Ressources déjà existantes	Formation sur mesure (Université de Genève [sans date]d) ; formation « Diffuser ses données » (Université de Genève 2019b) ; formation « Comment remplir le DMP du FNS » (Université de Genève 2019a)
Nouvelles ressources à mettre en place	Oui, proposer une formation sur le thème « Rendre ses données compréhensibles pour permettre leur réutilisation » ; ajouter une recommandation de fournir un README dans la formation et le guide d'utilisation de Yareta ; ajouter un onglet spécifique README sur la page web « données de recherche – collecter & organiser »
Nommage des fichiers	
Problématique	Intitulé pas assez clair
Ressources déjà existantes	Formation sur mesure (Université de Genève [sans date]d) ; formation « Organisation et nommage des fichiers de données de recherche » (Université de Genève [sans date]b) ; page web (Université de Genève [sans date]b)
Nouvelles ressources à mettre en place	Non
Responsabilité du dépôt	
Problématique	Une seule personne responsable du dépôt au sein du groupe

[34] Communication interne

Ressources déjà existantes	Non
Nouvelles ressources à mettre en place	Oui, proposer des formations plus ciblées en fonction du domaine de recherche. Proposer des formations plus ciblées sur l'utilisation de Yareta. S'assurer qu'elle a suivi la formation « Introduction à la gestion des données de recherche » (Université de Genève [sans date]b). Eventuellement lui proposer une formation sur mesure en fonction de ses besoins.
Objectif de Yareta	
Problématique	Répondre aux exigences du FNS
Ressources déjà existantes	Formation « Publier ses données avec Yareta » (Université de Genève 2020) ; page web (Université de Genève [sans date]j)
Nouvelles ressources à mettre en place	Non, mais insister sur les avantages pour le chercheur de rendre ses données accessibles lors des formations déjà existantes

7.3 Ressources fournies par les pages web

Hormis les formations, il y a deux ressources fournies sous la forme de page web qui pourraient être modifiées afin de fournir une information plus cohérente:

- Page web sur les licences
- Page web sur les formats

7.3.1 Licence

Sur Yareta, le chercheur a le choix entre sept licences CC différentes, alors que la page web de l'Université de Genève n'en mentionne que quatre (Tableau 13) (Université de Genève [sans date]c). De plus, la page web de l'Université de Genève propose des licences de types Open Data Commons (ODS) plus adaptées aux données et bases de données et qui ne sont pas spontanément proposées sur Yareta (sur demande, il y a la possibilité de rajouter d'autres licences dans Yareta). Il serait préférable d'harmoniser cette information de telle sorte à ce qu'il y ait une cohérence entre la ressource proposée par l'Université de Genève et les informations mentionnées sur Yareta.

Tableau 13 : Licences mentionnées par l'Université de Genève et par Yareta

Licence	Page web de l'Université de Genève	Yareta
Licence CC	CC BY CC BY-SA CC BY-ND CC BY-NC	CC BY CC BY-SA CC BY-ND CC BY-NC CC0 CC BY-NC-ND CC BY-NC-SA

Licence	Page web de l'Université de Genève	Yareta
Licence ODS	ODC Public Domain and Dedication License (PDDL) ODC Attribution License Open Database License (ODbL)	Sur demande

La proposition serait donc :

- de rajouter sur les pages web de l'Université de Genève les licences CCO, CC BY-NC-ND et CC BY-NC-SA
- de rajouter sur Yareta les licences ODS

7.3.2 Format

Concernant la page web sur les formats (Université de Genève [sans date]e) et comme mentionné dans la section 7.1.1, il serait préférable de n'avoir qu'une seule source pour les formats recommandés plutôt que deux sources différentes. En l'occurrence, les recommandations provenant du UK Data Service devrait être privilégiées.

Dans un deuxième temps, il pourrait être intéressant de s'assurer que tous les formats utilisés par les chercheurs de l'Université de Genève soient mentionnés sur cette page web, afin que le chercheur puisse trouver toute l'information dont il a besoin. En l'occurrence, des trois formats non listés par le UK Data Service et mentionnés dans les métadonnées METS (DNG, MATLAB, DICOM, voir section 5.2.3), seul le format DNG est présent dans le tableau de l'ETHZ (recommandé pour des données brutes en plus d'une copie en format TIFF).

7.4 Modifications à implémenter dans Yareta

Si c'est techniquement possible, les modifications suivantes pourraient être directement implémentées dans Yareta (Tableau 14):

Tableau 14 : Propositions de modification dans Yareta

Lien à la publication	
Actuellement dans Yareta	-
proposition	Ajouter un champ « lien à une publication » (métadonnée obligatoire). Si la réponse est « oui », apparition d'un champ « Publication » (métadonnée obligatoire).
Licence	
Actuellement dans Yareta	Licence nommée par leur nom complet (e.g. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International)
proposition	Nommer les licences par leurs acronymes (e.g. CC BY-NC-SA), et supprimer la mention « International » dans les licences

Licence	
Actuellement dans Yareta	Licence CC
proposition	Rajouter les licences ODS
Licence	
Actuellement dans Yareta	Liste déroulante proposant les différentes licences
proposition	Choix de la licence par question oui/non, sur le modèle de l'arbre de décision proposé dans le guide d'archivage
Modification du niveau d'accès	
Actuellement dans Yareta	-
proposition	Simplifier la procédure permettant le changement du niveau d'accès
Erreur dans l'extension d'un fichier	
Actuellement dans Yareta	-
proposition	Non-validation du dépôt et envoi d'un message d'erreur

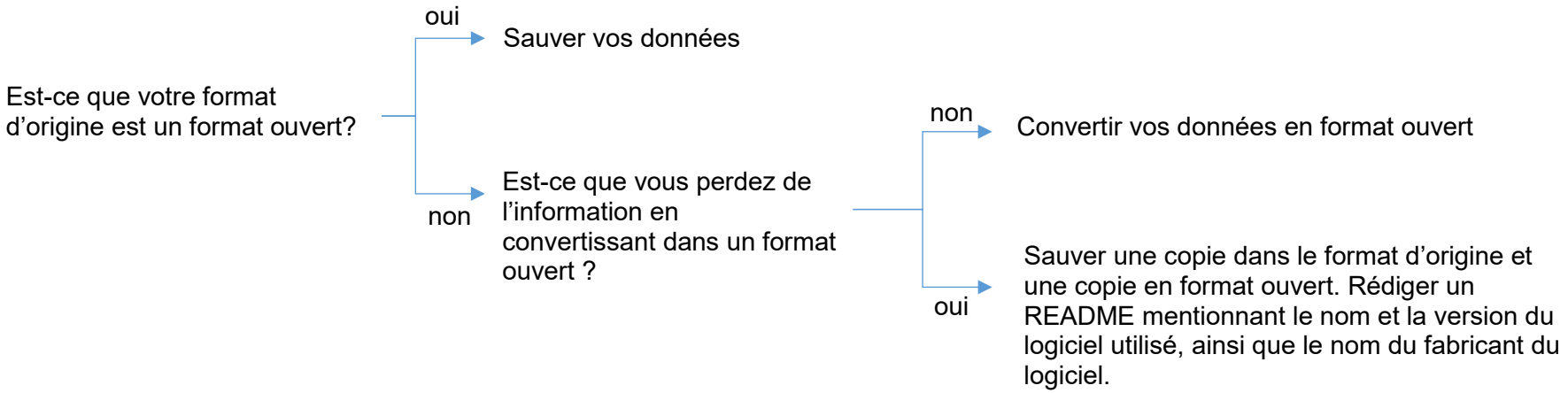
7.5 Livrable - Guide d'archivage

1. Format

But : permettre la conservation et la réutilisation des données

Bonne pratique : privilégier un format qui soit non-propriétaire, non crypté, non compressé et couramment utilisé dans votre domaine de recherche.

Si vous devez préserver vos données plus de 10 ans, assurez-vous de les sauver dans un format ouvert



Quel format utiliser pour quel type de donnée ?

Type de donnée	Format recommandé	Format acceptable	Format recommandé pour une préservation > 10 ans	Format adapté pour une préservation ≤ 10 ans
Tabular data with extensive metadata (variable labels, code labels, and defined missing values)	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file	proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)		
Tabular data with minimal metadata (column headings, variable names)	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements	delimited text (.txt) with characters not present in data used as delimiters widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)	.csv .tab	.xlsx .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) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti	Plain text, ASCII (.txt) .xml	.rtf .html .docx
Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif)	*.tif (TIFF 6.0)	.jpeg, .jpg, .jp2 .gif

Type de donnée	Format recommandé	Format acceptable	Format recommandé pour une préservation > 10 ans	Format adapté pour une préservation ≤ 10 ans
		TIFF other versions (.tif, .tiff) RAW image format (.raw) Photoshop files (.psd) BMP (.bmp) PNG (.png) (mais pas compressé d'après ETHZ) Adobe Portable Document Format (PDF/A, PDF) (.pdf)	PNG uncompressed (.png)	TIFF other versions (.tif, .tiff) .bmp
Audio data	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav)	.wav	.mp3
Video data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)	AVCHD video (.avchd)		.mp4 .mj2
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) widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0	PDF/A (.pdf) Plain text (.txt) .xhtml .html .xml	PDF (.pdf) .rtf

2. Contexte

But : Rendre vos données compréhensibles et réutilisables

2.1 Arborescence et nommage des fichiers

But : permet d'identifier facilement vos données et de les comprendre, pour vous et pour les autres

Bonne pratique : mettre en place une convention de nommage des fichiers au sein de votre groupe, la documenter, et l'utiliser de manière constante

Arborescence	Nommage des fichiers
Hiérarchie simple	Identifiant court et unique
Dossier général > dossier spécifique	Inclure un résumé du contenu
Structure simple et logique	Utiliser _ pour délimiter entre différents éléments
Eviter les dossiers trop remplis et les structures trop profondes	Utiliser - pour délimiter au sein d'un même élément
	Exemple : 20090915_Smith-John_AIG _POLICY.pdf
	Indiquer les dates en format YYYYMMDD
	Limiter les noms à 32 caractères
	Garder une trace des versions des documents
	Trouver la bonne balance entre trop et trop peu

2.2 Fournir un README

But : permettre de comprendre les données et la manière dont elles ont été générées afin de pouvoir les répliquer ou les réutiliser

Bonne pratique :

- Créer un README pour chaque jeu de données
- Utiliser un format *plain text* (.txt) ou .pdf si le formatage est important
- Nommer le fichier README
- Si possible, déposer le fichier README dans le dossier « documentation » de Yareta. Si cela n'est pas possible, déposer un document .txt dans le dossier « documentation » listant les fichiers du jeu de données (arborescence) et permettant de localiser le(s) fichier(s) README dans l'arborescence.

Modèle de README

Les informations minimales à fournir sont indiquées par une *

Information générale	• Titre du jeu de données* • Brève description des données* • Nom des chercheurs ayant participé à l'étude* • Convention de nommage des fichiers • Format des fichiers • Description de l'arborescence du dossier et des données qui se trouvent dans chaque sous-dossier • Relations entre les fichiers, ou la structure des fichiers • Information de contact • Information concernant la source de financement
Méthodologie	• Méthodes employées pour collecter les données et les analyser* • Date de collecte des données* • Définition des codes ou symboles utilisés* • Information spécifique aux appareils utilisés • Information concernant les logiciels utilisés (y compris le numéro de version) • Information concernant les standards utilisés et la calibration • Information concernant la localisation géographique de la collecte des données
Résultats / Données	• Définition des intitulés* • Date de création du fichier* • Mention des unités de mesure* • Définition des codes et des symboles utilisés pour les données manquantes* • Définition des abréviations utilisées* • Mention des formats spécifiques* • Date à laquelle le fichier a été modifié et description des modifications
Partage et niveau d'accès	• Mention des conditions d'accès aux données (licence), durant l'étude et après celle-ci* • Lien à la publication liée à ce jeu de données • Manière de citer le jeu de données

3. Compléter les métadonnées

But : permettre l'identification et la réutilisation des données

Bonnes pratiques :

- utiliser en priorité le standard de métadonnées spécifique à votre discipline (https://www.dcc.ac.uk/guidance/standards/metadata)
- Déposer vos données en priorité sur un dépôt spécifique à votre communauté de recherche, et si un tel dépôt n'existe pas, utiliser Yareta (Université de Genève 2020)

Yareta - Métadonnées obligatoires

	Quelles sont les informations à fournir
Titre	Si les données soutiennent une publication : titre de l'article Si les données ne soutiennent pas de publication : toute information permettant l'identification du jeu de données (e.g. objet de la recherche, type de données, lieu / date de collecte)
Description	Les informations permettant de comprendre l'objectif et le résultat de la recherche (e.g. type de données, « objet » de la donnée, éventuellement date de la collecte des données, mention de la publication si les données sont liées à une publication)
Date de publication	Date à partir de laquelle le dépôt doit être disponible en ligne de manière publique Format : dd/mm/yyyy
Contributeur	Mettre le numéro OrcID si disponible
Niveau d'accès	Public : accès ouvert Restreint : accès limité aux membres de l'unité organisationnelle Fermé : accès personnalisé (en cours d'implémentation)

Yareta - Métadonnées facultatives

	Quelles sont les informations à fournir
Mots clés	Mots clés en lien avec le jeu de données. Si votre jeu de données soutient une publication, reprenez les mots clés mentionnés dans la publication
Début de la collecte des données	Indiquer la date à laquelle vous avez commencé à collecter/générer les données Format : dd/mm/yyyy
Fin de la collecte des données	Indiquer la date à laquelle vous avez terminé de collecter/générer les données Format : dd/mm/yyyy
Embargo	Période pendant laquelle vos données doivent rester en accès restreint / fermé. Format : xx mois
Licence	Voir la section 4 de ce guide « choisir sa licence »

4. Choisir sa licence

But : permettre le partage et la réutilisation de vos données

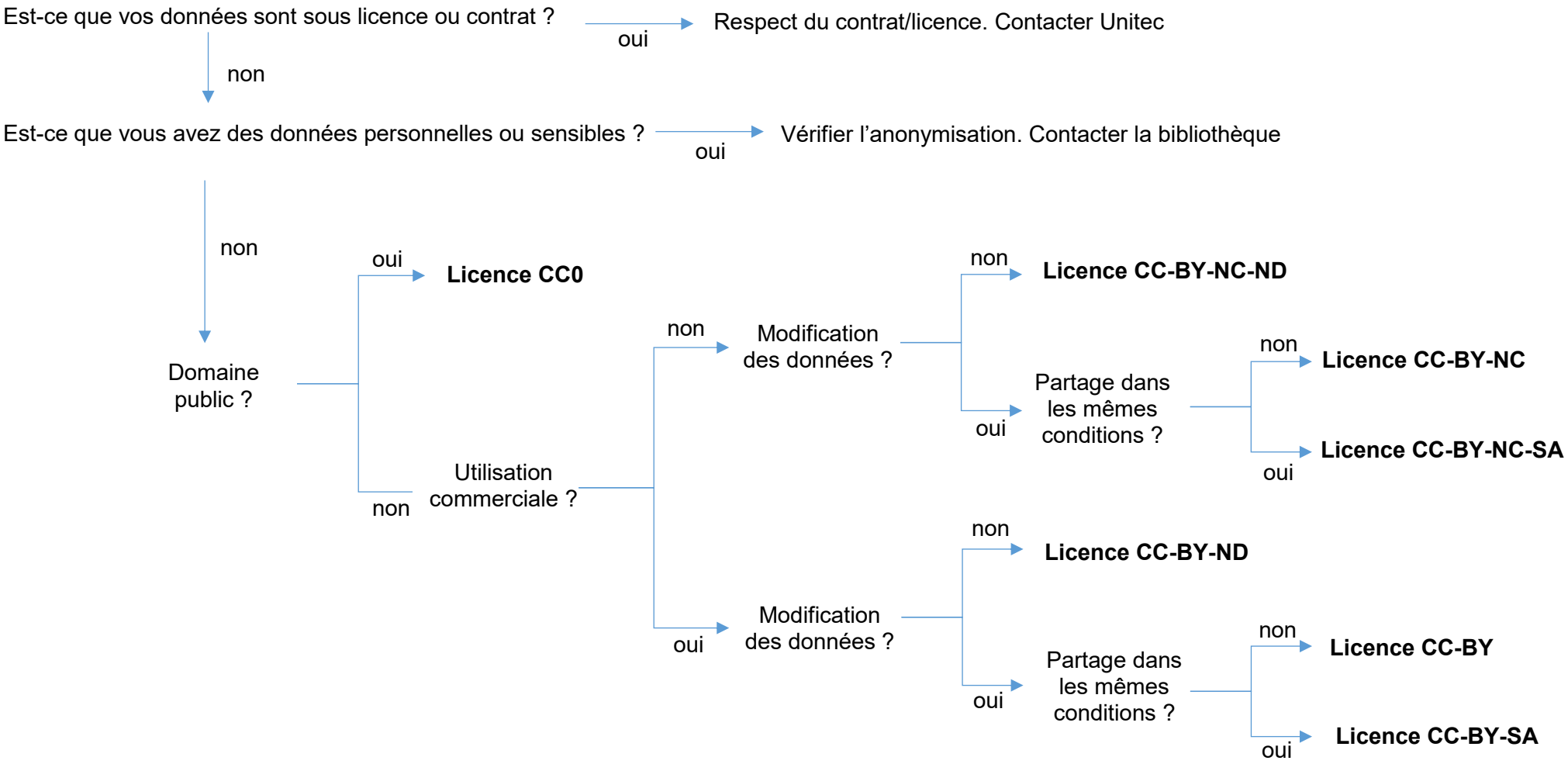
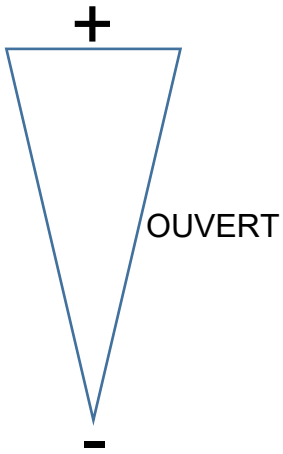


Tableau résumant les caractéristiques des différentes licences

Licence	Partage	Attribution	Utilisation commerciale	Modification	Partage à l'identique
	CC	BY	NC	ND	SA
CC0					
CC-BY					
CC-BY-SA					
CC-BY-NC					
CC-BY-NC-SA					
CC-BY-ND					
CC-BY-NC-ND					



Références section 1 – Format

- Formats de fichier - Université de Genève (https://www.unige.ch/researchdata/fr/preserver/all/formats-fichier/)
- File formats for archiving – ETHZ (https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving)
- Standard formats for long term access – UK Data Service (https://www.ukdataservice.ac.uk/manage-data/format/file-formats.aspx)
- Best practices for file formats - Université de Stanford (https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats)

Références section 2 - Contexte

- Nommer et organiser vos fichiers – Université de Genève (https://www.unige.ch/researchdata/fr/collecter-organiser/all/nommer-et-organiser-vos-fichiers/)
- Organisation et nommage des fichiers de données de recherche – Université de Genève (https://www.unige.ch/biblio/files/2615/8773/8108/2020_rdv_nommage_presentation.pdf)
- File Naming Conventions - Purdue University (https://guides.lib.purdue.edu/c.php?g=353013&p=2378293)
- Folder and file naming convention – 10 rules for Best Practices - eXadox (http://www.exadox.com/en/articles/file-naming-convention-ten-rules-best-practice)
- Comment remplir le DMP du FNS – Université de Genève (https://www.unige.ch/researchdata/files/4115/6872/2865/201909_atelier_DMP_FNS_presentation_fre.pdf)
- Guide to writing « readme » style metadata – Cornell University (https://data.research.cornell.edu/content/readme)

- Guideline for creating a REDME file – Delft Universtiy (http://resolver.tudelft.nl/uuid:66e210e1-c884-45d6-b9d4-711907680452)

Références section 3 – Compléter les métadonnées

- Université de Genève - https://www.unige.ch/researchdata/fr/collecter-organiser/all/creer-metadonnees/
- Université de Genève - Publier ses données avec Yareta (https://www.unige.ch/researchdata/files/3115/8860/8052/202004_atelier_Yareta_fr.pdf)
- Plateforme Yareta test
- DDC - Disciplinary metadata (https://www.dcc.ac.uk/guidance/standards/metadata)
- Université de Lausanne – Readme file (https://www.unil.ch/openscience/home/menuinst/open-research-data/gerer-ses-donnees-de-recherche/stockage--securite.html)

Références section 4 - Licence :

- Université de Genève – arbre de décision (https://www.unige.ch/researchdata/files/7115/1929/2646/2018_decision_tree_UNIGE.pdf)
- How To Attribute Creative Commons Photos (Foter, 2015)
- Gestion des données de recherche (EduTech Wiki, 2020)
- Choose a licence (version beta en test) (Creative Commons, [sans date]b)

7.6 Livrables – Scénarios de formation et ressources

README :

- Proposer une nouvelle formation sous la forme d'un atelier d'une heure en présentiel
- Proposer un onglet spécifique à cette thématique sur la page web de l'Université de Genève « Données de recherche - Collecter et organiser »

Licence :

- Proposer une nouvelle formation sous la forme d'une présentation de 15 minutes en ligne (en cours d'implémentation)
- Harmoniser l'information entre la page web de l'Université de Genève et Yareta

Format :

- Ne garder que le UK Data Service pour les formats recommandés
- S'assurer que tous les formats mentionnés dans les dépôts de Yareta soient listés sur cette page

Objectif de Yareta :

- Insister durant les formations existantes sur les avantages pour le chercheur de rendre ses données accessibles

Personne en charge du dépôt :

- Lui proposer de suivre la formation « Introduction à la gestion des données de recherche »
- Lui proposer une formation sur mesure adaptée à ses besoins

8. Discussion

Deux aspects seront principalement abordés dans cette discussion : l'accessibilité au grand public et l'adaptation du modèle de l'université de Delft (section 4.3) à l'Université de Genève.

Un public pour les données de recherche

Le mouvement de l'Open Data demande de rendre accessible au grand public les données financées par l'argent public (OCDE 2007 ; FNS [sans date]e). Cependant, le besoin ou l'intérêt pour le grand public d'accéder et de réutiliser ces données de recherche est sujet à questionnement (Académie des sciences naturelles, 2018 ; Faniel, Zimmerman, 2011). En l'occurrence, pour que les données soient accessibles au grand public il faudrait jouer sur trois contraintes.

Une contrainte technique : selon la taille du dépôt, il n'est pas toujours possible d'accéder aux données. Pour ma part, je ne pouvais pas télécharger des dépôts de plus de 1GB avec mon matériel informatique. Si le but est de permettre à tout un chacun d'accéder à des données de recherche, il faudrait que le téléchargement puisse se faire en plusieurs fois, ou alors que l'usager puisse visualiser les données avant de les télécharger, et choisir de télécharger uniquement les données qui l'intéressent. De même, la taille des fichiers peut également être problématique pour une personne possédant un matériel informatique standard car il peut être difficile de travailler sur un très gros fichier.

Une contrainte de temps : enrichir une donnée avec des métadonnées adéquates, rédiger un README, s'assurer que ses données soient dans un format ouvert, toutes ces activités sont essentielles si on veut qu'une donnée soit réutilisable, en particulier par le grand public. En effet, un chercheur du même domaine pourrait avoir accès à des logiciels couramment utilisés dans son domaine et contourner la contrainte du format ouvert. De même, ayant l'habitude de manipuler des données identiques, il sera à même de les comprendre sans avoir besoin d'une grande contextualisation. Cependant, cela ne sera pas possible pour le grand public et le chercheur sera obligé de consacrer du temps à ces activités.

Une contrainte de vulgarisation : la plupart des chercheurs interrogés partent du principe que leurs données sont compréhensibles pour une personne travaillant dans le même domaine, et que la publication ou le *Supplementary Information* sont suffisants pour permettre la réutilisation d'une donnée. Cela signifie qu'il y aurait un effort supplémentaire à faire pour rendre ses données compréhensibles pour le grand public.

Les activités concernant la GDR sont considérées comme peu intéressantes par les chercheurs et peu valorisantes pour les jeunes chercheurs (Académie des sciences naturelles 2018). Avant de contraindre les chercheurs à investir du temps dans ces activités, il serait judicieux de savoir si le besoin existe vraiment dans le grand public de vouloir réutiliser des données de recherche. Si ce besoin n'existe pas, il faudrait d'abord faire en sorte que les données soient réutilisables par les chercheurs eux-mêmes, et dans un deuxième temps, étendre cette réutilisation au grand public. Cette approche permettrait d'augmenter la motivation du chercheur à investir du temps dans des activités de GDR car cela donnerait du sens à ces activités. Et lorsque la pratique sera implémentée, on pourrait étendre cette réutilisation au grand public. De ce point de vue, il serait intéressant de faire des statistiques

sur la réutilisation des données déposées dans Yareta (fréquence de réutilisation et type d'utilisateur) afin de savoir qui réutilise ces données et à quelle fréquence.

Deux profils pour les données de recherche

Un des points qui est ressorti des entretiens est le fait qu'une seule personne soit en charge de déposer les données au sein du groupe. Au-delà de s'assurer que cette personne soit bien formée en GDR, nous pourrions transposer le modèle mis en place à l'université de Delft et proposer à cette personne de prendre un rôle de « Data Champion ». On pourrait également proposer que ces activités concernant les dépôts soient reconnues par l'institution, et qu'elles soient clairement mentionnées dans son cahier des charges à un pourcentage en adéquation avec cette tâche (par exemple à hauteur de 5%). De ce fait, cette activité ne serait pas « quelque chose à faire en plus », mais serait incluse dans son poste, et pourrait contrebalancer le manque de reconnaissance professionnelle souvent ressenti par les chercheurs (Plomp et al. 2019). Sachant qu'il existe déjà une communauté de « Data Champion » à l'EPFL, des contacts pourraient être pris pour profiter de leur expérience, ou pour directement intégrer les « Data Champion » de l'Université de Genève à leur communauté de « Data Champion ». Le cas échéant, si aucune collaboration n'est possible, nous pourrions utiliser le guide rédigé par l'université de Delft détaillant les six étapes pour créer une communauté de « Data Champion » (Clare, 2019).

Il me semble également important d'inclure un « Data Steward ». En effet, en partant de la pratique du chercheur, il est plus facile de se rendre compte de ce qui pose problème au chercheur, ou de ce qui peut être amélioré, alors que durant une formation théorique, ces problèmes peuvent tout à fait passer inaperçu. De plus, même si on ne comprend pas les données, le fait de télécharger les dépôts et de naviguer parmi les données nous donne des amorces pour ensuite aller discuter avec les chercheurs. En effet, si on se rend compte que le chercheur a utilisé trois licences différentes pour ses trois dépôts, il est aisé de lui demander la raison de ses choix, et de se rendre compte que les licences ont été choisies au hasard. De même, le fait d'être capable ou non d'ouvrir un fichier peut être une porte d'entrée pour discuter des formats ouverts. Finalement, l'arborescence ou les noms de fichiers sont rapidement visibles en téléchargeant un dépôt alors que durant une formation, il est probable que tous les chercheurs auront l'impression de maîtriser ces notions.

Le Data Steward pourrait avoir le même profil que celui proposé par l'université de Delft mais il serait préférable de l'intégrer dans la structure de la bibliothèque, et non pas dans les facultés car les formations et le soutien en GDR est actuellement sous la responsabilité de la bibliothèque. Je proposerai également que le cahier des charges de ce Data Steward soit réparti entre formation, service de soutien personnalisé et revue des dépôts de Yareta. La plupart des dépôts d'un même groupe sont très homogène (et donc un problème rencontré dans un dépôt se retrouvera certainement dans les autres dépôts du même groupe), il ne sera pas forcément nécessaire d'analyser chaque dépôt en détails, et consacrer 20% de son temps à faire cette revue devrait être suffisant. De plus, le fait que Yareta soit fait selon une méthode agile permettrait vraiment d'utiliser tout le potentiel d'un « Data Steward » puisque les améliorations ou changements découlant de la revue des dépôts pourraient être rapidement implémentés dans l'outil Yareta.

9. Conclusion

L'analyse des dépôts et les entretiens effectués durant ce travail ont permis de mettre en évidence trois points importants :

- la difficulté à réutiliser les jeux de données
- un manque de compréhension des licences CC
- la responsabilité du dépôt qui incombe à une seule personne dans le groupe

A partir de ce constat, les propositions suivantes ont été faites:

- élaboration d'un guide d'archivage portant sur quatre activités permettant de garantir une bonne préservation : format, contexte, métadonnées, licence
- Nouvelle formation sur le README
- Nouvelle ressource sur le README sous la forme de page web
- Nouvelles formations ciblées pour les personnes en charge du dépôt
- Modification de la page web sur les licences
- Modification de la page web sur les formats
- Modification / ajout d'information dans l'outil Yareta

Un des objectifs de ce travail était d'analyser la possibilité de réutiliser les jeux de données déposés sur Yareta, et l'analyse a montré qu'il y avait peu de dépôts réutilisables pour une personne du grand public. L'autre objectif était de fournir des services de soutien adaptés aux chercheurs, et des propositions ont été faites dans ce sens. De plus, une formation a déjà été mise en place par l'Université de Genève suite à ce travail (formation de 15 minutes sur les licences).

Bien que ce travail ait été restreint au site Uni Arve de l'Université de Genève, certaines des propositions (modification des pages web, modification dans l'outil Yareta) pourront être utilisées par toute la communauté de l'Université de Genève.

Un des apports de ce travail réside dans son approche pratique de la préservation (i.e. l'analyse des jeux de données déposés sur Yareta). En effet, cette approche donne une amorce pour discuter avec les chercheurs, et facilite par là-même l'échange et la qualité de l'information qui est transmise. De plus, cette approche permet également d'agir directement sur l'outil Yareta, et cet outil étant fait selon une méthode agile, les propositions de changement peuvent être discutées et implémentées rapidement. De même, sachant que Yareta en est à ses débuts, c'est le bon moment pour corriger ou améliorer les pratiques des chercheurs. Finalement, ce travail peut servir d'argument pour développer une position de « Data Steward », qui serait à mon sens le moyen le plus adéquat pour assurer une bonne qualité aux jeux de données.

Bibliographie

ACADÉMIE DES SCIENCES NATURELLES, 2018. *Open Data and Data Management – Issues and Challenges* [en ligne]. 29 octobre 2018. S.l. : s.n. [Consulté le 11 août 2020]. Disponible à l'adresse : https://sciencesnaturelles.ch/uuid/74106f4d-8a87-5134-8d0d-dc9478bda547?r=20200527115808_1565134741_f31eeb12-74e0-54a2-b008-713f18211549

AUSTRALIAN NATIONAL DATA SERVICE (ANDS), 2017a. *ANDS Guide - File Formats* [en ligne]. 10 janvier 2017. [Consulté le 31 mai 2020]. Disponible à l'adresse : www.ands.org.au/guides/file-formats

AUSTRALIAN NATIONAL DATA SERVICE (ANDS), 2017b. *What is research data* [en ligne]. 11 janvier 2017. [Consulté le 31 mai 2020]. Disponible à l'adresse : https://www.ands.org.au/__data/assets/pdf_file/0006/731823/Whatis-research-data.pdf

BAGNOUD, Gérard, 2016. *Archives des savoirs : De la gestion des données de recherche vers une gestion des données pour la recherche*. 23 février 2016 [en ligne]. [Consulté le 31 mai 2020]. Disponible à l'adresse : https://www.researchgate.net/publication/295672110_Archives_des_savoirs_De_la_gestion_des_donnees_de_recherche_vers_une_gestion_des_donnees_pour_la_recherche

BARI, Manon, BEZZI, Manuela, GUIRLET, Marielle et MAKHLOUF-SHABOU, Basma (dir), 2020. *Formation et éducation en gestion des données de recherche du point de vue du projet DLCM: dispositifs d'e-learning* [en ligne]. 19 janvier 2020. [Consulté le 26 juin 2020]. Disponible à l'adresse : http://doc.rero.ch/record/328462. 65

BAYKOUCHEVA, Svetla, 2015. *Managing scientific information and research data*. Chandos Publishing. Amsterdam : Chandos information professional series. ISBN 978-0-08-100195-0

BERLIN DECLARATION, 2003. Berlin declaration on open access to knowledge in the sciences and humanities. Max Planck Open Access [en ligne]. 22 octobre 2003. [Consulté le 25 juillet 2020]. Disponible à l'adresse : https://openaccess.mpg.de/Berlin-Declaration

BLUMER, Eliane et BURGI, Pierre-Yves, 2015. Data Life-Cycle Management Project: SUC P2 2015-2018. *Revue électronique suisse de science de l'information* [en ligne]. 2015. Vol. 16. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://archive-ouverte.unige.ch/unige:79346

BUDAPEST OPEN ACCESS INITIATIVE, 2002. Budapest Open Access Initiative | Read the Budapest Open Access Initiative. [en ligne]. 2002. [Consulté le 16 avril 2020]. Disponible à l'adresse : https://www.budapestopenaccessinitiative.org/read

BURGI, Pierre-Yves, 2015. Data Life-Cycle Management: The Swiss Way. In : *Bulletin de l'Académie suisse des sciences humaines et sociales*. 2015. Vol. 4, p. 48-50

BURGI, Pierre-Yves, 2019. Le Projet de Loi 12146 : Infrastructures et services numériques pour la recherche. *Revue électronique suisse de science de l'information* [en ligne]. 2019. Vol. 20. [Consulté le 4 mars 2020]. Disponible à l'adresse : https://archive-ouverte.unige.ch/unige:128845

BURGI, Pierre-Yves et BLUMER, Eliane, 2018. Le projet DLCM : gestion du cycle de vie des données de recherche en Suisse. *Bibliotheken der Schweiz: Innovation durch Kooperation. Festschrift für Susanna Bliggenstorfer anlässlich ihres Rücktrittes als Direktorin der Zentralbibliothek Zürich* [en ligne]. S.l. : De Gruyter. p. 235-249. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://archive-ouverte.unige.ch/unige:105931

BURGI, Pierre-Yves, BLUMER, Eliane et MAKHLOUF-SHABOU, Basma, 2017. Research data management in Switzerland: National efforts to guarantee the sustainability of research outputs. *IFLA Journal*. 1 mars 2017. Vol. 43, n° 1, p. 5-21

BURGI, Pierre-Yves et CAZEAUX, Hugues, 2019. *Yareta, une nouvelle solution numérique pour archiver et partager vos données de recherche* [en ligne]. 2019. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://www.unige.ch/eresearch/fr/services/yareta/

CCSDS, 2012. Reference Model for an Open Archival Information System (OAIS). [en ligne] juin 2012. p. 135. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://public.ccsds.org/Pubs/650x0m2.pdf

CCSDS, 2017. *Modèle de référence pour un Système ouvert d'archivage d'information (OAIS)* [en ligne]. octobre 2017. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://public.ccsds.org/Pubs/650x0m2(F).pdf

CLARE, Connie, 2019. *The Real World of Research Data* [en ligne]. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://zenodo.org/record/3584373#.Xzl8B-gzbIU

CORNELL UNIVERSITY, [sans date]. Guide to writing « readme » style metadata. *cornell.edu* [en ligne]. [Consulté le 10 juillet 2020]. Disponible à l'adresse : https://data.research.cornell.edu/content/readme

CORTI, Louise, VAN DEN EYNDEN, Veerle, BISHOP, Libby et MORGAN-BRETT, Bethany, 2011. *Managing and sharing data - Training resources* [en ligne]. septembre 2011. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://ukdataservice.ac.uk/media/622416/trainingresourcespack.pdf

CREATIVE COMMONS, [sans date]a. A propos des licences. *creativecommons.org* [en ligne]. [sans date] [Consulté le 30 juillet 2020]. Disponible à l'adresse : https://creativecommons.org/licenses/

CREATIVE COMMONS, [sans date]b. Choose a License. *creativecommons.org* [en ligne]. [sans date] [Consulté le 12 juillet 2020]. Disponible à l'adresse : https://beta-chooser.creativecommons.org

CREATIVE COMMONS FRANCE, [sans date]. 6 LICENCES gratuites | Creative Commons France. *creativecommons.fr/* [en ligne]. [sans date]. [Consulté le 12 août 2020]. Disponible à l'adresse : http://creativecommons.fr/licences/

DELFT UNIVERSITY OF TECHNOLOGY, [sans date]a. About 4TU.ResearchData. In : *data.4tu.nl* [en ligne]. [sans date]. [Consulté le 4 août 2020]. Disponible à l'adresse : https://data.4tu.nl/info/en/about/organisation/

DELFT UNIVERSITY OF TECHNOLOGY, [sans date]b. Faculties and disciplines. In : *https://www.tudelft.nl/* [en ligne]. [sans date]. [Consulté le 15 juillet 2020]. Disponible à l'adresse : https://www.tudelft.nl/en/research/faculties-and-disciplines/

DELFT UNIVERSITY OF TECHNOLOGY, 2017. *Guideline for creating a README file* [en ligne]. décembre 2017. [Consulté le 21 avril 2020]. Disponible à l'adresse : http://resolver.tudelft.nl/uuid:66e210e1-c884-45d6-b9d4-711907680452

DELFT UNIVERSITY OF TECHNOLOGY, 2020a. *Data preservation and archiving*. mai 2020. Support de cours : MOOC on edX « Open Science: Sharing Your Research with the World », Delft University of Technology, mai 2020

DELFT UNIVERSITY OF TECHNOLOGY, 2020b. *Open Science: Sharing Your Research with the World*. mai 2020. Support de cours : MOOC on edX « Open Science: Sharing Your Research with the World », Delft University of Technology, mai 2020

DELFT UNIVERSITY OF TECHNOLOGY, 2020c. *The advantage of being an Open Researcher*. mai 2020. Support de cours : MOOC on edX « Open Science: Sharing Your Research with the World », Delft University of Technology, mai 2020

DIGITAL CURATION CENTER (DCC), [sans date]a. DCC. *www.dcc.ac.uk/* [en ligne]. [sans date]. [Consulté le 16 août 2020 a]. Disponible à l'adresse : https://www.dcc.ac.uk/

DIGITAL CURATION CENTER (DCC), [sans date]b. Disciplinary Metadata | DCC. *www.dcc.ac.uk/* [en ligne]. [sans date]. [Consulté le 19 juillet 2020 b]. Disponible à l'adresse : https://www.dcc.ac.uk/guidance/standards/metadata

DLCM, [sans date]. About OLOS. *www.dlcm.ch* [en ligne]. [sans date]. [Consulté le 12 août 2020]. Disponible à l'adresse : https://www.dlcm.ch/olos/about.

EDUTECH WIKI, 2020. Gestion des données de recherche. *edutechwiki.unige.ch* [en ligne]. 11 février 2020. [Consulté le 12 juillet 2020]. Disponible à l'adresse : http://edutechwiki.unige.ch/fr/Gestion_des_donn%C3%A9es_de_recherche#Quelles_licences_utiliser_.3F_Qui_d.C3.A9cide_de_l.E2.80.99ouverture_des_donn.C3.A9es_.3F_Comment_proc.C3.A9der_en_cas_de_recherche_internationale_.3F

EPFL, [sans date]. *RDM Walkthrough Guide* [en ligne]. [sans date]. S.l. : s.n. Disponible à l'adresse : https://www.epfl.ch/campus/library/wp-content/uploads/2019/09/RDM_Walkthrough_Guide_20190930.pdf

ETHZ, [sans date]a. File formats for archiving. *library.ethz.ch* [en ligne]. [sans date]. [Consulté le 9 juillet 2020]. Disponible à l'adresse : https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving

ETHZ, [sans date]b. Repository for Publications and Research Data. *ethz.ch* [en ligne]. [sans date]. [Consulté le 5 juillet 2020]. Disponible à l'adresse : https://www.research-collection.ethz.ch/

FANIEL, Ixchel M. et ZIMMERMAN, Ann, 2011. Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *International Journal of Digital Curation*. 11 mars 2011. Vol. 6, n° 1, p. 58-69. [Consulté le 17 mai 2020]. Disponible à l'adresse : DOI 10.2218/ijdc.v6i1.172.

FNS, [sans date]a. Archivage et conservation des données de recherche. In : *www.snf.ch/* [en ligne]. [sans date]. [Consulté le 15 août 2020]. Disponible à l'adresse : http://www.snf.ch/fr/pointrecherche/faq/Pages/faq-open-research-data-archivierung-und-aufbewahrung-forschungsdaten.aspx

FNS, [sans date]b. Data Management Plan (DMP) - Directives pour les chercheuses et chercheurs. *www.snf.ch* [en ligne]. [sans date]. [Consulté le 17 mai 2020]. Disponible à l'adresse : http://www.snf.ch/fr/leFNS/points-de-vue-politique-de-recherche/open_research_data/Pages/data-management-plan-dmp-directives-pour-les-chercheuses-et-chercheurs.aspx

FNS, [sans date]c. Open Access to Publications - SNF. *www.snf.ch* [en ligne]. [sans date]. [Consulté le 25 juillet 2020]. Disponible à l'adresse : http://www.snf.ch/fr/leFNS/points-de-vue-politique-de-recherche/open-access/Pages/default.aspx

FNS, [sans date]d. Open Research Data. *Open Research Data* [en ligne]. [sans date]. [Consulté le 17 mai 2020]. Disponible à l'adresse : http://www.snf.ch/fr/leFNS/points-de-vue-politique-de-recherche/open_research_data/Pages/default.aspx

FNS, [sans date]e. Open Science. *Open Science* [en ligne]. [sans date]. [Consulté le 9 mai 2020]. Disponible à l'adresse : http://www.snf.ch/fr/pointrecherche/dossiers/open-science/Pages/default.aspx

FNS, 2017. Open Research Data : les requêtes devront inclure un plan de gestion des données. *www.snf.ch/* [en ligne]. 6 mars 2017. [Consulté le 25 juillet 2020]. Disponible à l'adresse : http://www.snf.ch/fr/pointrecherche/newsroom/Pages/news-170306-open-research-data-bientot-une-realite.aspx

FOTER, 2015. How To Attribute Creative Commons Photos. *Foter blog* [en ligne]. 4 mars 2015. [Consulté le 12 juillet 2020]. Disponible à l'adresse : https://foter.com/blog/how-to-attribute-creative-commons-photos/

FRANCE ARCHIVES, [sans date]. La pérennisation et ses enjeux. *francearchives.fr/* [en ligne]. [sans date]. [Consulté le 31 juillet 2020]. Disponible à l'adresse : https://francearchives.fr/fr/section/88482503

FRANCE ARCHIVES, 2018. Modèle et normes de l'archivage électronique. *francearchives.fr/* [en ligne]. 13 septembre 2018. [Consulté le 17 juillet 2020]. Disponible à l'adresse : https://francearchives.fr/fr/article/91524937

GAILLARD, Rémi, 2014. *De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?* [en ligne]. Thèse Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques. [Consulté le 16 avril 2020]. Disponible à l'adresse : http://eprints.rclis.org/22746/

IACCONI, Eugénie, 2018. *Les ressources administratives, patrimoniales, culturelles et scientifiques à l'Université de Genève: définition d'une politique de préservation numérique*. [en ligne]. Carouge : Haute école de gestion de Genève. Travail de Master. [Consulté le 16 avril 2020]. Disponible à l'adresse : https://doc.rero.ch/record/324541/files/HEG_TM_EI_20190412_corrige.pdf

JOUDREY, Daniel N. et TAYLOR, Arlene G., 2017. *The Organization of Information*. 4th Edition. ABC-CLIO. ISBN 978-1-4408-6129-1

MANTRA, 2020. *MANTRA Research Data Management Training* [en ligne]. 2020. [Consulté le 16 avril 2020]. Disponible à l'adresse : https://mantra.edina.ac.uk/

MCKIERNAN, Erin C, et al. How open science helps researchers succeed. RODGERS, Peter (éd.), *eLife*. 7 juillet 2016. Vol. 5, p. e16800. [Consulté le 16 avril 2020]. Disponible à l'adresse : DOI 10.7554/eLife.16800

OCDE, 2007. *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics* [en ligne]. 2007. [Consulté le 31 mai 2020]. Disponible à l'adresse : http://www.oecd.org/fr/sti/inno/38500823.pdf

PINFIELD, Stephen, COX, Andrew M. et SMITH, Jen, 2014. Research Data Management and Libraries: Relationships, Activities, Drivers and Influences. LAUNOIS, Pascal (éd.), *PLoS ONE*. 8 décembre 2014. Vol. 9, n° 12, p. e114734. [Consulté le 31 mai 2020]. Disponible à l'adresse : DOI 10.1371/journal.pone.0114734

PLOMP, Esther et al. 2019. Cultural obstacles to research data management and sharing at TU Delft. *Insights*. 9 octobre 2019. Vol. 32, n° 1, p. 29. [Consulté le 31 mai 2020]. Disponible à l'adresse : DOI 10.1629/uksg.484

PRYOR, Graham (éd.), 2012. *Managing research data*. London : Facet Publ. ISBN 978-1-85604-756-2

RICE, Robin et SOUTHALL, John, 2016. *The Data Librarian's Handbook*. S.l. : Facet Publishing. ISBN 978-1-78330-047-1

RILEY, Jenn, 2017. *Understanding metadata: what is metadata, and what is it for?* [en ligne]. [Consulté le 17 juillet 2020]. Disponible à l'adresse : http://www.niso.org/publications/understanding-metadata-riley

SCHNEIDER, René, 2018. *Gestion des données de la recherche. Cours du Master en sciences de l'information.* Carouge. 2018

STANFORD UNIVERSITY, [sans date]. Best practices for file formats. *library.stanford.edu/* [en ligne]. [sans date]. [Consulté le 16 juillet 2020]. Disponible à l'adresse : https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats

SWISSUBASE, [sans date]. SWISSUbase - À propos du projet. *swissubase.ch* [en ligne]. [sans date]. [Consulté le 14 juin 2020]. Disponible à l'adresse : https://swissubase.ch/fr/about-the-project/

SWISSUNIVERSITIES, [sans date]a. Open Access - Numérique. Mondial. Accessible. *swissuniversities.ch* [en ligne]. [sans date]. [Consulté le 9 mai 2020]. Disponible à l'adresse : https://www.swissuniversities.ch/fr/themes/digitalisation/open-access

SWISSUNIVERSITIES, [sans date]b. Open Science – le programme pour des sciences ouvertes. In : *swissuniversities.ch* [en ligne]. [sans date]. [Consulté le 9 mai 2020]. Disponible à l'adresse : https://www.swissuniversities.ch/fr/themes/digitalisation/open-science

SWISSUNIVERSITIES, [sans date]c. *SWISSUbase* [en ligne]. [sans date]. Disponible à l'adresse : https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK_P-2/DOC_SWISSUBase_final.pdf

TENOPIR, Carol et al. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*. 29 juin 2011. Vol. 6, n° 6, p. e21101. DOI 10.1371/journal.pone.0021101

TENOPIR, Carol et al. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE* [en ligne]. 26 août 2015. Vol. 10, n° 8. [Consulté le 6 avril 2020]. Disponible à l'adresse : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4550246/

TIÈCHE, Julien et DUBOIS, Alain, 2015. *La mesure des dimensions de la qualité des archives électroniques: apport des textes normatifs en matière d'archivage électronique à long terme*. [en ligne]. Carouge : Haute école de gestion de Genève. Travail de Bachelor. [Consulté le 16 avril 2020]. Disponible à l'adresse : https://doc.rero.ch/record/258018/files/TDB_Tieche_Julien.pdf

UK DATA SERVICE, [sans date]a. File formats and software. *ukdataservice.ac.uk* [en ligne]. [sans date]. [Consulté le 9 juillet 2020]. Disponible à l'adresse : https://www.ukdataservice.ac.uk/manage-data/format/file-formats.aspx

UK DATA SERVICE, [sans date]b. Research data lifecycle. *ukdataservice.ac.uk* [en ligne]. [sans date]. [Consulté le 12 août 2020]. Disponible à l'adresse : https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx

UNIVERSITÉ DE GENÈVE, [sans date]a. Définitions - Research Data - UNIGE. [en ligne]. [sans date]. [Consulté le 26 juillet 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/footer/definitions/

UNIVERSITÉ DE GENÈVE, [sans date]b. Données de recherche. *unige.ch/biblio* [en ligne]. [sans date]. [Consulté le 19 juillet 2020]. Disponible à l'adresse : https://www.unige.ch/biblio/fr/formation/donnees-de-recherche/

UNIVERSITÉ DE GENÈVE, [sans date]c. Droits et licences. *unige.ch/researchdata* [en ligne]. [sans date]. [Consulté le 19 juillet 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/partager/all/droits/

UNIVERSITÉ DE GENÈVE, [sans date]d. Formations et Présentations. *unige.ch/researchdata* [en ligne]. [sans date]. [Consulté le 19 juillet 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/services/all/formations-presentations/

UNIVERSITÉ DE GENÈVE, [sans date]e. Formats de fichier. *unige.ch/researchdata* [en ligne]. [sans date]. [Consulté le 12 juillet 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/preserver/all/formats-fichier/

UNIVERSITÉ DE GENÈVE, [sans date]f. Le dépôt cantonal des données de recherche Yareta. *unige.ch/researchdata* [en ligne]. [sans date]. [Consulté le 4 août 2020]. Disponible à l'adresse : https://www.unige.ch/eresearch/fr/projets/yareta/

UNIVERSITÉ DE GENÈVE, [sans date]g. Le Programme. *unige.ch/eresearch/* [en ligne]. [sans date]. [Consulté le 4 août 2020]. Disponible à l'adresse : https://www.unige.ch/eresearch/fr/propos/le-programme/

UNIVERSITÉ DE GENÈVE, [sans date]h. Nommer et organiser vos fichiers - Research Data - UNIGE. *unige.ch/eresearch/* [en ligne]. [sans date]. [Consulté le 12 juillet 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/collecter-organiser/all/nommer-et-organiser-vos-fichiers/

UNIVERSITÉ DE GENÈVE, [sans date]i. Que conserver ? *unige.ch/eresearch/* [en ligne]. [sans date]. [Consulté le 15 août 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/preserver/all/que-conserver

UNIVERSITÉ DE GENÈVE, [sans date]j. Yareta - Archivage et préservation des données de recherche. [en ligne]. [sans date]. [Consulté le 19 juillet 2020]. Disponible à l'adresse : https://www.unige.ch/eresearch/fr/services/yareta/

UNIVERSITÉ DE GENÈVE, [sans date]k. Yareta: The Research Data Repository of Geneva's Higher Education Institutions. *yareta.unige.ch* [en ligne]. [sans date]. [Consulté le 5 juillet 2020]. Disponible à l'adresse : https://yareta.unige.ch/

UNIVERSITÉ DE GENÈVE, 2018. *Politique institutionnelle sur la gestion des données de recherche* [en ligne]. 25 juin 2018. [Consulté le 20 avril 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/footer/politique/

UNIVERSITÉ DE GENÈVE, 2019a. Comment remplir le DMP du FNS. In : [en ligne]. Septembre 2019. Disponible à l'adresse : https://www.unige.ch/researchdata/files/4115/6872/2865/201909_atelier_DMP_FNS_presentation_fre.pdf

UNIVERSITÉ DE GENÈVE, 2019b. Diffuser ses données. In : [en ligne]. 2019. Disponible à l'adresse : https://www.unige.ch/researchdata/files/3815/5741/6132/2019_midi_Diffuser_donnees_presentation_fr_20190429.pdf

UNIVERSITÉ DE GENÈVE, 2019c. *Où diffuser ses données de recherche ?* [en ligne]. 2019. Disponible à l'adresse : https://www.unige.ch/researchdata/files/5415/5791/3801/2019_midi_researchdatarepositories_memoV2_fr.pdf

UNIVERSITÉ DE GENÈVE, 2019d. Yareta : Une nouvelle solution numérique pour archiver et partager vos données de recherche - Research Data - UNIGE. In : [en ligne]. 14 juin 2019. [Consulté le 17 mai 2020]. Disponible à l'adresse : https://www.unige.ch/researchdata/fr/actualites/yareta/

UNIVERSITÉ DE GENÈVE, 2020. Publier ses données de recherche avec Yareta. [en ligne]. mai 2020. Disponible à l'adresse : https://www.unige.ch/researchdata/files/3115/8860/8052/202004_atelier_Yareta_fr.pdf

UNIVERSITÉ DE LAUSANNE, [sans date]a. Cycle de vie et types de données. *www.unil.ch* [en ligne]. [sans date]c. [Consulté le 15 août 2020]. Disponible à l'adresse : https://www.unil.ch/openscience/fr/home/menuinst/open-research-data/les-donnees-de-recherche/cycle-de-vie-et-types-de-donnees.html

UNIVERSITÉ DE LAUSANNE, [sans date]b. L'Open Science à l'UNIL: Archivage & partage. *unil.ch* [en ligne]. [sans date]a. [Consulté le 5 juillet 2020]. Disponible à l'adresse : https://www.unil.ch/openscience/fr/home/menuinst/open-research-data/gerer-ses-donnees-de-recherche/archivage--partage.html

UNIVERSITÉ DE LAUSANNE, [sans date]c. Organisation & description. *www.unil.ch* [en ligne]. [sans date]b. [Consulté le 14 août 2020]. Disponible à l'adresse : https://www.unil.ch/openscience/fr/home/menuinst/open-research-data/gerer-ses-donnees-de-recherche/organisation--description.html

UNIVERSITY OF BASEL, [sans date]a. Preserving. *researchdata.unibas.ch* [en ligne]. [sans date]. [Consulté le 14 juin 2020]. Disponible à l'adresse : https://researchdata.unibas.ch/en/preserve-store/

UNIVERSITY OF BASEL, [sans date]b. Sharing. In : *researchdata.unibas.ch* [en ligne]. [sans date]. [Consulté le 5 juillet 2020]. Disponible à l'adresse : https://researchdata.unibas.ch/en/publish-and-share/

UNIVERSITY OF BERN, [sans date]a. BORIS: Bern Open Repository and Information System. *boris.unibe.ch* [en ligne]. [sans date]. [Consulté le 5 juillet 2020]. Disponible à l'adresse : https://boris.unibe.ch/

UNIVERSITY OF BERN, [sans date]b. Research Data Management. In : *unibe.ch* [en ligne]. [sans date]. [Consulté le 14 juin 2020]. Disponible à l'adresse : https://www.unibe.ch/university/services/university_library/services/open_science/research_data_management/index_eng.html

UNIVERSITY OF NORTH CAROLINE, [sans date]. Metadata for Data Management: A Tutorial: Standards/Schema. [en ligne]. [Consulté le 27 juillet 2020]. Disponible à l'adresse : https://guides.lib.unc.edu/metadata/standards

UNIVERSITY OF ZURICH, [sans date]. Data Repositories. *hbz.uzh.ch* [en ligne]. [sans date]. [Consulté le 5 juillet 2020]. Disponible à l'adresse : https://www.hbz.uzh.ch/en/open-access-und-open-science/daten-repositories.html

VAN DEN EYNDEN, Veerle, 2011. *Managing and sharing data: a best practice guide for researchers* [en ligne]. Colchester : UK Data Archive. [Consulté le 31 mars 2020]. Disponible à l'adresse : http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

VAN DEN EYNDEN, Veerle et al. 2016. *Survey of Wellcome researchers and their attitudes to open research* [en ligne]. 31 octobre 2016. [Consulté le 30 juillet 2020]. Disponible à l'adresse : https://wellcome.figshare.com/articles/journal_contribution/Survey_of_Wellcome_researchers_and_their_attitudes_to_open_research/4055448

WHYTE, Angus, 2015. Where to keep research data. *dcc.ac.uk* [en ligne]. 28 décembre 2015. [Consulté le 6 août 2020]. Disponible à l'adresse : https://www.dcc.ac.uk/guidance/how-guides/where-keep-research-data

WILKINSON, Mark D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. décembre 2016. Vol. 3, n° 1, p. 160018. DOI 10.1038/sdata.2016.18

Annexe 1 : Définition des principes FAIR

TO BE FINDABLE :

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

Source : FORCE11, [sans date]. The FAIR data principles. *force11.org* [en ligne]. [Consulté le 4 août 2020]. Disponible à l'adresse : https://www.force11.org/group/fairgroup/fairprinciples

Annexe 2 : Détails des dépôts versés sur Yareta*

<table>
<tr><td rowspan="11">88 dépôts</td><td rowspan="9">78 dépôts publics</td><td rowspan="3">31 dépôts entièrement analysés dont 1 dépôt Cellosaurus (métadonnées Datacite & METS & fichiers téléchargés)</td><td>15 dépôts dont au moins un fichier ne s’ouvre pas avec mon matériel informatique</td><td></td></tr>
<tr><td rowspan="2">16 dépôts dont les fichiers s’ouvrent</td><td>9 dépôts liés à une publication</td></tr>
<tr><td>7 dépôts non liés à une publication et sans README</td></tr>
<tr><td rowspan="3">10 dépôts partiellement analysés</td><td>3 dépôts</td><td>3 dépôts partiellement analysé (métadonnées Datacite & fichiers téléchargés)</td></tr>
<tr><td rowspan="2">7 dépôts non téléchargés car trop lourds</td><td>3 dépôts partiellement analysés (métadonnées Datacite & METS)</td></tr>
<tr><td>4 dépôts partiellement analysés (métadonnées Datacite)</td></tr>
<tr><td rowspan="2">37 dépôts pas du tout analysés</td><td>33 dépôts Cellosaurus</td><td>Pas du tout analysés car ces dépôts sont similaires au dépôt Cellosaurus entièrement analysé</td></tr>
<tr><td>4 dépôts DLCM</td><td>Pas du tout analysés car ce sont des dépôts créés pour les formations Yareta</td></tr>
<tr><td rowspan="2">7 dépôts restreints</td><td colspan="2">5 dépôts</td><td>Partiellement analysés (métadonnées Datacite & METS)</td></tr>
<tr><td colspan="2">2 dépôts DLCM</td><td>Pas du tout analysés car ce sont des dépôts créés pour les formations Yareta</td></tr>
<tr><td rowspan="2">3 dépôts fermés</td><td colspan="2">2 dépôts</td><td>Partiellement analysés (métadonnées Datacite & METS)</td></tr>
<tr><td colspan="2">1 dépôt</td><td>Partiellement analysé (métadonnées Datacite)</td></tr>
</table>

* Situation au 16 avril 2020

Annexe 3 : Grille d'analyse des dépôts versés sur Yareta

Titre	
Description	
Lié à une publication (O/N)	
Mention du Dataset dans l'article (O/N)	
Mention de la publication dans Yareta	
Format – Doc DLCM niveau -1 (métadonnées DataCite)	
Format - Doc DLCM niveau -2 (métadonnées METS)	
Licence Accès	
Date de collecte des données	
README file contenant	
• Convention de nom	
• Contexte de la récolte des données (résumé, objectif, hypothèse, …)	
• Méthode de collecte des données	
• Structure, lien entre les fichiers	
• Descriptions des variables,…	
• Définition des codes, acronymes	
Utilisation de métadonnées spécifiques	
Description des changements	
Clarté des noms des fichiers	
Téléchargement des données (Taille + O/N)	
Nbre d'*	*Critère de qualité assigné par Yareta en fonction du format, et permettant de signaler au chercheur si son format est adéquat ou pas.* *Maximum : ******
Remarque	

Annexe 4 : Analyse de dépôts versés sur Yareta

Exemple d'analyse de deux dépôts versés sur Yareta en accès libre ou fermé.

1. **Analyse d'un dépôt en accès libre**

Titre : *xxx*	
Description : *xx est un corpus de parole, aligné et annoté, développé pour l'étude des proéminences syllabiques en français. Il inclut 24 enregistrements échantillonnés en 7 genres (ou styles) de parole et produits par des locuteurs francophones (issus de Belgique, de France et de Suisse). Il comporte 12 locutrices et 16 locuteurs, en 70 minutes de parole*.	
Lié à une publication (O/N)	Au premier abord non, car il n'y a aucune mention d'un article dans Yareta, et pas de publication 2019 de ce chercheur dans Archive Ouverte. Cependant, le document word déposé avec le jeu de données mentionne 13 articles fait à partir de ce corpus (2007-2011)
Mention du Dataset dans l'article (O/N)	Non, mais les articles datent d'avant Yareta (2007-2011)
Mention de la publication dans Yareta	Référence aux articles dans le document word + pdf déposé avec les données
Format – Doc DLCM niveau -1 (métadonnées DataCite)	• application/octet-stream • audio/x-wave • application/vnd.openxmlformats-officedocument.wordprocessingml.document • application/pdf • text/xml Le document word + pdf mentionne que les outils utilisés pour annoter le corpus sont également distribués librement
Format - Doc DLCM niveau -2 (métadonnées METS)	• <fits:externalIdentifier type="**puid**" toolversion="**6.4**" toolname="**Droid**">fmt/141</fits:externalIdentifier> PRONOM identifier: MIME: audio/x-wav • <fits:externalIdentifier type="puid" toolversion="6.4" toolname="Droid">fmt/189</fits:externalIdentifier> PRONOM identifier : Microsoft Office Open XML *OOXML is not an open format but the specification and development rights have been released under Microsoft's Covenant not to Sue. OOXML is platform independent.* • <fits:externalIdentifier type="puid" toolversion="6.4" toolname="Droid">fmt/19</fits:externalIdentifier> PRONOM : PDF (1.5) Le document word + pdf mentionne que tous les fichiers sont lisibles dans praat (téléchargeable et code source disponible) (www.praat.org)
Licence Accès	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public
Date de collecte des données	"Issued">2019-07-05T00:00:00Z</ "CollectedStart">2006-12-31T23:00:00Z< "CollectedEnd">2010-12-30T23:00:00Z</ "Created">2019-07-05T09:03:01.182Z</

	"Updated">2019-09-02T20:17:52.633Z</ "Accepted">2019-09-02T20:17:54.348Z< Date de début/fin de collecte et correspond à la description
README file contenant :	Pas de README mais un document word et pdf fourni avec les données et qui donne : • Description détaillée des enregistrements • Formats des fichiers distribués • Inventaire des symboles utilisés pour l'annotation • Signification des symboles • Licence • Référence • Publications • Constitution et annotation du corpus • Études faites à partir du corpus
• Convention de noms	Tous les fichiers sont nommés de manière cohérente et le document word + pdf contient un tableau résumant les enregistrements. Donc, avec le nom du fichier on sait à quel type d'enregistrement les données correspondent.
• Contexte de la récolte des données (résumé, objectif, hypothèse, …)	La description du dépôt est reprise dans le document word + pdf, mais il n'y a pas d'information supplémentaire.
• Méthode de collecte des données	Pas spécifié dans le document word mais plusieurs articles ayant utilisés ce corpus sont référencés.
• Structure, lien entre les fichiers	-
• Descriptions des variables, …	Oui, dans le document word et pdf
• Définition des codes, acronymes	Oui
Utilisation de métadonnées spécifiques	-
Description des changements	Non
Clarté des noms des fichiers	Oui
Téléchargement des données (Taille + O/N)	*Taille* Oui
Nbre d'*	ComplianceLevel 3 ou 4 en fonction des données (correspond à ** ou ***)
Remarque	Bonne description du jeu de données 1 dossier documentation et 1 dossier research data Dossier documentation : 1 doc word et pdf Bon exemple de données pouvant être réutilisées

2. Analyse d'un dépôt en accès fermé

Titre : *xxx*	
Description: *Titre de la publication*	
Lié à une publication (O/N)	Oui, trouvé l'article dans Archive Ouverte Accès libre à la version postprint Accès restreint à UniGE pour la version publiée
Mention du Dataset dans l'article (O/N)	Oui *Supplementary Material* *General synthetic procedures and characterization of the synthetic products (pdf file) and microarray data (Excel file) are available on the WWW under https:// doi.org/10.xx. Raw data associated with experiments has been deposited and is available (https://yareta.unige.ch/xxx).*
Mention de la publication dans Yareta	Non, mais on n'a pas accès aux données
Format – Doc DLCM niveau -1 (métadonnées DataCite)	• image/jpeg • application/x-sqlite3 • image/tiff • image/vnd.adobe.photoshop • text/tab-separated-values • text/xml
Format - Doc DLCM niveau -2 (métadonnées METS)	A cause de contraintes techniques dues à mon matériel informatique, ce document n'a pas pu être analysé
Licence Accès	Creative Commons Attribution 4.0 International Fermé (Public depuis le 13.06.2020 00:43)
Date de collecte des données	"Issued">2019-11-28T00:00:00Z</datacite:date> "CollectedStart">2019-11-27T23:00:00Z</datacite:date> "CollectedEnd"></datacite:date> "Created">2019-11-28T14:35:13.428Z</datacite:date> "Updated">2019-12-12T22:43:47.809Z</datacite:date> "Accepted">2019-12-12T22:43:53.351Z</datacite:date> "Available">2020-06-12T22:43:49.899Z</datacite:date> Date de début, pas de date de fin et correspond à la veille du dépôt.
README file contenant	Na
• Convention de nom	Na
• Contexte de la récolte des données (résumé, objectif, hypothèse, …)	Na
• Méthode de collecte des données	Na
• Structure, lien entre les fichiers	Na

Descriptions des variables, …	Na
• Définition des codes, acronymes	Na
Utilisation de métadonnées spécifiques	Na
Description des changements	Na
Clarté des noms des fichiers	Na
Téléchargement des données (Taille + O/N)	*Taille* Non (accès fermé)
Nbre d'*	A cause de contraintes techniques dues à mon matériel informatique, ce critère n'a pas pu être analysé
Remarque	

Annexe 5 : Guide d'entretien

Nom du répondant	
Nom du groupe de recherche	Date : Durée de l'entretien :

Guide/formation

1. Avez-vous suivi une formation pour l'utilisation de Yareta? Si non, pourquoi ?

2. Utilisez-vous le guide Yareta lorsque vous déposer vos données ?

Processus pour déposer des données

3. Qui a la responsabilité de déposer les données sur Yareta (chef de groupe? 1er auteur?)

4. Avez-vous déjà des règles/pratiques communes établies dans votre groupe (e.g. convention de nommage pour les fichiers/dossiers) ou avez-vous dû créer des règles spécifiquement pour Yareta ?

5. Avez-vous une/des conventions au sein de votre groupe pour remplir / déposer sur Yareta, par exemple :

- convention pour le titre/description du jeu de données
- convention concernant la mention du dataset dans la publication, ou la mention de la publication dans Yareta

6. Quand (à quelle étape du projet), déposez-vous les données sur Yareta, par exemple :

- au moment de la collecte/analyse des données? dépôt avec/sans nettoyage des données?
- au moment de la publication ? dépôt avec/sans nettoyage des données?
- après la publication de l'article? dépôt avec/sans nettoyage des données?

7. Comment choisissez-vous les données que vous allez déposer sur Yareta?

8. Où allez-vous chercher les données à déposer (sur votre propre ordinateur, centralisées sur le NAS, centralisées ailleurs ?)

9. Est-ce que vous devez préparer/modifier vos données pour le dépôt ? Si oui, quelle préparation/changement devez-vous faire (créer une arborescence, changer le nom des fichiers/dossier, changer le format, …) ?

10. Pouvez-vous décrire le processus de dépôt des données sur Yareta ?

11. Quel est le format des données que vous déposez sur Yareta ? S'agit-il du format d'origine ? Migrez-vous au contraire vos données dans un autre format ? Si oui, lequel ? Si oui, pour quelle raison ?

12. Comment choisissez-vous la licence, pourquoi ?

13. S'il n'existe pas de README, pourquoi?

14. Pourquoi déposez-vous vos données sur Yareta (ou sur un autre dépôt e.g. Zenodo)? Comprenez-vous le but d'une telle plateforme ?

15. Si applicable à ce groupe de recherche : pourquoi déposez-vous des données sur Yareta avec un accès restreint/fermé ?

Besoins actuels et futurs

16. Que pensez-vous du processus actuel de dépôt des données sur Yareta ? Quelles sont ses forces et ses faiblesses ?

17. Avez-vous des besoins spécifiques pour déposer vos données sur Yareta (qqn qui soit avec vous pendant que vous faites le dépôt ? un guide d'utilisation ?)

18. Lorsqu'on dépose ses données sur Yareta, il y a maintenant la possibilité de mentionner la discipline et les mots-clés. Que mettriez-vous comme discipline et mots-clés ? Quelle est à votre avis la plus-value de ces compléments ?

ELN

19. L'Electronic Lab Notebook (ELN) pourrait faciliter la préparation des données pour une préservation à long terme. Utilisez-vous un cahier de laboratoire électronique dans votre groupe ?

Annexe 6 : Formats recommandés – Université de Genève et ETHZ

Ce tableau reprend les formats mentionnés sur les pages web de l'Université de Genève[35] (provenant du UK Data Service[36]) et de l'ETHZ[37], et met en évidence dans les formats proposés par le UK Data Service:

- En jaune : les formats recommandés par l'ETHZ pour +/- 10 ans
- En bleu : les formats recommandés par l'ETHZ pour moins de 10 ans
- En violet : les formats non recommandés par l'ETHZ pour de l'archivage

Type de donnée	UK Data Service – Recommended	UK Data Service – Acceptable	ETHZ – Recommended (< ou > 10 years)	ETHZ – suitable to only a limited extent (< 10 years)	ETH not suitable for archiving
Tabular data with extensive metadata (variable labels, code labels, and defined missing values)	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file	proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)	-	-	-
Tabular data with minimal metadata (column headings, variable names)	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements	delimited text (.txt) with characters not present in data used as delimiters widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)	Spreadsheet or table : Comma- or tab delimited text files (*.csv)	Spreadsheet or table : Excel *.xlsx (container format) OpenDocument spreadsheets (*.ods)	Excel *.xls, *.xlsb (binary formats)

[35] https://www.unige.ch/researchdata/fr/preserver/all/formats-fichier/
[36] https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats
[37] https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving

Type de donnée	UK Data Service – Recommended	UK Data Service – Acceptable	ETHZ – Recommended (< ou > 10 years)	ETHZ – suitable to only a limited extent (< 10 years)	ETH not suitable for archiving
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) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti	Text : PDF/A (*.pdf) Plain Text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) coded as ASCII, UTF-8, or UTF-16 using byte order mark XML (inclusive XSD/XSL/XHTML etc.; with included or accessible schema and character encode explicitly specified)	Text : PDF (*.pdf) with embedded fonts Plain text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) (ISO 8859-1 coded) Rich Text Format (*.rtf) HTML and XML (The ASCII text is readable over long term; try to avoid external links.) Not accepted for publication, OK for supplementary materials: Word *.docx PowerPoint *.pptx LaTeX, TeX (The ASCII text is readable over long term; open source software required for formatting and the resulting PDF should be included.)	Text : Word *.doc PowerPoint *.ppt

Type de donnée	UK Data Service – Recommended	UK Data Service – Acceptable	ETHZ – Recommended (< ou > 10 years)	ETHZ – suitable to only a limited extent (< 10 years)	ETH not suitable for archiving
				OpenDocument formats (*.odm, *.odt, *.odg, *.odc, *.odf)	
Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif) TIFF other versions (.tif, .tiff) RAW image format (.raw) Photoshop files (.psd) BMP (.bmp) PNG (.png) (mais pas compressé d'après ETHZ) Adobe Portable Document Format (PDF/A, PDF) (.pdf)	TIFF (*.tif) (uncompressed, preferentially TIFF 6.0, Part 1: baseline TIFF). TIFF is preferred as compared to PNG or JPEG2000. Portable Network Graphics (*.png, uncompressed) JPEG2000 (*.jp2, lossless compression) Digital-Negative-Format (*.dng) to keep raw data of digital fotos in addition to an second copy in TIFF forma	TIFF (*.tif) (compressed) GIF (*.gif) BMP (*.bmp) JPEG/JFIF (*.jpg) JPEG2000 (lossy compression) (*.jp2)	-
Audio data	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav)	WAV (*.wav) (uncompressed, pulse-code modulated)	Advanced Audio Coding (*.mp4) MP3 (*.mp3)	-
Video data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)	AVCHD video (.avchd)	FFV1 codec (version 3 or later) in Matroska container (*.mkv)	MPEG-2 (*.mpg,*.mpeg) MP4, which is also called MPEG-4 Part 14 (*.mp4) QuickTime Movie (*.mov) [2)] Audio Video Interleave (*.avi)	Windows Media Video (*.wmv)

Type de donnée	UK Data Service – Recommended	UK Data Service – Acceptable	ETHZ – Recommended (< ou > 10 years)	ETHZ – suitable to only a limited extent (< 10 years)	ETH not suitable for archiving
				Motion JPEG 2000 (*.mj2, *.mjp2)	
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) widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0	Text : PDF/A (*.pdf) Plain Text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) coded as ASCII, UTF-8, or UTF-16 using byte order mark XML (inclusive XSD/XSL/XHTML etc.; with included or accessible schema and character encode explicitly specified)	Text : PDF (*.pdf) with embedded fonts Plain text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) (ISO 8859-1 coded) Rich Text Format (*.rtf) HTML and XML (The ASCII text is readable over long term; try to avoid external links.) Not accepted for publication, OK for supplementary materials: Word *.docx PowerPoint *.pptx LaTeX, TeX (The ASCII text is readable over long term; open source software required for formatting and the resulting PDF should be included.) OpenDocument formats (*.odm, *.odt, *.odg, *.odc, *.odf)	Text : Word *.doc PowerPoint *.ppt