Ouverture des données de la recherche dans la communauté en Intelligence Artificielle : exemple du Plan de gestion des données ANITI

Candice Fillaud

Master 1

Mention Information et Médiation Scientifique et Technique

IMST Master Information et Médiation Scientifique et Technique

Mémoire de stage

Ouverture des données de la recherche dans la communauté en Intelligence Artificielle : exemple du Plan de gestion des données ANITI

Candice FILLAUD

Année universitaire 2020/2021

Sous la direction de Violaine REBOUILLAT

Attachée temporaire d'enseignement et de recherche en Sciences de l'information et de la communication

ANITI Toulouse et URFIST Occitanie

Tutrices professionnelles : Corinne JOFFRE, secrétaire générale d'ANITI

et Amélie BARRIO conservatrice des bibliothèques, co-responsable de l'URFIST Occitanie

Université de Lyon | Lyon 1

Remerciements

Je tiens d'abord à remercier mes deux cotutrices, Amélie Barrio et Corinne Joffre, qui m'ont permis de réaliser ce stage si complet et intéressant. Je voudrais remercier tout particulièrement Amélie Barrio avec qui j'ai travaillé tout au long de mon stage. Je la remercie pour son accompagnement, sa confiance, sa bienveillance, ses conseils et les riches échanges que nous avons eus. Je pense qu'il est difficile de trouver une tutrice aussi passionnante et investie, c'est en grande partie grâce à Amélie que j'ai réalisé ce stage avec un réel plaisir.

Un grand merci aux membres de l'équipe projet avec qui j'ai pu échanger et mener les entretiens pour la rédaction du plan de gestion des données : Emilie Marchand, Aricia Bassinet, Loeticia Moya, Chloée Fabre, Yann Sérot, Sandrine Fourquin, Soraya Demay, Philippe Eyraud, Marie-Françoise Brémond et Aubin Buffière. Merci pour leur accueil et leur accompagnement, leurs conseils.

Je remercie les porteurs de chaires, les chercheurs associés, les post-doctorants et les doctorants de l'institut ANITI que j'ai pu rencontrer. Qui parfois ont pris plus de temps lors des entretiens pour nous emmener au cœur de leurs recherches et nous ont appris beaucoup sur la communauté IA. Merci pour les échanges et leur temps précieux.

Je remercie également ma tutrice pédagogique Violaine Rebouillat pour son soutien permanent, cela a toujours été un plaisir de travailler avec elle. Son accompagnement dans la rédaction de mon mémoire et ses conseils m'ont été très précieux.

Merci également à Guillaume Sire co-responsable URFIST de m'avoir gracieusement prêté son bureau pour l'intégralité de mon stage, merci pour son accueil, sa bonne humeur communicante et les anecdotes échangées.

Bien évidemment je remercie mes proches qui ont été là pour moi lors de mes périodes de questionnement.

Résumé

Depuis 2019 l'Agence nationale de recherche a mis en place un plan de gestion des données obligatoire pour tout projet de recherche financé. ANITI est un institut avec financement ANR PIA3 2019-2023, l'élaboration d'un PGD fait partie intégrante du processus de recherche. C'est dans ce cadre-là que l'URFIST Occitanie a été mandatée pour l'organisation et la rédaction du PGD. Ce mémoire se focalise sur la communauté des chercheurs en IA, leurs pratiques en termes de gestion et diffusion des données. Il suit et décrit la méthodologie employée pour élaborer ce PGD et détermine en quoi ce document permet de valoriser les pratiques des chercheurs et quelles en sont les limites.

Mots-clés

Données de la recherche ; Plan de Gestion des Données ; Pratiques des chercheurs ; Valorisation des données ; Science ouverte

Abstract

Since 2019 the National Research Agency has implemented a mandatory data management plan for any funded research project. ANITI is an institute with ANR PIA3 2019-2023 funding, the development of a DMP is an integral part of the research process. It is within this framework that URFIST Occitanie has been mandated to organize and write the DMP. This thesis focuses on the AI research community, their practices in terms of data management and dissemination. It follows and describes the methodology used to elaborate this DMP and determines in what way this document allows to valorize the researchers practices and what are its limits.

Keywords

Research data ; Data management plan ; Researchers practices ; Data valuation ; Open Science

Liste des abréviations

ANITI : Artificial and Natural Intelligence Toulouse Institute

ANR : Agence Nationale de Recherche

CéSO : Comité de réflexion pour la Science Ouverte

CNRS : Centre National de Recherche Scientifique

DMP (PGD) : Data Management Plan (Plan de Gestion des Données)

DMP OPIDoR : Data Management Plan pour l'Optimisation du Partage et de l'Interopérabilité des Données de Recherche

DPD : Délégué à la Protection des Données

ENAC : École Nationale de l'Aviation Civile

FAIR : Findable, Accessible, Interopable, Reusable

INIST : Institut de l'Information Scientifique et Technique

INP : Institut National Polytechnique

INSA : Institut National des Sciences Appliquées

IST : Information Scientifique et Technique

LAAS : Laboratoire d'Analyse et d'Architecture des Systèmes

LIP : Lyon Ingénierie Projets

OATAO : Open Archive Toulouse Archive Ouverte

OCDE : Organisation de Coopération et de Développement Économique

ONERA : Office National d'Études et de Recherches Aérospatiales

OpenAIRE : Open Access Infrastructure of Research in Europe

RGPD : Règlement Général sur la Protection des Données

URFIST : Unité Régionale de Formation à l'Information Scientifique et Technique

SICD : Service Inter-établissement de Coopération Documentaire

UT1 : Université Toulouse 1

UT2J : Université Toulouse Jean Jaurès

UT3 : Université Toulouse 3

Sommaire

Introduction

Depuis le début des années 2000, propulsés par la révolution du numérique (Beckouche, 2017), des changements économiques et politiques s'opèrent vers une ouverture des données accompagnée d'une amplification du libre accès au savoir (*Berlin Declaration*„ 2003). Le mouvement de la science ouverte gagne de plus en plus de terrain et semble s'engager fortement pour l'ouverture des données de la recherche prônant de nombreux bienfaits sur la société comme l'intégrité scientifique, la transparence, la reproductibilité, l'innovation et le retour sur investissement pour les agences de financement. Un écosystème se crée alors, impliquant de nombreux acteurs autour des données de la recherche rassemblant les chercheurs et communautés scientifiques, les financeurs, les établissements supérieurs de la recherche, les éditeurs, de nouveaux intermédiaires de l'information scientifique et technique (IST) et des médiateurs (Pain, 2016). De nouvelles politiques se mettent en place régulièrement, l'année 2021 marque l'annonce du Deuxième Plan national pour la science ouverte suivant trois axes : généraliser l'accès ouvert aux publications ; structurer, partager et ouvrir les données de la recherche ; ouvrir et promouvoir les codes sources produits par la recherche. Dans le deuxième axe est stipulée la mise en place effective des plans de gestion des données garantissant la préservation économe, l'ouverture ou le partage de données documentées, en créant les conditions de leur réutilisation et de leur valorisation. Le plan de gestion de données est un document rendu obligatoire par les agences de financements comme l'Agence nationale de la recherche (ANR), renseignant les actions mises en œuvre tout au long du cycle de vie de la donnée. C'est un document qui doit être rédigé au commencement du projet de recherche et mis à jour régulièrement jusqu'à son terme recherche de manière à expliciter la mise à disposition des données en correspondant au maximum aux principes FAIR (Wilkinson et al. 2016).

L'institut Artificial and Natural Intelligence Toulouse Institute (ANITI) est un projet pluridisciplinaire de recherche relevant d'un financement ANR, la mise en place d'un plan de gestion des données conditionne donc son financement. Pour réaliser son PGD, ANITI a fait appel à l'Unité régionale de formation à l'information scientifique et technique d'Occitanie (URFIST Occitanie), regroupant une équipe de professionnels de l'IST présents dans les établissements académiques partenaires d'ANITI. C'est dans ce cadre que j'ai réalisé mon stage de Master 1, avec l'équipe projet nos missions étaient de recenser les pratiques des chercheurs en termes de gestion et diffusion des données afin d'élaborer et rédiger le PGD d'ANITI. Dans un second temps nous avons été chargés de créer un modèle de guide des bonnes pratiques à destination des chercheurs d'ANITI. Dans ce mémoire je ciblerai donc la communauté des chercheurs en intelligence artificielle (IA) que j'ai côtoyés durant ces cinq mois de stage et dont la culture de la donnée est hétérogène selon les disciplines et déploierai ainsi la problématique suivante : En quoi un PGD permet-il de saisir les pratiques pluridisciplinaires des chercheurs en IA et de favoriser les bonnes pratiques de gestion et diffusion des données de la recherche ?

Pour répondre à cette problématique, je poserai dans un premier temps le contexte de l'ouverture des données de la recherche et ferai un rappel historique sur l'origine et la mise en place des plans de gestion des données. J'aborderai dans un deuxième temps une partie

plus méthodologique, sur la manière dont nous avons procédé avec l'équipe projet pour rédiger le PGD ANITI. Je m'attacherai ensuite à décrire les spécificités de la communauté IA et ses pratiques en matière de données de la recherche. Enfin je montrerai en quoi le plan de gestion des données permet une valorisation des pratiques des chercheurs et étayerai mes propos d'un retour d'expérience critique sur l'utilisation du modèle de PGD de l'ANR et de l'outil DMP OPIDoR.

I. Ouverture des données de la recherche : le plan de gestion des données

Encore inexploitées récemment, les données de la recherche sont en train d'acquérir une valeur d'échange dans le cadre des politiques d'ouverture. Dans cette partie nous allons définir la notion complexe des « données de la recherche », les politiques d'ouverture qui l'accompagnent et ce à quoi correspond le plan de gestion des données parmi ces évolutions.

1. Définition des données de la recherche

Le terme « donnée de la recherche » n'a été défini que très récemment. L'une des premières définitions a été posée par l'Organisation de coopération et de développement économique (OCDE) en 2007 dans les Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics *:*

> *« Dans le cadre de ces Principes et Lignes directrices [pour l'accès aux données de la recherche financée sur fonds publics], les « données de la recherche » sont définies comme des 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. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche ».*

Faisant partie intégrante de la recherche, étant présentes partout autour de nous, les données restent un terme complexe à définir du fait de leur hétérogénéité. La notion de « donnée » varie selon les disciplines et les objets de recherche et ne sera pas définie de la même façon par un paléontologue ou par un chercheur en chimie qui auront tous les deux une culture de la donnée bien différente. Par manque de consensus véritable sur la définition des Données l'OCDE publie également une liste en 2007, en complément de la définition, de ce qui n'est pas une donnée : « carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs ». La définition passe donc par des exemples, une énumération exhaustive, les données sont également différentes selon les domaines, leur utilisation etc.

Si l'on revient sur la définition de l'OCDE, les données peuvent se présenter sous plusieurs formes : chiffres, textes, images, sons, qu'elles soient numériques ou non. Dans tous les cas, les données sont une ressource qui permet au chercheur de produire un résultat de recherche. Les données peuvent être classées sous forme de typologie, par exemple en fonction de leur méthode d'acquisition (F. Cabrera, 2015). Selon Borgman et Leonelli (2015), la donnée ne peut être considérée indépendamment de son contexte ; définir le contexte s'avère alors primordial pour définir la donnée. Pour reprendre les termes de C. Borgman, la nature d'une donnée dépend donc de la discipline scientifique, de l'objectif de recherche ou encore de la méthodologie et de l'instrumentation utilisées (Borgman, 2012).

2. Politiques d'ouverture des données

Prenant part au mouvement de la science ouverte, la définition des données par l'OCDE s'inscrit dans une déclaration de principe et d'ouverture publiée en 2004. Dans cette déclaration est reconnu l'impact bénéfique des données en libre accès. Trente pays dont la France et plusieurs pays de l'Union européenne s'engagent à mettre en place des systèmes d'accès libre aux données de la recherche financée sur fonds publics.

Au niveau national, l'ouverture des données de la recherche publique est posée par la loi pour une République numérique en 2016, qui rend les données ouvertes librement réutilisables (Robin 2017). Deux ans après, est instauré le Plan national pour la science ouverte, qui propulse les données de la recherche au cœur du changement de politique scientifique, celles-ci constituant le deuxième axe du plan : « Structurer et ouvrir les données de la recherche ». L'objectif est de sortir de l'ombre ces produits de la recherche et de leur donner une valeur d'échange, en les ouvrant de manière adaptée selon le principe : « aussi ouvert que possible, aussi fermé que nécessaire ». Leur ouverture se limite aux exceptions légitimes encadrées par la loi comme les données sensibles soumises au secret professionnel, industriel ou commercial ou encore les données à caractère personnel et les données soumises au droit d'auteur. Dans ce cadre l'ANR adopte elle aussi une politique en faveur de la science et des données ouvertes : à partir de 2019, elle impose la rédaction d'un plan de gestion de données aux projets de recherche qu'elle finance.

L'annonce du deuxième Plan national pour la science ouverte (2021-2024) introduit de nouvelles mesures, parmi lesquelles la mise en place d'une politique d'accompagnement à la diffusion des données de la recherche financée sur fonds publics, telle que prévue par la loi pour une République numérique. D'autres normes verront le jour d'ici 2024, dont la création d'un entrepôt national fédérant les entrepôts de données existants, l'adoption d'une politique d'ouverture des données structurée autour d'un réseau d'administrateurs des données et l'incitation à la réutilisation des données grâce notamment à la création d'un prix permettant de récompenser le travail de préparation des données à leur réutilisation. Une des mesures à laquelle nous prêtons une attention particulière est la généralisation des plans de gestion des données à partir de 2021 pour tout projet de recherche financé.

3. Plan de gestion des données

Le plan de gestion des données, par sa définition, est un document décrivant le cycle de vie des données. C'est un document évolutif qui doit être pensé en amont et tout au long du projet de recherche. Ce livrable vise à décrire les différentes étapes entre la collecte et l'archivage, notamment comment les données de recherche sont produites, réutilisées, stockées, sécurisées, disséminées ou conservées à long terme. La mise en place de ce document reste très récente pour les organismes de financement français, en réalité le PGD a d'abord été un document utilisé par les chercheurs eux-mêmes. À partir de 1966 et jusque dans les années 2000, les chercheurs l'ont d'abord utilisé au sein de projets en aéronautique et ingénierie complexes puis dans diverses disciplines techniques et scientifiques. Les PGD n'ont pas été utilisés comme livrables à proprement parler, mais comme des documents de

travail relatant les activités de recherche et les développements prévus. Ils servaient d'outils de gestion de projet permettant aux projets complexes de faire face à leurs exigences en matière de gestion des données pendant les étapes de collecte et analyse des données (Smale al. 2018). Un autre exemple est celui de Jayroe (1973), c'est une procédure d'analyse et de gestion des données qui décrit les données à destination d'autres chercheurs du même domaine pour qu'ils puissent entreprendre des projets similaires. Ces documents étaient donc utilisés par des chercheurs dans le but d'aider d'autres chercheurs du même domaine à entreprendre des projets similaires sur des données complexes. Les notions de partage, stockage à long terme et d'archivage n'étaient pas l'objectif que peuvent avoir les actuels PGD, mais ils avaient plutôt pour vocation d'aider d'autres chercheurs dans un domaine restreint, à organiser, collecter, stocker sur le court terme des données de projets complexes. Avec la révolution numérique que nous avons connue au début des années 2000, est paru le rapport de Lord et Macdonald (2003) sur l'impact du passage à l'ère du numérique. Ce rapport évoquait l'absence de financement concernant les dépôts de données, ce manque d'investissement de la part des gouvernement et institutions pouvaient mener à la perte de nombreux jeux de données. Les auteurs du rapport ont donc recommandé à ces organismes de mettre en place des stratégies de gestion, partage et conservation des données pour des données d'une future valeur potentielle afin de les identifier et de les archiver et ainsi pallier les soucis numériques pouvant être rencontrés.

Pour des raisons en partie économiques, la National Science Foundation (NSF) annonce que les soumissions de projets financés devraient s'accompagner d'un DMP à partir de janvier 2011 pour s'assurer de l'archivage et du partage des données. En Europe pour tous les projets financés par la Commission européen dans le cadre du programme *Open Research Data Pilot* (Koulocheri 2017) H2020 la rédaction d'un DMP (*Data Management Plan*) est devenue obligatoire en 2013. En France, les agences de financement ont également mis en place l'élaboration d'un PGD : depuis 2019 l'ANR a adopté une politique pour les données et demande que les projets rédigent un plan de gestion des données. La première version du plan doit être remise à l'ANR dans les six premiers mois, c'est la version initiale. Deux autres versions doivent également être déposées : la version intermédiaire à mi-projet et la version finale au terme du projet. L'ANR n'oblige toutefois pas les chercheurs à ouvrir leurs données au terme du projet ; celles-ci peuvent restées confidentielles sans être forcément partagées. Ce document divise néanmoins, d'un côté il est perçu par les chercheurs comme une contrainte administrative qui s'ajoute à leur charge de travail mais apparaît politiquement comme un document permettant d'améliorer la gestion et l'ouverture des données. Plus généralement le PGD permettrait aux chercheurs et aux projets de tirer des avantages théoriques revendiqués (Smale al.2018) que ça soit des avantages professionnels, économiques et institutionnels. Ces bénéfices popularisés n'ont néanmoins jamais été démontrés, les avantages professionnels pour les chercheurs ont seulement été supposés.

Le plan de gestion des données est également décrit comme l'élément clé pour rendre les données de la recherche le plus étroitement conformes aux principes FAIR (faciles à trouver, accessibles, Interopérables, Réutilisables) (Wilkinson et al. 2016), de manière à ce que les données soient le plus exploitables possibles par les pairs et la communauté scientifique. Notamment qu'elles soient accompagnées de métadonnées, de description dans des formats

les plus compatibles possible, pour résumer qu'elles s'accompagnent de bonnes pratiques de gestion. Le plan de gestion de données a eu un impact positif sur le travail de jeunes chercheurs comme le montre le retour d'expérience de Delphine Thonney (Dällenbach, Thonney 2019) qualifiant l'exercice d'utile et permettant de se « *confronter à la réalité ainsi qu'à la nécessité de bien planifier et penser sa recherche (aussi) dans ce domaine* » ou encore celui de Gwendoline Torterat post-doctorante en archéologie (Torterat 2020) :

> *« Si la rédaction d'un plan de gestion des données n'est pas chose aisée en soi lorsque l'on ne dispose d'aucune formation, nous encourageons tous ceux qui le souhaitent, même timidement, à se lancer. Il nous apparaît comme une phase préliminaire indispensable pour tout projet de recherche, y compris ceux qui ne disposent pas d'un volume de données très conséquent. Une version simplifiée pourrait être proposée pour les projets doctoraux par exemple, ou ceux qui ne disposent que de très peu de données »*

II. Méthodologie : élaboration du plan de gestion des données d'ANITI

Cette deuxième partie consiste à décrire le contexte de la cotutelle qui a été mise en place pour mener à bien le projet d'élaboration d'un plan de gestion de données pour ANITI, elle abordera également la méthodologie que nous avons appliquée avec l'équipe projet.

1. Contexte de la cotutelle URFIST - ANITI

Le projet Artificial and Natural Intelligence Toulouse Institute est un projet toulousain (cf. Annexe 1) bénéficiant d'un financement ANR PIA3 2019-2023. ANITI est donc l'un des quatre instituts interdisciplinaires d'intelligence artificielle (3IA) mis en place pour une durée de quatre ans au sein du Programme Investissements d'Avenir du Plan Villani (*Le Programme d'investissements d'avenir* 2018). L'institut regroupe plus de 200 chercheurs issus de 33 laboratoires de recherche et une trentaine d'entreprises, s'articulant autour de trois programmes de recherche (IA Acceptable, IA Certifiable et IA Collaborative), divisés en 24 chaires présentes dans différents établissements partenaires.

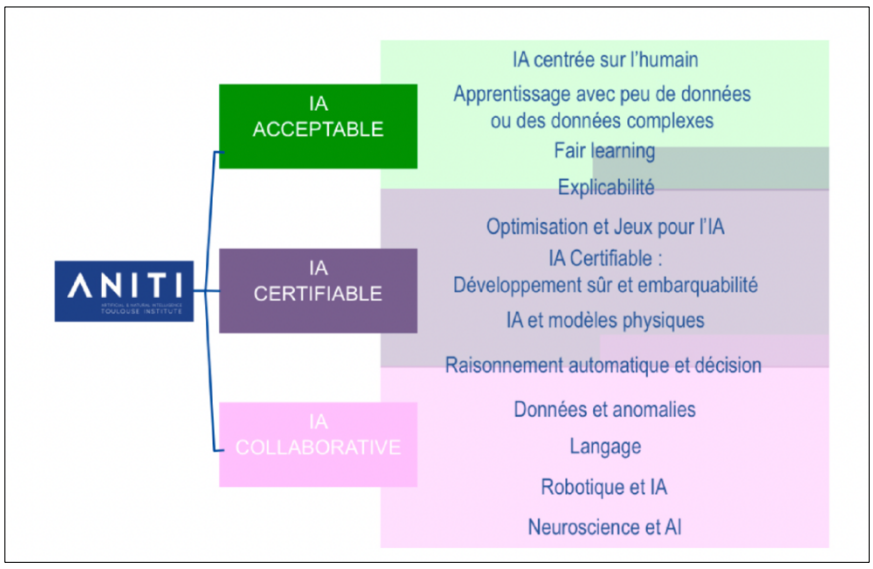


Figure 1 : Organisation des chaires ANITI
(site ANITI : https://aniti.univ-toulouse.fr/)

En tant que projet financé par l'ANR, ANITI est soumis à l'obligation de rédaction d'un plan de gestion de données. L'enjeu de l'élaboration du PGD est considérable de par son obligation, le fait qu'il doit couvrir l'ensemble du projet ANITI en un seul livrable et également par l'aspect pluridisciplinaire du projet, qui regroupe des disciplines comme les mathématiques, l'informatique, les neurosciences ou encore les sciences humaines et sociales. L'élaboration du PGD nécessite une organisation en amont afin de recueillir les jeux de données et les pratiques des chercheurs auprès de la totalité des chaires ANITI.

Pour mener à bien le projet, en décembre 2020, dans une lettre de mission relative à la conception du plan de gestion des données ANITI (cf. Annexe 2), Corinne Joffre, secrétaire générale ANITI, propose et décrit la mission de rédaction du PGD. Les deux objectifs définis sont de fournir une version à mi-projet du PGD avant la fin d'année 2021 ainsi que de faire remonter ce qui doit être mis en place pour une bonne gestion des données (stockage sécurisé et pérenne, ouverture des données, « FAIRisation » des données, etc.) à ANITI et au CéSO. Il est précisé que la collecte d'informations et de jeux de données passera par des entretiens avec les chercheurs ANITI, afin d'engranger un maximum d'informations et d'établir une solide vision des données ainsi que de leurs pratiques associées.

La lettre est adressée à Amélie Barrio, co-responsable de l'URFIST Occitanie, en raison de sa capacité à piloter ce type de projet et à mobiliser une équipe réseau. L'équipe mobilisée pour la rédaction du PGD ANITI a donc été composée de 12 professionnels de l'IST. Ces professionnels font partie du réseau des services d'appui à la recherche présents dans les services communs de documentation (SCD) des établissements partenaires : UT1, UT2, UT3, INP, INSA, ISAE, LAAS-CNRS, ONERA, ENAC. En tant que professionnels de l'IST, l'équipe était déjà familière avec la notion de « données de la recherche » et les récentes exigences de l'ANR. Ce choix de collaboration a permis de mettre les connaissances et l'expérience de l'équipe à profit, permettant d'informer les chercheurs non avertis sur les nouvelles exigences lors des entretiens.

C'est également dans ce cadre-là que mon recrutement en tant que stagiaire à la rédaction du PGD ANITI a eu lieu. J'ai rejoint l'équipe projet avec pour missions de recenser les pratiques en matière de gestion des données des établissements académiques partenaires d'ANITI et en proposer une analyse pour la mise en place de la première version du PGD d'ANITI. Je devais également participer à la collecte d'informations auprès des équipes de recherche d'ANITI, participer aux actions de l'équipe support d'ANITI et de l'URFIST et proposer, dans un deuxième temps, un livrable de bonnes pratiques à destination des chercheurs.

Le rôle d'ANITI a été d'appuyer la prise de contact avec les chercheurs pour la réalisation des entretiens, en fournissant les listes de contacts des porteurs de chaires, chercheurs associés, post-doctorants et doctorants. Afin d'informer les chercheurs ANITI du projet de PGD, un mail d'introduction a été envoyé par les co-directeurs scientifiques d'ANITI, M. Kaaniche et M. Asher. Ce mail adressé aux 24 porteurs de chaire ANITI précisait les enjeux du PGD pour l'institut et justifiait la délégation du projet à l'URFIST Occitanie et aux personnels du SCD des établissements partenaires. L'équipe d'ANITI intervenait aussi dans les cas où les porteurs de chaires ou chercheurs associés ne répondaient pas malgré nos relances (10 cas sur 24). Pour le reste des chercheurs la prise de contact s'est avérée relativement facile grâce à la prévention en amont de la direction ANITI sur ce projet.

2. Méthode de réalisation du PGD

Comme mentionné dans la lettre de mission, la collecte d'informations auprès des chercheurs ANITI s'est effectuée sous forme d'entretiens sociologiques semi-directifs

incitant le chercheur à davantage parler que celui qui pose des questions (Combessie 2007). La première étape du projet PGD a été la création d'une grille d'entretien utilisée par l'ensemble de l'équipe projet lors des entretiens avec les chercheurs (cf. Annexe 3). L'objectif de cette grille était de faciliter les entretiens en se servant de ce document comme trame pour structurer et uniformiser les différentes entrevues. Cette grille a été divisée en sept grandes parties : données réutilisées, données produites, sécurité, stockage des données, diffusion des données, conservation des données, coûts, incluant des sous-parties, précisions ou exemples et s'appuyant des informations demandées dans le modèle de PGD de l'ANR. Une fois les grilles rédigées en deux versions (français et anglais), est venue la question de la prise de contact et de la méthodologie des entretiens.

D'une part nous nous sommes mis d'accord sur la répartition des entretiens avec les chercheurs. Celle-ci s'est faite de manière logique : les membres de l'équipe projet prenaient contact avec les chercheurs de leur établissement partenaire, certains professionnels connaissant en effet déjà les chercheurs, ce qui rendait les échanges plus spontanés et plus fluides. Pour la répartition, nous avons utilisé un tableau de suivi des chercheurs que j'avais préalablement crée et déposé sur notre SCOUT, outil collaboratif spécifique dédié à l'équipe projet et nous permettant de partager tous nos documents et de les organiser. Les entretiens devaient idéalement être enregistrés (enregistrement vocal uniquement) puis retranscrits, permettant ainsi de ne pas perdre d'éléments importants mentionnés lors de l'entretien, d'avoir une vision précise pour faciliter la rédaction du PGD ou encore de revenir vers les chercheurs sur des éléments précis non approfondis. Dans la mesure du possible nous proposions des entretiens en présentiel ou distanciel étant donné les contraintes sanitaires ou contraintes de temps pour les chercheurs. C'est par un mail commun que nous avons contacté dans un premier temps les porteurs de chaire, pour avoir un point de vue global sur chaque chaire. Dans ce mail commun était rappelé le contexte dans lequel l'équipe intervenait (recontextualisation du PGD dans le cadre d'ANITI) et explicité le déroulé des entretiens en précisant qu'un entretien pouvait durer théoriquement entre 30min et 1h30. Un accord de consentement (cf. Annexe 3) validé au préalable par le délégué à la protection des données[1] (DPD ou *DPO* en anglais) en pièce jointe du mail, garantissant la pseudonymisation des enregistrements et laissant le choix aux chercheurs d'accepter ou non l'enregistrement. La grille d'entretien était également envoyée au préalable pour donner au chercheur un ordre d'idée des questions qui allaient lui être posées (quelques chercheurs ont transmis la grille à leurs doctorants, ce qui nous a permis de compléter les entretiens avec leurs réponses).

La première phase d'entretiens s'est déroulé avec les 24 porteurs de chaires. Une fois cette première phase terminée nous avons pu « cartographier » les recherches au sein des chaires et identifier les nouvelles personnes à contacter pour un entretien. Très souvent les porteurs de chaires nous ont redirigés vers leurs doctorants, en charge de la production des données dans le cadre de leur thèse, ou encore vers les chercheurs associés qui pouvait eux aussi manipuler des données au côté des doctorants. Au total 43 entretiens ont été réalisés

[1] « Qualifié de « chef d'orchestre » par la CNIL, le délégué à la protection des données, aussi appelé DPO pour « Data Protection Officer », est la personne en charge de la protection des données à caractère personnel au sein des organismes publics ou privés. » https://datalegaldrive.com/dpo/definition-missions/

dont cinq en anglais ; parmi eux 32 ont été retranscrits, certains chercheurs n'ayant pas souhaité être enregistrés ou en raison d'un problème technique lors de l'enregistrement qui nous a empêché de garder trace de certains entretiens. Lors des entretiens nous avons essayé de composer une équipe de deux membres équipe projet selon les disponibilités de chacun. Sur les 44 entretiens j'ai pu en (co-)animer 33. Pour beaucoup d'entre nous, avoir des chercheurs en entretien pour recueillir leurs pratiques en termes de gestion des données était une première, le fait de former un binôme ou trinôme nous a aidés et a facilité les échanges, permettant ainsi de ne pas oublier d'éléments de la grille ou de répondre plus facilement aux questions des chercheurs.

Mon rôle au sein de l'équipe a été de participer à un maximum d'entretiens afin d'avoir une vision globale pour la rédaction du PGD. Je m'occupais également de retranscrire l'intégralité des entretiens enregistrés (français et anglais). Ce travail, d'une moyenne de 72h, a occupé une grande partie de mon stage. Pour chaque entretien, j'étais également chargée de compléter les grilles d'entretien avec les réponses données par les chercheurs. La plupart du temps, celles-ci étaient remplies en direct lors de l'entretien par l'un des membres du binôme, mais elles étaient également retravaillées par la suite, afin qu'y soient apportées des informations complémentaires sur les jeux de données, isolées au cours du travail de retranscription. Ma position et ma vision d'ensemble m'ont permis de coordonner l'équipe projet : rappel des chercheurs à contacter, demande de telle ou telle information, appui aux entretiens, etc. J'avais également la charge de mettre à jour les outils collaboratifs : tableau de suivi des entretiens, tableau de recensement des besoins des chercheurs, dépôt des grilles, accord de consentement, enregistrements audio, retranscriptions…

Au fur et à mesure des entretiens, nous avons acquis une meilleure connaissance des jeux de données d'ANITI. Nous pouvions en dresser une première cartographie. Avec ma tutrice Amélie Barrio, nous avons réfléchi à la structuration que pourrait prendre le PGD, tout en sachant que je le rédigerai via l'outil DMP OPIDoR sur le modèle « ANR – Modèle de PGD (français) ». L'idée de départ était une organisation par chaire, nous avons finalement opté pour une organisation par jeu de données, qui s'est avérée la plus adéquate avec l'outil DMP OPIDoR. Après validation auprès du comité scientifique d'ANITI, j'ai pu entamer la rédaction en commençant par les jeux de données réutilisés les plus communs, jusqu'aux plus spécifiques générés. Au total plus de 40 jeux de données ont été recensés, comprenant des jeux de données réutilisés au sein de chaires différentes, des jeux de données produits ou encore des jeux de données de partenaires industriels. Le nombre de jeux de données recensés à vocation à évoluer au cours du projet.

III. La communauté des chercheurs en IA : pluridisciplinarité et diversité des cultures de la donnée

En regroupant plus de 200 chercheurs de laboratoires différents, les disciplines peuvent varier au sein même des chaires : droit, informatique, mathématiques... La pluridisciplinarité est la force de cet institut où les chercheurs ont également leur propre culture de la donnée et leurs propres habitudes en termes de gestion et partage. Cette deuxième partie permet de représenter la culture globale de la donnée dans la communauté IA et de montrer la réalité entre les politiques mises en œuvre sur le plan de gestion des données et le ressenti ainsi que l'expérience des chercheurs avec ce récent document.

1. Cultures de la donnée dans la communauté IA

La particularité de cette communauté est l'hétérogénéité des disciplines qu'elle peut fédérer dans le cadre d'un même projet. Au sein d'un institut pluridisciplinaire la culture de la donnée ne s'adapte pas forcément à la culture de certains par rapport aux autres notamment sur le rôle, la place et l'importance des données. Les chercheurs en droit, en économie et en mathématiques ont une culture de la donnée différente des chercheurs en informatique et robotique par exemple. Il a été plus compliqué de recenser les jeux de données des économistes ou des chercheurs en droit, soit parce qu'ils pensaient ne pas mobiliser de jeux de données dans leurs recherches, soit par manque de consensus sur la notion des « données ». Les chercheurs ne considéraient pas utiliser des données dans leurs recherches ne sachant pas définir le périmètre couvert par cette notion. À la différence des communautés en archéologie ou en géologie qui peuvent globalement s'entendre sur le périmètre de leurs données de recherche et de leur importance et qui auront des outils et des entrepôts utilisés communément. La communauté IA est un rassemblement de communautés différentes ayant un dénominateur commun : l'intelligence artificielle. Chaque communauté garde ses pratiques, par exemple ne dispose pas d'entrepôt ou d'outil universel à l'ensemble de la communauté. Un chercheur en neurosciences publiera ses résultats dans un entrepôt comme OpenNeuro[2], un mathématicien créant des modèles déposera sur GitHub etc... Le plan de gestion de données révèle ces pratiques pluridisciplinaires que l'on retrouve avec des jeux de données singuliers, les lieux de stockage, les choix de conservation, des politiques d'ouverture et de partage de données plus ou moins ancrées.

Pour la plupart des chercheurs en apprentissage et modélisation ce qui est ressorti lors des entretiens est une culture de la donnée ouverte déjà ancrée. Que ce soit en *machine learning,* en *deep learning*, en apprentissage les données réutilisées sont souvent ouvertes en licence libre et dans un format *open source*. Les seules limites à l'ouverture sont les partenariats avec les industries privés comme Airbus, Safran, ATMOS ou encore Vitesco technologies, pour lesquels les jeux de données partagés restent la propriété intellectuelle de l'entreprise et n'ont pas vocation à être ouvertes à la communauté. Les chercheurs en apprentissage entraînent

[2] https://openneuro.org/

leurs algorithmes sur des *benchmarks*, souvent des données images avec annotations ou non, son ou texte. Les jeux de données les plus connus sont : MNIST[3], Cifar10[4], Cifar100, imageNet[5], ils sont téléchargeables directement sur les sites internet dédiés ou répertoriés sur des bases de données publiques telles que GitHub, Kaggle[6], UCI[7]. Dans le domaine de l'optimisation et de la résolution de problème les données ne sont pas forcément volumineuses mais également en libre accès sur des bases de données comme PSP lib[8]. Pour produire un jeu de données qui deviendra une référence pour la communauté permettant de comparer des algorithmes, la production est coûteuse et chronophage pour les chercheurs et les doctorants. Ce qui justifie principalement la réutilisation des jeux de données de référence déjà disponibles. En robotique les chercheurs ne déposent pas systématiquement leurs données dans des entrepôts sauf si les éditeurs ou agences de financement forcent la main, car prendre le temps de déposer des données dans un entrepôt spécialisé signifie les adapter avec un format open source, les compresser les documenter suffisamment, et les chercheurs n'ont pas forcément le temps pour ça. Leurs recherches sont également très spécifiques et les chercheurs ne considèrent pas forcément que les déposer soient utiles pour d'autres chercheurs, comme indique ce chercheur d'ANITI :

> *« Donc cet effort-là de le faire et de le produire n'a jamais été important chez nous et on ne l'a fait dans notre cas que quand on en avait besoin pour faire nos propres tests, on a généré les données pour nos besoins propres, surtout on n'a pas fait l'effort derrière de les rendre utilisables par d'autres justement à cause du fait que ça soit du temps. Un doctorant n'aura pas sa thèse parce qu'il a produit de la donnée. »*

Néanmoins le manque de données en robotique est assez révélateur et montre que les chercheurs, même en pensant que leurs données ne seront pas utiles à la communauté, gagneraient à les partager pour combler ce manque de données au sein de leur discipline. Un autre point est soulevé ici, le fait qu'un doctorant n'aura pas sa thèse car il produit de la donnée, produire de la donnée ne signifie pas forcément la production d'une meilleure thèse ou de meilleurs résultats scientifiques. Cela montre encore une fois de plus que les données de la recherche ont encore de la marge avant d'être entièrement valorisées. Les personnes les plus à même de produire et utiliser de la donnée sont les doctorants qui se retrouvent au cœur des manipulations et des recherches, à manipuler des données et les rendre FAIR. Une mise en valeur du travail de la donnée serait donc tout aussi valorisant pour les doctorants commençant leur carrière de chercheur.

En IA les méthodes d'évaluation sont évidemment les publications dans les revues scientifiques, mais également les publications sous forme de communications à lors de

[3] http://yann.lecun.com/exdb/mnist/
[4] https://www.cs.toronto.edu/~kriz/cifar.html
[5] https://www.image-net.org/
[6] Kaggle: Your Machine Learning and Data Science Community; [cité le 20 juill 2021]. Disponible: https://www.kaggle.com/
[7] UCI Machine Learning Repository; [cité le 20 juill 2021]. Disponible: https://archive.ics.uci.edu/datasets
[8] http://www.om-db.wi.tum.de/psplib/

conférences internationales telles que NeurIPS[9]. Lors de ces conférences il est fortement recommandé d'associer ses données aux résultats afin de s'assurer de la reproductibilité de la recherche, cela assure aussi la reconnaissance des travaux de recherche donc une certaine légitimité au chercheur. L'association du code et de l'algorithme est aussi perçue comme une suite logique de publication des résultats, généralement les chercheurs déposent leurs publications sur des serveurs *open source* tels que l'archive ouverte HAL qui est la plateforme de dépôt en ligne développée par le Centre pour la Communication Scientifique Directe (CCSD) en 2001 ou encore l'archive ouverte de prépublications arXiv, les serveurs universitaires comme OATAO[10]. Dans un bon nombre de cas, les chercheurs poussent les doctorants à publier en *open source* en ajoutant le code à l'article ce qui peut aider à la publication, comme en témoigne ce chercheur ANITI : « *Disons que c'est un argument qui permet de convaincre un éditeur d'accepter de publier le papier aussi* ». Tout en désirant se faire citer si jamais le travail est réutilisé, dans ce cas-là une licence peut s'appliquer.

2. Vision transversale de la donnée : les chercheurs face au PGD

Les mouvances, les changements de politique autour des données de la recherche, leur ouverture, leur partage sont bien connus des professionnels de l'information scientifique et technique. En étant les premiers concernés par ces évolutions, les chercheurs restent encore peu avertis pour la plupart à moins d'avoir commencé un projet de recherche nécessitant un PGD. Pour les chercheurs l'élaboration d'un PGD est une nouveauté tout comme l'organisation en amont du cycle de vie de la donnée. Cette partie évoque la vision transversale qui existe de la donnée entre professionnels de l'IST et les chercheurs, mais également entre les chercheurs eux-mêmes en ce qui concerne le PGD.

La majeure partie des chercheurs rencontrées lors des entretiens - soit plus de la moitié - n'avaient jamais entendu parler du PGD ou de ce qu'il représentait et ne connaissaient encore moins les enjeux que ce document représentait. Nous avons donc pris le temps au début des entretiens, d'expliquer en quoi les chercheurs étaient désormais dans l'obligation de remettre ce livrable pour tout projet financé par l'ANR, et également en quoi le PGD pouvait leur être utile, non seulement sur des aspects concernant le financement de leurs recherches. Certains chercheurs étaient engagés dans d'autres projets de recherche en parallèle nécessitant PGD, dans ce cas-là nous les avons aidés dans l'élaboration du PGD en leur transmettant la grille d'entretien et en répondant à toutes leurs questions concernant la rédaction. Néanmoins et dans tous les cas, la rédaction du plan de gestion de données était perçue comme une contrainte administrative s'ajoutant au travail du chercheur et non comme un modèle positif d'ouverture des données comme en témoigne ce chercheur d'ANITI : « *C'est une des choses requises par les États/politiques qu'on peut pleinement qualifier dans la plupart des cas comme un "waste of time and complete pain in the ass."* ». Cet argument est très révélateur de la différence entre ce que les politiques de science ouverte envisagent de ce qu'est un PGD, de

[9] « *Neural Information Processing Systems Foundation* est une société à but non lucratif dont le but est de favoriser l'échange des avancées de la recherche en intelligence artificielle et en apprentissage automatique, principalement en organisant une conférence universitaire interdisciplinaire annuelle avec les normes éthiques les plus élevées pour une communauté diversifiée et inclusive » https://nips.cc/

[10] Open Archive Toulouse Archive Ouverte https://oatao.univ-toulouse.fr/

ce qu'il représente et ce que les chercheurs eux-mêmes en pensent, eux dont la une réalité de terrain est toute autre. Et cet argument n'est pas isolé : un bon nombre de chercheurs ressentent la même chose comme cet autre chercheur d'ANITI : « *c'est un nouvel exercice inventé par les sphères supérieures, qui est à la fois pénible, obligatoire et pas universellement utile.* ». L'utilité du PGD est souvent remise en cause par les chercheurs car la plupart du temps il n'y a pas de réflexion en amont sur la manière dont les données produites sont gérées, stockées. Ce qui est directement en lien avec le manque d'information, de formation et d'expérience sur le PGD qui est censé aider à la réflexion et l'organisation des données. L'autre perception qui ressort c'est le sentiment de travailler à sens unique sur ce livrable. En effet, le PGD demandé par l'ANR est un document qui fait l'objet de pas ou peu de retours de la part de l'agence de financement. Ce travail n'est pas non plus valorisé pour le moment ; il est seulement obligatoire pour mener son projet de recherche à bien et bénéficier de la totalité de son financement de recherche. Les professionnels de l'IST préconisent la publication d'un *data paper* pour obtenir une forme de reconnaissance du travail du PGD (Puren, 2019) mais pour l'instant celui-ci n'est pas valorisé.

Les chercheurs ont parfois besoin d'aide dans leurs méthodes de gestion, partage et diffusion des données. En effet, dans plusieurs entretiens il est revenu des questions sur les types de licences applicables aux données, mais également sur les entrepôts : où déposer les données de manière pérenne ? ou encore des questions sur le stockage des données. En soi, malgré la volonté d'ouvrir les données et de développer de nouvelles politiques d'ouverture les chercheurs ne sont pas suffisamment formés et informés à la gestion des données et ce n'est pas le cas uniquement pour les chercheurs de la communauté IA, comme évoquait Smale et al. 2018 :

> *« Researchers report feeling as though they need help with data management, sharing, and archiving (Brandt, 2007). There likely is a service gap in delivering education in data management skills. This gap may be leading researchers to engage in poor data management practices. »*
>
> *Traduction libre : « Les chercheurs disent avoir l'impression d'avoir besoin d'aide pour la gestion, le partage et l'archivage des données (Brandt, 2007). Il est probable qu'il existe des lacunes dans la prestation de services d'éducation en matière de gestion des données. Ces lacunes peuvent conduire les chercheurs à adopter de mauvaises pratiques de gestion des données. »*

Depuis quelques années et face aux besoins des chercheurs, des services d'accompagnement ont été mis en place par les pôles IST et les bibliothèques universitaires, qui à la différence d'un système de cotutelle comme l'URFIST et ANITI, guident mais ne rédigent pas le PGD. Parmi ces services on trouve les services institutionnels recensés dans le répertoire[11] des Services Opérationnels de Soutien à la rédaction des Plans de Gestion des Données (SOS-PGD). Ces services d'accompagnement sont là pour leur expliquer ce qu'est le PGD et ses enjeux, répondre aux besoins des chercheurs et à leurs questions, mais ne vont

[11] https://scienceouverte.couperin.org/sos-pgd/

pas jusqu'à rédiger le PGD soit par manque de temps, soit par volonté de ne pas substituer au chercheur qui connait ses données mieux que quiconque. C'est dans ce sens que d'autres services d'accompagnement ont vu le jour, allant jusqu'à la construction et la rédaction du PGD pour pallier les besoins des chercheurs. C'est ce que révèle la Semaine Data SHS avec la table ronde "Retours d'expérience autour du plan de gestion de données en sciences humaines et sociales", organisée par la MSH Lyon St-Etienne le 11 décembre 2020[12]. Tout comme la cotutelle de l'URFIST et d'ANITI, des équipes de professionnels se mobilisent pour accompagner et informer, ainsi les chercheurs se voient décharger de la rédaction du PGD puisque cette tâche peut être déléguée. D'autres services existent comme le Lyon Ingénierie Projets (LIP) qui proposent leur expertise dans la gestion de projet à des chercheurs, enseignants-chercheurs, entreprises ou encore à des organismes financiers et de les accompagner dans leurs projets, prenant notamment en compte la rédaction de PGD. Ces services seront d'ailleurs probablement amenés à se développer dans les prochaines années et les compétences en gestion, diffusion des données et rédaction de PGD de plus en plus demandés dans les recrutements (exemple des profils de *data scientist, data librarian*).

[12] https://www.msh-lse.fr/semaine-data-shs-2020/

IV. Favoriser les pratiques de gestion et diffusion des données grâce au PGD : limites et propositions

Cette partie décrit mon expérience autour du modèle de plan de gestion des données de l'ANR, ainsi que de l'outil de rédaction DMP OPIDoR en évoquant notamment les limites rencontrées au cours du stage et d'éventuelles propositions. Enfin cette partie montre comment le PGD a-t-il pu faire valoriser les pratiques des chercheurs d'ANITI.

1. Limites du plan de gestion des données de l'ANR

Au cours des entretiens nous avons échangé avec des chercheurs ayant réalisé des PGD dans le cadre de projets H2020 et de projets ANR. Leurs retours étaient unanimes sur le fait que la rédaction de PGD au sein des projets H2020 était plus limpide, car ceux-ci sont relus et un retour de la part des agences de financement est envoyé sur chaque version du PGD, ne laissant pas l'impression au chercheur de travailler en sens unique.

De plus, sur la plateforme européenne OpenAIRE, un regroupement de plus de 800 PGD[13] comprenant les différentes versions suivant la vie du projet est consultable[14]. Ces PGD ont été réalisés dans le cadre de projets H2020 et permettent grâce à leur libre consultation dans l'archive ouverte de l'Université de Vienne, de pouvoir se faire une idée de comment rédiger un bon PGD. Dans le cadre de mon stage et pour utiliser cette ressource, j'ai créé un tableau répertoriant plus de 30 « bons » PGD avec parfois les 3 versions d'un même projet, afin que ceux-ci puissent être exploités par d'autres chercheurs (cf. Annexe 5 : Tableau DMP H2020). Ce tableau ne classe pas particulièrement les PGD mais permet de connaître leurs noms, la version, le sujet, le nombre de pages et d'y accéder directement à partir du lien URL. Avoir accès aux différentes versions d'un PGD permet ainsi de voir les améliorations ou corrections potentielles effectuées entre les phases des projets. Dans le cadre des PGD ANR il serait potentiellement utile de pouvoir accéder à ce même type de ressource.

Si l'on se met à la place du chercheur qui ne connaît pas parfaitement les exigences de l'ANR mais qui doit constituer un PGD au début de son projet, le fait qu'il n'existe pas d'exemple précis de PGD avec vérification et validation des organismes financeurs ne favorise pas la rédaction de « bon » PGD. Sans retour de la part de l'ANR sur les premières versions des PGD et sans exemple concret disponible publiquement pour donner un ordre d'idée de ce qui est attendu, un chercheur peut se sentir plus ou moins désemparé sur ce document qui conditionne son financement. Sur un autre outil comme DMP Tool un challenge de « meilleur » PGD[15] a été mené par un jury de bibliothécaires en élisant les 10 meilleurs PGD. Ceux-ci sont disponibles publiquement sur des entrepôts comme Zenodo ou direction avec l'URL du PGD, ils sont accompagnés d'un commentaire de la part du jury

[13] https://www.openaire.eu/blogs/establishing-a-collection-of-841-horizon-2020-data-management-plans
[14] *Open Access Infrastructure of Research in Europe* https://www.openaire.eu/mission-and-vision
[15] https://blog.dmptool.org/2021/05/19/dmp-competition-winners-dmps-so-good-they-go-to-11/

et permet aux autres chercheurs d'accéder aux PGD. La possibilité de publier ses PGD avec un retour dessus permettrait de valoriser ce travail.

Un autre point à relever est que le travail de rédaction du plan de gestion des données ANR n'est actuellement pas mis en valeur et malgré les témoignages d'aide à la gestion et diffusion il reste bel et bien une contrainte administrative pour les chercheurs. La volonté des chercheurs en IA de vouloir partager et ouvrir leurs données est néanmoins présente et constitue une suite logique dans leur publication. Comme l'atteste l'extrait traduit ci-dessous :

> *« Though encouraging sharing of data is one of the main drivers of funding bodies mandating DMPs, evidence of translation from DMP completion to better managed data to more shared data is as yet untested which raises the question – are there alternate mechanisms besides DMPs that can provide a better means by which data sharing is achieved ? »*
>
> *Traduction libre : « Bien que l'encouragement du partage des données soit l'une des principales raisons pour lesquelles les organismes de financement rendent les DMP obligatoires, la preuve de la traduction de l'achèvement du DMP en une meilleure gestion des données et un plus grand partage des données n'a pas encore été vérifiée, ce qui soulève la question suivante : existe-t-il des mécanismes autres que les DMP qui peuvent fournir un meilleur moyen de réaliser le partage des données ? »* (Smale al. 2018)

Une solution pourrait être envisagée notamment avec la publication de jeux de données dans les *data paper* si tant est qu'ils soient reconnus comme de vrais articles scientifiques, entrant dans l'évaluation des chercheurs. Les PGD pourraient constituer une trame de travail aidant les chercheurs à se familiariser avec les principes FAIR et les orientant dans leurs choix pour ensuite publier leurs données dans des *data paper* qui, eux, seront relus et pris en compte, contribuant ainsi à développer plus encore la valeur d'échange des données (Reymonet, 2017).

2. DMP OPIDoR : un outil de rédaction peu modulable

Le premier questionnement autour du projet PGD ANITI a été la structuration du PGD, comment allions-nous organiser le PGD pour que tous les jeux de données soient représentés et que la pluridisciplinarité ressorte de ce document ? L'outil DMP OPIDoR propose une organisation par « produit de recherche » (cf. Annexe 6) donc plutôt par jeux de données individuels. Afin de mettre en valeur la pluridisciplinarité de l'institut, nous avons organisé les jeux de données en commençant par les jeux les plus réutilisés au sein des chaires, en allant par ordre décroissant jusqu'aux jeux les plus spécifiques comme indiqué dans la partie méthodologie. Une recommandation de la part de la direction d'ANITI nous a été communiqué de séparer les jeux de données réutilisés, produits, aux jeux de données industriels mais malheureusement nous ne pouvons pas procéder en les regroupant sur l'outil ce qui oblige à préciser dans le nom et les descriptions la spécialisation du jeu de données

pour pouvoir le différencier. Lors de la réflexion autour de la structuration du PGD nous avons cherché des exemples de PGD qui avaient été rédigés dans le cadre d'un financement ANR pour nous donner une idée de ce que pouvait être un « bon » PGD. Hormis les recommandations de l'ANR[16] sur le modèle à utiliser et la rédaction, il n'existe que très peu d'exemples de PGD rendus publics. Sur DMP OPIDoR il existait 17 DMP publics du modèle ANR qu'il est possible de consulter. Même si DMP OPIDoR n'a pas vocation à conserver les PGD sur la plateforme, une nouvelle catégorie de documents est prévue sur l'archive ouverte HAL afin d'accueillir les PGD à terme. Dans l'outil DMP OPIDoR, il est précisé que ces DMP sont partagés publiquement par leurs propriétaires et qu'ils n'ont pas été vérifiés pour leur qualité, leur exhaustivité ou leur adhésion aux lignes directrices des financeurs. En effet, la plupart ne font qu'une dizaine de page à peine sans exhaustivité dans les réponses.

Lors d'une formation OpenAIRE en juillet 2021, un autre outil de rédaction nous a été présenté, l'outil Argos[17] : Argos est un service qui permet la création de plans de gestion des données automatisés, guidant les chercheurs vers les principes FAIR et configurable en fonction de la discipline de recherche. Cet outil n'est pas encore disponible pour les modèles ANR français mais a pour objectif de se développer en une version française compatible avec les modèles des certains organismes de recherche français et d'établir une intégration à l'archive ouverte HAL également. Un futur et deuxième outil verrait ainsi le jour pour les projets chargés de rédiger un PGD dans le cadre de financement ANR.

3. Valorisation des pratiques des chercheurs, le guide des bonnes pratiques

Le plan de gestion des données et les entretiens avec les chercheurs auront permis de cibler les points où il y a un manque d'information et de connaissance pour une bonne gestion et diffusion des données. Cibler ces points à ANITI et à l'URFIST d'avoir une vue d'ensemble sur les éventuelles zones à éclaircir et les formations ou des ressources numériques en ligne à proposer. Par exemple pendant mon stage j'ai pu réadapter en français une infographie (cf. Annexe 7) sur les coûts des données de la recherche, qui était disponible sur la plateforme OpenAIRE. Cette infographie peut aider les chercheurs. Elle ne se limite d'ailleurs pas aux chercheurs en intelligence artificielle puisqu'elle peut concerner tous les chercheurs utilisant de la donnée. Elle permet d'estimer en amont les coûts à prévoir pour la gestion et le stockage des données.

Tout au long des entretiens nous avons récupéré les différentes questions dans un tableau de recensement des besoins (cf. Annexe 8) dédié aux chercheurs d'ANITI pour les aider dans leurs démarches d'ouverture des données. La production du livrable des bonnes pratiques permettra aux chercheurs d'instaurer plus facilement une gestion compatible avec les principes FAIR et les aidera à ouvrir leurs données que ce soit sur les questions de licences,

[16] https://anr.fr/fileadmin/documents/2019/ANR-modele-PGD.pdf
[17] https://argos.openaire.eu/splash/

de coûts de stockage ou d'entrepôts pour assurer la pérennité des données. Lorsque nous évoquions avec les chercheurs la question des licences appliquées aux jeux de données réutilisés ou produits, ils ne savaient pas quelles licences étaient appliquées ou à quoi elles correspondaient. De même pour les entrepôts de données, certains connaissaient déjà des entrepôts adaptés à leur discipline mais d'autres non, ce qui nous donnait la possibilité de discuter avec eux des entrepôts et des critères de choix en leur conseillant de privilégier les entrepôts disciplinaires. L'articulation du livrable des bonnes pratiques se fait en trois parties : une partie de rappel sur les enjeux et principes FAIR suivie d'une partie englobant tous les contacts ANITI que ce soit pour les outils mis à disposition pour les chercheurs d'ANITI (lieu de stockage, lieu de partage etc…), les personnes à contacter pour les problématiques RGPD, pour les achats de jeux de données etc… La dernière partie est une foire aux questions inspirée du tableau de recensement des besoins complété au fur et à mesure des entretiens. Ce livrable sera par la suite prolongé par un module de type étude de cas (cf. Annexe 9), destiné à conseiller les chercheurs de la communauté d'IA sur le choix du format des données, de l'entrepôt, des espaces de stockage selon la volumétrie et des licences.

Conclusion

La réalisation de ce stage de cinq mois aura permis la rédaction de la première version du plan de gestion des données d'ANITI et aura également pu nourrir une réflexion critique sur le modèle de l'ANR quant à la mise en valeur des pratiques de gestion des données des chercheurs de la communauté IA. À travers ce mémoire, nous avons souhaité montrer en quoi un PGD permet-il de saisir les pratiques pluridisciplinaires des chercheurs en IA et de favoriser les bonnes pratiques de gestion et diffusion des données de la recherche ? Plusieurs points ont été abordés. Nous avons d'abord vu que, grâce à la méthodologie employée et au travail de rédaction, la pluridisciplinarité de l'institut ANITI pouvait être mise en valeur au travers des différents jeux de données recensés. La communauté en intelligence artificielle est une communauté hétérogène aux cultures et à la gestion des données plurielles, rendant le PGD lui-même hétérogène sur la représentation des jeux de données qu'il décrit : jeux de données images, enregistrement sonores, jeux de donnés produits et réutilisés en libre accès ou encore des jeux de données appartenant aux partenaires industriels privés non diffusables. Nous avons ensuite fait le constat qu'il existe une hétérogénéité dans les pratiques mêmes de gestion des données : les jeux de données sont déposés dans des entrepôts disciplinaires respectifs en fonction des disciplines de recherche, rappelant ainsi que la communauté IA rassemble en son sein plusieurs communautés ayant leurs habitudes propres de partage et de diffusion des données. Enfin nous avons montré comment, en réalisant les entretiens avec les chercheurs d'ANITI nous avons également pu les accompagner et les orienter dans la gestion de leurs données en leur faisant part des bonnes pratiques de gestion. Grâce aux échanges avec les chercheurs, nous avons pu recenser leurs besoins et ainsi proposer un modèle de guide des bonnes pratiques ainsi qu'une étude de cas. Ces deux livrables pourront leurs être communiqués, afin qu'ils s'approprient les bonnes pratiques de gestion en rendant leurs données les plus FAIR possible.

Le PGD n'est donc pas seulement un livrable destiné aux agences de financement, puisqu'autour de la rédaction de ce livrable une discussion à double sens entre les professionnels de l'IST et les chercheurs ANITI a pu s'instaurer : d'un côté avec l'équipe projet nous avons pu rédiger le PGD grâce à l'aide des chercheurs et de l'autre nous avons pu répondre à leurs demandes en les aidant à mieux diffuser et gérer leurs données.

Sur le plan personnel, le stage m'a fait prendre conscience de la complexité d'une communauté de recherche comme celle des chercheurs en intelligence artificielle. En se voulant pluridisciplinaire, l'institut ANITI fait face à de nombreux enjeux notamment ceux de rassembler au sein d'un même projets des communautés distinctes et de rendre la collaboration efficace sur ce projet d'envergure. Les échanges avec les chercheurs et professionnels de l'IST m'ont permis de consolider mes connaissances développées lors de mon Master 1 et de mettre en pratique les aspects théoriques de la gestion des données de la recherche.

Bibliographie

BECKOUCHE, Pierre, 2017. La révolution numérique est-elle un tournant anthropologique ? *Le Débat* [en ligne]. 2017. Vol. 193, n° 1, pp. 153-166. [Consulté le 12 mai 2021] DOI. Disponible à l'adresse : https://www.cairn.info/revue-le-debat-2017-1-page-153.htm

BORGMAN, Christine L., 2012. The conundrum of sharing research data. [en ligne]. 2012. [Consulté le 4 janvier 2021]. Disponible à l'adresse : https://onlinelibrary-wiley-com.docelec.univ-lyon1.fr/doi/10.1002/asi.22634

BORGMAN, Christine L., 2020. *Qu'est-ce que le travail scientifique des données ? : Big data, little data, no data* [en ligne]. Marseille : OpenEdition Press. [Consulté le 4 janvier 2021]. Encyclopédie numérique. ISBN 979-10-365-6541-0. Disponible à l'adresse : http://books.openedition.org/oep/14692

CABRERA, Francisca, 2015. *Les données de la recherche en Sciences humaines et sociales : enjeux et pratiques. Enquête exploratoire.* [en ligne]. other. Institut National des techniques et sciences de la documentation - Conservatoire National des Arts et des Métiers ; CNAM, Paris. [Consulté le 5 juillet 2021]. Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_01117375

COMBESSIE, Jean-Claude, 2007. II. L'entretien semi-directif. In : [en ligne]. Paris : La Découverte. pp. 24-32. Repères. [Consulté le 4 avril 2021]. ISBN 978-2-7071-5241-1. Disponible à l'adresse : https://www.cairn.info/la-methode-en-sociologie--9782707152411-p-24.htm

DÄLLENBACH, par David et THONNEY, Delphine, 2019. Data Management Plan : retour d'expérience. *Recherche d'IdéeS* [en ligne]. 29 octobre 2019. [Consulté le 17 juillet 2021]. Disponible à l'adresse : https://campus.hesge.ch/blog-master-is/data-management-plan-retour-dexperience/

KOULOCHERI, Eleni, 2017. What is the EC Open Research Data Pilot? *OpenAIRE* [en ligne]. 22 novembre 2017. [Consulté le 20 juillet 2021]. Disponible à l'adresse : https://www.openaire.eu/what-is-the-open-research-data-pilot

LEONELLI, Sabina, 2015. What Counts as Scientific Data? A Relational Framework |. *Philosophy of Science* [en ligne]. décembre 2015. Vol. Volume 82, n° 5. [Consulté le 22 novembre 2022]. Disponible à l'adresse : https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/what-counts-as-scientific-data-a-relational-framework/33A05C7F71958D1FEC37AE89C2866A72

LORD, Philip et MACDONALD, Alison, 2003. The JISC Committee for the Support of Research (JCSR). . 2003. pp. 85.

NEYLON, Cameron, 2017. Compliance Culture or Culture Change ? The role of funders in improving data management and sharing practice amongst researchers. *Research Ideas and Outcomes* [en ligne]. 19 octobre 2017. Vol. 3, pp. e21705. [Consulté le 14 août 2021]. DOI 10.3897/rio.3.e21705. Disponible à l'adresse : https://riojournal.com/article/21705/

PAIN, Marilou, 2016. *Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation : étude de cas pour l'archive HAL* [en ligne]. other. Enssib. [Consulté le 12 mai 2021]. Disponible à l'adresse : https://memsic.ccsd.cnrs.fr/mem_01374509

PUREN, Marie, 2019. *Le data paper* [en ligne]. École thématique. Paris, France : EHESS - Paris. [Consulté le 12 mai 2021]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-03185756

REYMONET, Nathalie, 2017. Améliorer l'exposition des données de la recherche : la publication de data papers. 6 janvier 2017.

ROBIN, Agnès, 2017. Les données scientifiques au prisme du dispositif open data. Communication Commerce Électronique [en ligne]. Septembre 2017. Vol. 9, n° étude 14, pp. 7. [Consulté le 2 juillet 2022]. Disponible à l'adresse : https://hal.umontpellier.fr/hal-01845205

SMALE, Nicholas, UNSWORTH, Kathryn, DENYER, Gareth et BARR, Daniel, 2018. *The History, Advocacy and Efficacy of Data Management Plans* [en ligne]. 17 octobre 2018. bioRxiv. [Consulté le 7 avril 2021]. Disponible à l'adresse : https://www.biorxiv.org/content/10.1101/443499v1

TORTERAT, Gwendoline, 2020. Rédaction d'un plan de gestion de données : retour d'expérience. Partie 2. *Le blog d'Huma-Num et de ses consortiums* [en ligne]. 2020. [Consulté le 12 mai 2021]. Disponible à l'adresse : https://humanum.hypotheses.org/6196

WALLIS, Jillian C., ROLANDO, Elizabeth et BORGMAN, Christine L., 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE* [en ligne]. juil 2013. Vol. 8, n° 7, pp. e67332. [Consulté le 12 juillet]. DOI 10.1371/journal.pone.0067332. Disponible à l'adresse : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332

WILKINSON, Mark D., DUMONTIER, Michel, AALBERSBERG, IJsbrand Jan, APPLETON, Gabrielle, AXTON, Myles, BAAK, Arie, BLOMBERG, Niklas, BOITEN, Jan-Willem, DA SILVA SANTOS, Luiz Bonino, BOURNE, Philip E., BOUWMAN, Jildau, BROOKES, Anthony J., CLARK, Tim, CROSAS, Mercè, DILLO, Ingrid, DUMON, Olivier, EDMUNDS, Scott, EVELO, Chris T., FINKERS, Richard, GONZALEZ-BELTRAN, Alejandra, GRAY, Alasdair J. G., GROTH, Paul, GOBLE, Carole, GRETHE, Jeffrey S., HERINGA, Jaap, 'T HOEN, Peter A. C., HOOFT, Rob, KUHN, Tobias, KOK, Ruben, KOK, Joost, LUSHER, Scott J., MARTONE, Maryann E., MONS, Albert, PACKER, Abel L., PERSSON, Bengt, ROCCA-SERRA, Philippe, ROOS, Marco,

VAN SCHAIK, Rene, SANSONE, Susanna-Assunta, SCHULTES, Erik, SENGSTAG, Thierry, SLATER, Ted, STRAWN, George, SWERTZ, Morris A., THOMPSON, Mark, VAN DER LEI, Johan, VAN MULLIGEN, Erik, VELTEROP, Jan, WAAGMEESTER, Andra, WITTENBURG, Peter, WOLSTENCROFT, Katherine, ZHAO, Jun et MONS, Barend, 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [en ligne]. 15 mars 2016. Vol. 3, n° 1, pp. 160018. [Consulté le 5 mai 2021]. DOI 10.1038/sdata.2016.18. Disponible à l'adresse : https://www.nature.com/articles/sdata201618

Sitographie

CéSO - Comité de réflexion pour la science ouverte | Université de Toulouse. [en ligne]. [Consulté le 23 mai 2021]. Disponible à l'adresse : https://www.univ-toulouse.fr/recherche-dynamique/ceso-comite-de-reflexion-pour-science-ouverte

CIFAR-10 and CIFAR-100 datasets. [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://www.cs.toronto.edu/~kriz/cifar.html

Datasets - UCI Machine Learning Repository. [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://archive.ics.uci.edu/datasets

DPO et RGPD : définition, rôle & missions. Data Legal Drive [en ligne]. [Consulté le 23 juillet 2022]. Disponible à l'adresse : https://datalegaldrive.com/dpo/definition-missions/

Establishing a collection of 841 publicly available Horizon 2020 Data Management Plans, 2021. OpenAIRE [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://www.openaire.eu/blogs/establishing-a-collection-of-841-horizon-2020-data-management-plans

Kaggle: Your Machine Learning and Data Science Community. [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://www.kaggle.com/

MARIAPRAETZELLIS, 2021. DMP Competition Winners: DMPs so good they go to 11. DMPTool Blog [en ligne]. 19 mai 2021. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://blog.dmptool.org/2021/05/19/dmp-competition-winners-dmps-so-good-they-go-to-11/

OpenNeuro. [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://openneuro.org/

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (2004). Déclaration sur l'accès aux données de la recherche financée par des fonds publics. https://legalinstruments.oecd.org/fr/instruments/157

Plan national pour la Science Ouverte. In : [en ligne]. [Consulté le 21 mai 2021]. Disponible à l'adresse : https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte.

Plan de gestion de données – DoRANum. [en ligne]. [Consulté le 23 mai 2021]. Disponible à l'adresse : https://doranum.fr/plan-gestion-donnees-dmp/

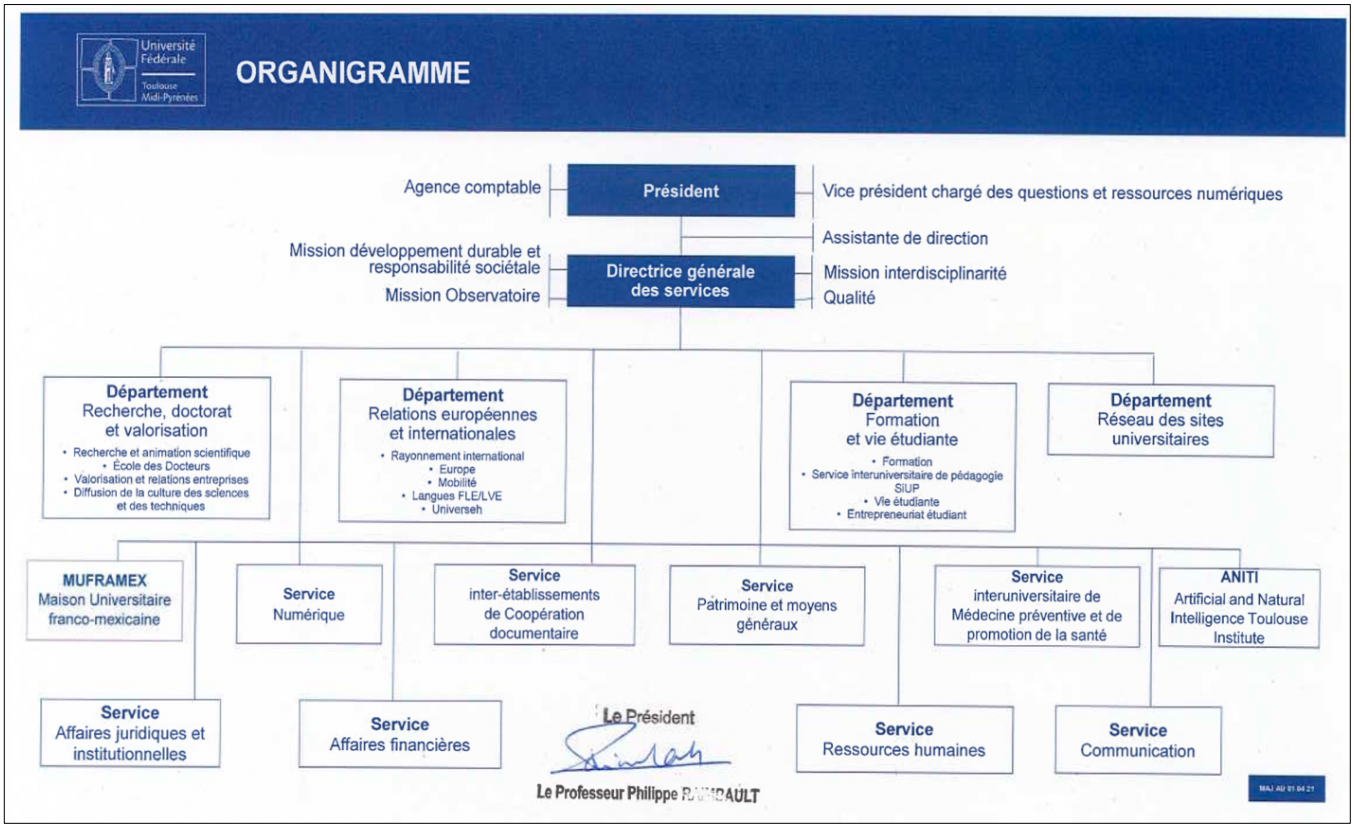
SOS-PGD – Science ouverte France. [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://scienceouverte.couperin.org/sos-pgd/

Welcome to OATAO (Open Archive Toulouse Archive Ouverte) - oatao. [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://oatao.univ-toulouse.fr/

Welcome to PSPLIB. [en ligne]. [Consulté le 20 juin 2023]. Disponible à l'adresse : https://www.om-db.wi.tum.de/psplib/main.html

Annexes

Annexe 1 : Organigramme ANITI



Annexe 2 : Lettre de mission ANITI

ANITI — Université Fédérale Toulouse Midi-Pyrénées — TOULOUSE INSTITUTE

À Toulouse, le 8 décembre 2020

RÉDACTEUR : Corinne JOFFRE
SOUS-COUVERT DE : Odile JANKOWIAK-GRATTON
COPIE POUR INFO : Catherine ROUSSY, Guillaume SIRE, Copil CéSO

Objet : Lettre de mission relative à la conception du Plan de Gestion des Données ANITI

La convention attributive d'aide ANR n°ANR-19-PI3A-0004 pour ANITI prévoit à l'article 9.5 une clause sur le Plan de Gestion des Données (PGD). Ce document doit être fourni à l'ANR à trois temps de l'avancement du projet : tout d'abord une version initiale durant le premier semestre après le démarrage du projet, une mise à jour à mi-parcours puis une version à la date de fin de projet. Le porteur de projet est libre d'utiliser le modèle de son choix, l'ANR ayant proposé un modèle accessible sur le site DMP OPIDoR. L'enjeu est fort car la convention précise que la non-transmission d'un tel document peut conduire à l'interruption du versement de l'aide conformément aux dispositions prévues à l'Article 13 de la Convention.

ANITI est articulé autour de 24 chaires de recherche, rattachées à 11 établissements d'enseignement supérieur et de recherche, avec des pratiques en matière de gestion de données propres et plus ou moins avancées. ANITI, de par son domaine scientifique, est amené à traiter de nombreux jeux de données, y compris des jeux de données industrielles. En termes de méthode, l'UFTMiP n'ayant pas eu l'occasion de définir de pratiques en matière de gestion des données, il est proposé de construire le PGD ANITI en concertation avec les établissements impliqués pour capitaliser sur les pratiques déjà en place et proposer des compléments là où ce serait pertinent.

L'essentiel de la mission porte sur le recueil des pratiques auprès des 24 chaires ANITI, dont les disciplines sont très variées (informatique, mathématique, neurosciences, SHS, droit, etc.), et le pilotage de cette approche collaborative avec les établissements académiques partenaires. Il s'agira de concevoir et de conduire des entretiens visant à établir une solide vision des données concernées par le PGD et des pratiques associées puis d'en produire une analyse permettant de (i) collecter les informations et rédiger le PGD ANITI et (ii) identifier ce qui doit être mis en place pour la bonne gestion des données (stockage sécurisé et pérenne, ouverture des données, « FAIRisation » des données, etc.) pour les faire remonter à ANITI et au CéSO.

Amélie Barrio, co-responsable de l'Unité Régionale de Formation à l'Information Scientifique et Technique (URFIST) Occitanie, dispose des compétences clés ainsi que du réseau de contacts pertinents dans les établissements partenaires pour mener à bien cette mission.

C'est pourquoi nous lui confions cette mission qui se déroulera du 1er janvier 2021 au 31 décembre 2023, à temps partiel avec l'appui temporaire de stagiaires financés par ANITI.

Un point d'avancement sera fait régulièrement avec la secrétaire générale d'ANITI, Corinne Joffre, qui sera son interlocutrice sur cette mission.

Odile Jankowiak-Gratton
Directrice générale des services

Université Fédérale de Toulouse

Signature

Annexe 3 : Grille d'entretien chercheurs

Grille commune de recension des jeux de données

Données réutilisées	Objectif dans le cadre du projet	Contextualiser la réutilisation jeux de données par rapport au projet	
	Origine des données	Licence de réutilisation Paternité	
	Type de données	Numérique (base de données, tableurs), textuel (documents), image, audio, vidéo et/ ou médias composites	
	Nature de données	Nature selon le jeu de données (algorithmes..)	
	Format des données	Manière dont les données sont codées pour le stockage (pdf xls, doc txt, rdf, autre)	
	Volumétrie des données	Espace de stockage requis (octets), quantité de fichiers	
	Logiciels de réutilisation	Un logiciel est-il nécessaire à l'utilisation des données	
	Standard de métadonnées	Type de standard utilisé	
Données générées	Contexte de collecte	Contexte au sein du projet	
	Méthode de collecte	Méthodologies, logiciels utilisés pour recueillir de nouvelles données	
	Types de données	Numérique (base de données, tableurs), textuel (documents), image, audio, vidéo et/ ou médias composites	
	Nature des données	Nature selon le jeu de données (algorithmes..)	
	Format des données	Manière dont les données sont codées pour le stockage (pdf xls, doc txt,rdf, autre)	
	Volumétrie des données	Espace de stockage requis (octets), quantité de fichiers	
	Logiciels de réutilisation	Logiciel nécessaire ou non à la réutilisation de la donnée, quel type ?	
	Propriété intellectuelle	Propriété industrielle, littéraire et artistique	
Stockage des données	Lieu, Cloud	Décrire l'endroit où les données sont sauvegardées au cours du processus de recherche	

	Fréquence de sauvegarde	Backup de l'ensemble des données et métadonnées	
	Nommage	Convention de nommage (exemple recommandé sur le site DoRaNum)	
	Droit d'accès	Qui aura accès aux données et comment l'accès est contrôlé	
	Récupération en cas d'accident	Plateforme de sauvegarde	
Sécurité	Identification des données sensibles	Données soumises au secret, protection du patrimoine scientifique et technologique	
	Politique de protection	Droit d'usage/Droit d'accès Politique institutionnelle de protection	
	Processus de cryptage	Procédures de cryptage pour sécuriser les documents	
Diffusion des données	Diffusable ou non	Licence de diffusion *Creative Commons*	
	Lieu de diffusion	Entrepôt (*data requisitory*), article de recherche	

	Principes FAIR (Findable, Accessible, Interoperable, Reusable)	Principes FAIR et ouverture des données Identifiants pérennes Standard métadonnée	
	Question des data paper (valorisation de la donnée)	Publication dans un article type data paper ou non	
Conservation des données	Lieu de conservation	Dépôt dans un entrepôt (zenodo, figshare, mendeley data, dryad, research data) ou archive	
	Stockage	Plan de préservation des données Durée d'archivage pérenne des données	
	Embargo	Délais de publication et disponibilité	
Coûts	Coût de la conservation	Frais d'entrepôt, d'archivage	
	Coût du stockage	Frais de stockage, coût matériel	
	Coût en terme de ressources humaines	Temps personnel	
	Achat de jeux de données spécifiques	Coût de l'achat de jeux de données nécessaires à la recherche	

Annexe 4 : Modèle d'accord de consentement

FORMULAIRE DE CONSENTEMENT POUR LA PARTICIPATION A UNE ENQUÊTE

TITRE DE LA RECHERCHE : Plan de gestion de données ANITI

Établissement responsable de l'enquête : Université fédérale de Toulouse Midi-Pyrénées (UFTMiP)

Coordinatrice du projet : Amélie BARRIO, URFIST, amelie.barrio@univ-toulouse.fr

Je soussigné(e), Madame/Monsieur _____, exerçant la fonction de _____, au sein de _____, déclare accepter, librement et de façon éclairée de participer à la réalisation de l'enquête ayant pour objectif la rédaction du Plan de gestion des données (PDG) de l'institut interdisciplinaire en intelligence artificielle (ANITI).

La rédaction d'un PGD est un vrai défi pour tenir compte des différentes cultures disciplinaires, des outils et méthodes mises en place par les établissements, mais aussi des pratiques de chacun. La réalisation du présent PGD passe par des entretiens qualitatifs menés par une équipe de professionnels de l'information afin de pouvoir remplir les indications demandées par l'Agence Nationale de la Recherche (ANR). Ces entretiens seront ensuite analysés par l'équipe projet. Pour ce faire, les entretiens seront enregistrés puis transcrits pour pouvoir en garder une trace et permettre à l'ensemble de l'équipe projet d'en prendre connaissance. Une fois transcrits, les entretiens seront pseudonymisés et conservés jusqu'au rendu du PGD définitif en 2024.

En acceptant de participer à l'enquête, je m'engage à répondre à des questions posées en entretien en présentiel (enregistrement audio) ou à distance (enregistrement vidéo), concernant la gestion et diffusion éventuelle de données de recherche de l'institut ANITI.

J'ai eu la possibilité de poser toutes les questions que je souhaitais à l'équipe projet qui m'a expliqué la nature, les objectifs, les risques potentiels et les contraintes liées à ma participation à cette enquête. Je pourrai à tout moment demander des informations complémentaires à l'UFTMiP qui m'a proposé de participer à cette enquête, centralisées aux coordonnées suivantes : urfist@univ-toulouse.fr.

Je connais la possibilité qui m'est réservée d'interrompre ma participation à cette enquête à tout moment sans avoir à justifier ma décision et je ferai mon possible pour en informer Amélie BARRIO, dont les coordonnées se trouvent ci-dessus.

Au cours de cette enquête, j'accepte que soient recueillies des données personnelles (nom, prénom, adresse mail, chaire ANITI de rattachement). Je comprends que les informations recueillies sont strictement confidentielles et à usage exclusif de l'UFTMiP et de l'Agence Nationale de la Recherche (ANR). J'accepte que l'équipe projet qui collabore à cette enquête ait accès à l'information dans le respect le plus strict de la confidentialité.

J'accepte que les données enregistrées à l'occasion de cette enquête, puissent faire l'objet d'un traitement informatisé sous la responsabilité du délégué à la protection des données de l'UFTMiP.

J'ai été informé que mes données seront pseudonymisées et que mon identité n'apparaîtra dans aucun rapport et que toute information me concernant sera traitée de façon confidentielle. J'accepte que les données enregistrées à l'occasion de cette enquête puissent être conservées dans une base de données et faire l'objet d'un traitement informatisé non nominatif par l'UFTMiP. Les données de l'enquête seront stockées et conservées pendant une durée totale de 3 ans avant d'être détruites.

J'ai bien noté que, conformément aux dispositions de la loi relative à l'informatique, aux fichiers et aux libertés et au règlement UE 2016/679 du Parlement Européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (RGPD), je dispose des droits suivants :

- un droit d'accès à mes données,
- un droit à l'information sur les données personnelles me concernant collectées et traitées,
- un droit à la rectification pour corriger des données personnelles incorrectes me concernant,
- un droit d'effacement des données personnelles me concernant uniquement,
- un droit à la limitation du traitement : dans ce cas, mes données pourront uniquement être stockées mais ne seront pas utilisées pour l'enquête, sauf consentement exprès,
- ainsi qu'un droit d'opposition au traitement de mes données personnelles et à la transmission des données couvertes par le secret professionnel susceptibles d'être utilisées dans le cadre de cette enquête.

Ces droits s'exercent auprès du délégué à la protection des données de l'UFTMiP, dpd@univ-toulouse.fr.

Si je le souhaite et à ma demande, les résultats de l'enquête me seront communiqués.

Mon consentement ne décharge en rien les organisateurs de l'enquête de leurs responsabilités, je conserve tous les droits garantis par la loi.

Fait à ....., le xx/xx/xxxx,

En deux (2) exemplaires originaux
Un exemplaire conservé par le signataire
Un exemplaire conservé par l'UFTMiP

Le signataire

Madame/Monsieur Prénom, NOM

Signature
Suivie de la mention « lu et approuvé »

Annexe 5 : Tableau DMP H2020

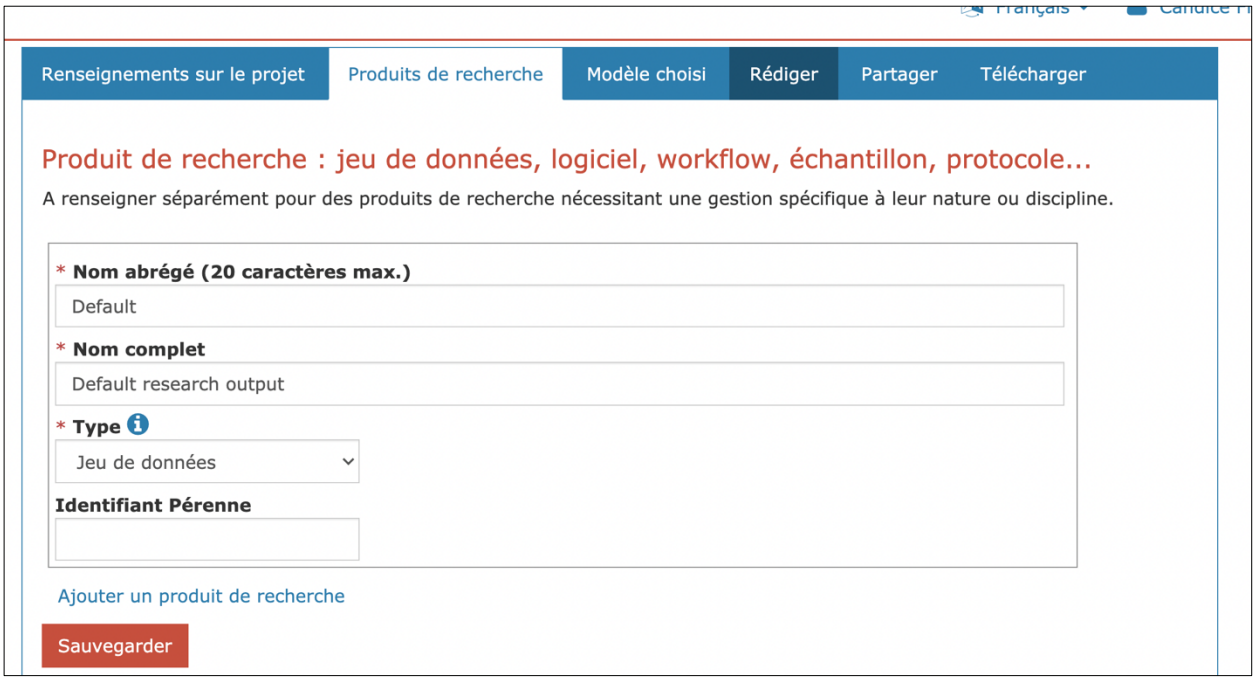
Nom du DMP	Version	Sujet	Nombre de pages du DMP	Lien
DTOceanPlus project	Première version	Outils de conception avancés pour les systèmes d'énergie océanique Innovation, développement et déploiement	22	https://services.phaidra.univie.ac.at/api/object/o:1139997/diss/Content/get
	Version finale		24	https://phaidra.univie.ac.at/detail/o:1139995#?page=1&pagesize=10&collection=o:1140797
ZAero Data Management Plan	Version 1	Fabrication de pièces composites sans défaut dans l'industrie aérospatiale	11	https://phaidra.univie.ac.at/detail/o:1139754#?page=2&pagesize=10&collection=o:1140797
CATANA Data management plan	Première version	Aéroélasticité et aéroacoustique composites	14	https://phaidra.univie.ac.at/detail/o:1140078#?page=3&pagesize=10&collection=o:1140797
ENLIVEN Data management plan	Version 1	Encourager l'apprentissage tout au long de la vie pour une Europe inclusive et vivante	16	https://services.phaidra.univie.ac.at/api/object/o:1139743/diss/Content/get
ROSSINI data management plan	Version 1	Intelligence et actionnement pour améliorer la qualité du travail dans la fabrication	31	https://phaidra.univie.ac.at/detail/o:1139306#?q=Artificial%20Intelligence&page=3&pagesize=10&collection=o:1140797
PoliVisu Data Management Plan	Version 1.0	Développement de politiques basées sur l'analyse et la visualisation de données géospatiales avancées	36	https://phaidra.univie.ac.at/detail/o:1139879#?q=Artificial%20Intelligence&page=3&pagesize=10&collection=o:1140797
FREME Data management plan	Version 1	Cadre ouvert de services électroniques pour l'enrichissement multilingue et sémantique du contenu numérique	27	https://phaidra.univie.ac.at/detail/o:1139603#?q=freme&page=1&pagesize=10&collection=o:1140797
	Version 2		22	https://services.phaidra.univie.ac.at/api/object/o:1139602/diss/Content/get

Nom du DMP	Version	Sujet	Nombre de pages du DMP	Lien
MELOA Data management plan	Version 1	Appareil d'océanographie extra léger polyvalent/multi-capteur	13	https://services.phaidra.univie.ac.at/api/object/o:1139227/diss/Content/get
OPERA Data management plan	Première version	Une expérience d'exploitation en pleine mer pour réduire le coût de l'énergie houlomotrice	30	https://services.phaidra.univie.ac.at/api/object/o:1139216/diss/Content/get
	Version finale		29	https://phaidra.univie.ac.at/detail/o:1139215#?q=%22OPERA%20Data%20management%20plan%20(final%20version)%22&page=1&pagesize=10&collection=o:1140797
Open4Citizens Data management Plan	Version initiale (ébauche)	Donner aux citoyens les moyens d'utiliser utilement les données ouvertes	15	https://phaidra.univie.ac.at/detail/o:1139705#?q=%22Open4Citizens%20Data%20Management%20Plan%20(Draft)%22&page=1&pagesize=10&collection=o:1140797
	Version mi-parcours		20	https://phaidra.univie.ac.at/detail/o:1139704#?q=%22Open4Citizens%20Data%20Management%20Plan%20(Midterm)%22&page=1&pagesize=10&collection=o:1140797
	Version finale		35	https://phaidra.univie.ac.at/detail/o:1139446#?q=%22Open4Citizens%20Data%20Management%20Plan%20(Final)%22&page=1&pagesize=10&collection=o:1140797

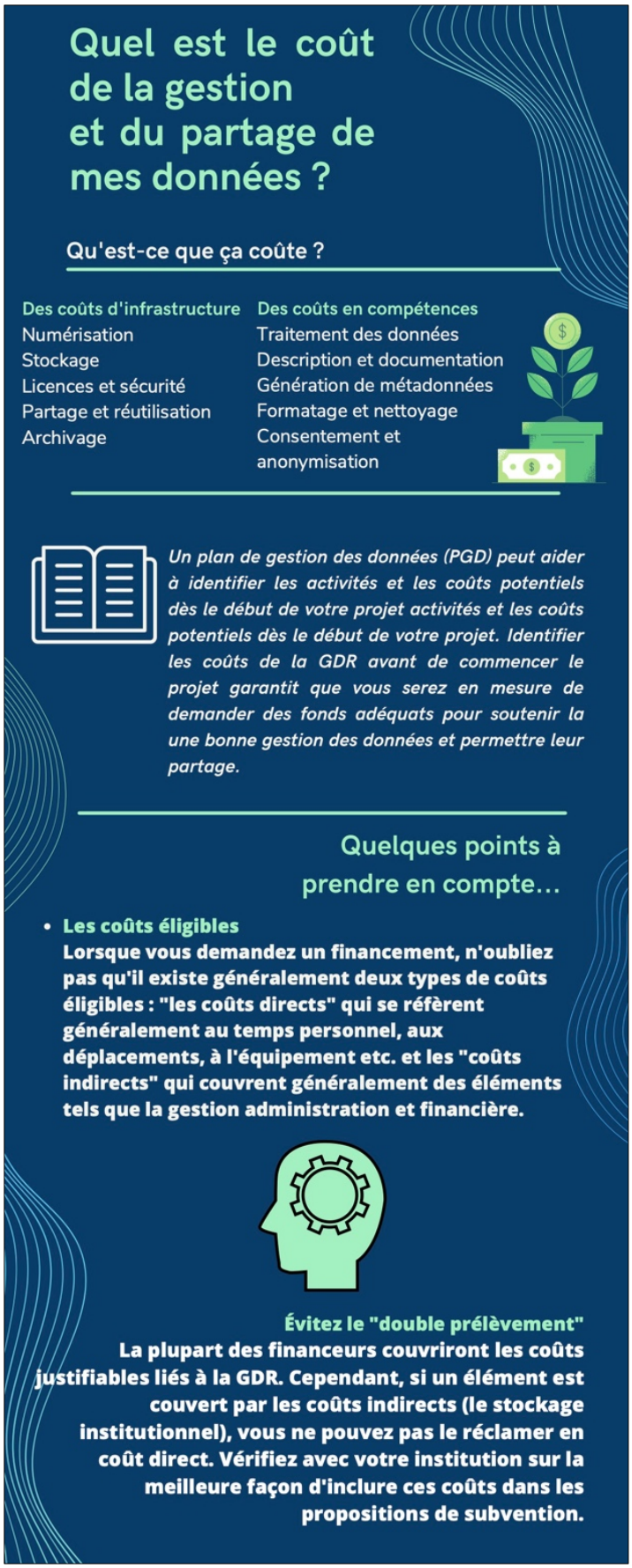
Nom du DMP	Version	Sujet	Nombre de pages du DMP	Lien
DARWIN Data management plan	Version initiale (ébauche)	S'attendre à l'inattendu et savoir comment réagir	22	https://phaidra.univie.ac.at/detail/o:1139683#?q=DARWIN&page=1&pagesize=10&collection=o:1140797
	Version finale		26	https://phaidra.univie.ac.at/detail/o:1139682#?q=DARWIN&page=1&pagesize=10&collection=o:1140797
RADICLE Data management plan	Version initiale (ébauche)	Système de contrôle dynamique en temps réel pour le soudage au laser	6	https://phaidra.univie.ac.at/detail/o:1139407#?q=RADICLE&page=1&pagesize=10&collection=o:1140797
	Version mi-parcours		10	https://phaidra.univie.ac.at/detail/o:1139406#?q=RADICLE&page=1&pagesize=10&collection=o:1140797
	Version finale		18	https://phaidra.univie.ac.at/detail/o:1139404#?q=RADICLE&page=1&pagesize=10&collection=o:1140797
PLANMAP Data management plan	Première version	Cartographie planétaire	80	https://phaidra.univie.ac.at/detail/o:1139189#?q=%22PLANMAP%20Data%20Management%20Plan%20%E2%80%93%201st%20update%22&page=1&pagesize=10&collection=o:1140797

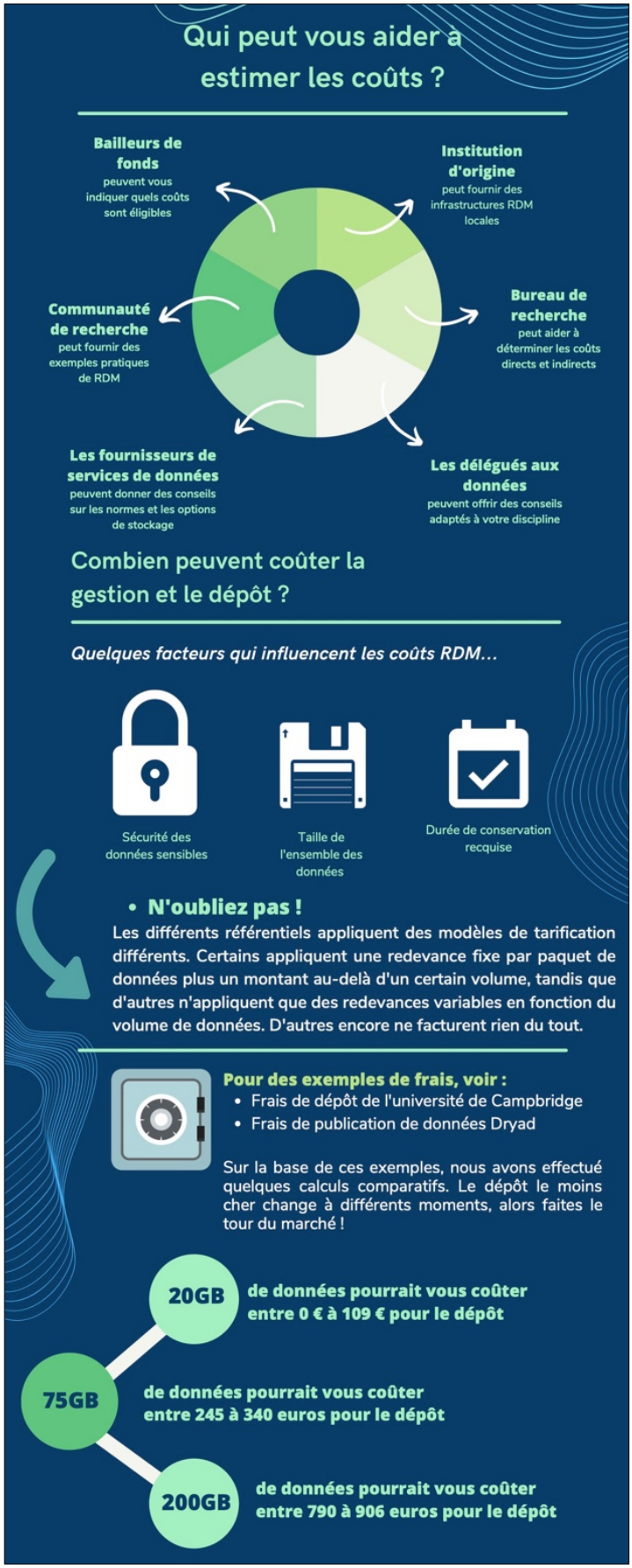
Nom du DMP	Version	Sujet	Nombre de pages du DMP	Lien
RINGO Data management plan	Version initiale (ébauche)	Préparation de l'ICOS aux besoins d'observations globales intégrées	6	https://phaidra.univie.ac.at/detail/o:1139159#?q=%22RINGO%20Initial%20Data%20Management%20Plan%20(DMP).%22&page=1&pagesize=10&collection=o:1140797
	Version mi parcours		9	https://phaidra.univie.ac.at/detail/o:1139157#?q=%22RINGO%20First%20Updated%20Data%20Management%20Plan%22&page=1&pagesize=10&collection=o:1140797
	Version finale		10	https://phaidra.univie.ac.at/detail/o:1139783#?q=%22RINGO%20Final%20Data%20Management%20Plan%22&page=1&pagesize=10&collection=o:1140797
SOLUS Data management plan	Première version	Diagnostics optiques et ultrasonores intelligents du cancer du sein	8	https://services.phaidra.univie.ac.at/api/object/o:1139394/diss/Content/get
	Version mise à jour		10	https://services.phaidra.univie.ac.at/api/object/o:1139797/diss/Content/get
DESTINATIONS Project Data management plan	Version initiale	DÉFIS SOCIÉTAUX - Des transports intelligents, verts et intégrés	35	https://services.phaidra.univie.ac.at/api/object/o:1139716/diss/Content/get
	Version mise à jour (V2)		20	https://services.phaidra.univie.ac.at/api/object/o:1139325/diss/Content/get
	Version mise à jour (V3)		23	https://services.phaidra.univie.ac.at/api/object/o:1139205/diss/Content/get
SUPREMA Data management plan	Version 1	Soutien à la modélisation politiquement pertinente de l'agriculture	29	https://services.phaidra.univie.ac.at/api/object/o:1139907/diss/Content/get

Annexe 6 : Capture d'écran DMP OPIDoR



Annexe 7 : Infographie coûts données de la recherche




Qui peut vous aider à estimer les coûts ?
Bailleurs de fonds
peuvent vous indiquer quels coûts sont éligibles
Institution d'origine
peut fournir des infrastructures RDM locales
Communauté de recherche
peut fournir des exemples pratiques de RDM
Bureau de recherche
peut aider à déterminer les coûts directs et indirects
Les fournisseurs de services de données
peuvent donner des conseils sur les normes et les options de stockage
Les délégués aux données
peuvent offrir des conseils adaptés à votre discipline
Combien peuvent coûter la gestion et le dépôt ?
Quelques facteurs qui influencent les coûts RDM...
Sécurité des données sensibles
Taille de l'ensemble des données
Durée de conservation recquise
N'oubliez pas !
Les différents référentiels appliquent des modèles de tarification différents. Certains appliquent une redevance fixe par paquet de données plus un montant au-delà d'un certain volume, tandis que d'autres n'appliquent que des redevances variables en fonction du volume de données. D'autres encore ne facturent rien du tout.
Pour des exemples de frais, voir :
Frais de dépôt de l'université de Campbridge
Frais de publication de données Dryad
Sur la base de ces exemples, nous avons effectué quelques calculs comparatifs. Le dépôt le moins cher change à différents moments, alors faites le tour du marché !
20GB
de données pourrait vous coûter entre 0 € à 109 € pour le dépôt
75GB
de données pourrait vous coûter entre 245 à 340 euros pour le dépôt
200GB
de données pourrait vous coûter entre 790 à 906 euros pour le dépôt

Annexe 8 : Modèle tableau de recensement des besoins des chercheurs

Nom chercheur	Chaire	Typologie du besoin (licence, contrat...)	Remarques (besoin exprimé)	Proposition de réponse

Annexe 9 : Étude de cas, communauté IA

Étude de cas
Gestion des données en Intelligence Artificielle

1) Étude de cas

Contexte : J'intègre un projet de recherche en intelligence artificielle en laboratoire pluridisciplinaire dans lequel je vais être amené à utiliser/produire des données afin de pouvoir entraîner mes algorithmes. Je compte publier mes résultats plus tard et pour cela je vais accompagner ma publication de mes données. Le but est d'organiser mes données afin de répondre aux principes FAIR et de les mettre à disposition de manière pérenne.

Organisation de mes données :

- Format : Quel serait le meilleur format adapté ?

Un format open source sera privilégié : .csv / .xml afin d'assurer la compatibilité avec des langages comme Python, Matlab, R etc…

- Stockage : Où stocker mes données lors de mon projet de recherche ? (Disque dur, clé USB, serveurs locaux, serveur labo) ?

Privilégier l'utilisation d'un GitLab sécurisé, soit du laboratoire affilié, soit du projet de recherche s'il en existe un. Si les données sont volumineuses prévoir aussi le coût que cela peut entraîner.

- Où stocker des données sensibles ?

Prévoir un lieu de stockage avec contrôle et accès limité à équipe de recherche. Si besoin, utiliser un processus de cryptage.

- Question sur les coûts de stockage → les données que je vais générer vont devoir être stockées en grande partie, comment prévoir le coût ?

Se renseigner avec les institutions : partenariats ou trouver une solution avec les affiliations.

- Question : Dois-formuler une convention de nommage ?

S'appuyer sur la convention de nommage existante ou celle appliquée au sein du laboratoire
Toujours suivre le même modèle sur les documents au cours du projet de recherche.

- Licence : Quelle licence appliquée sur mes données ? (Licence ouverte, licence à des fins non commerciales etc…)

Licence CC-BY-SA 4.0 Share A Like : https://creativecommons.org/licenses/by-sa/4.0/
La licence Share A Like permet de reproduire et republier les données sans restrictions, ainsi que de les modifier même à des fins commerciales à condition d'utiliser les données sous la même licence que l'originale.

Privilégier une licence CC-BY-NC 4.0. Cette licence permet à d'autres d'utiliser, d'adapter et de s'appuyer sur votre travail à des fins non commerciales. Bien que leurs nouvelles œuvres doivent également vous reconnaître et être non commerciales, ils ne sont pas tenus d'accorder une licence à leurs œuvres dérivées selon les mêmes conditions.
https://creativecommons.org/licenses/?lang=fr-FR
https://creativecommons.org/licenses/by-nc/4.0/

Pour un jeu de données que l'on souhaite ouvrir sans restriction, utiliser une licence ouverte telle que : etalab https://www.etalab.gouv.fr/licence-ouverte-open-licence
La « Licence Ouverte / Open License » présente les caractéristiques suivantes :

1. Une grande liberté de réutilisation des informations :
    - Une licence ouverte, libre et gratuite, qui apporte la sécurité juridique nécessaire aux producteurs et aux réutilisateurs des données publiques ;
    - Une licence qui promeut la réutilisation la plus large en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données ;
    - Une licence qui s'inscrit dans un contexte international en étant compatible avec les standards des licences Open Data développées à l'étranger et notamment celles du gouvernement britannique (Open Government Licence) ainsi que les autres standards internationaux (ODC-BY, CC-BY 2.0).
2. Une exigence forte de transparence de la donnée et de qualité des sources en rendant obligatoire la mention de la paternité.
3. Une opportunité de mutualisation pour les autres données publiques en mettant en place un standard réutilisable par les collectivités territoriales qui souhaiteraient se lancer dans l'ouverture des données publiques.

Licences GNU applicable pour les codes ou logiciels, bases de données
Comment utiliser une licence GNU pour vos logiciels ?

Les principes de la licence GNU-GPLv3 :
« Personne ne doit être limité par les logiciels qu'il utilise. Il y a quatre libertés que tout utilisateur doit posséder :

- la liberté d'utiliser le logiciel à n'importe quelle fin,
- la liberté de modifier le programme pour répondre à ses besoins,
- la liberté de redistribuer des copies à ses amis et voisins,
- la liberté de partager avec d'autres les modifications qu'il a faites »

Il existe d'autres versions de licence GNU comme la GNU-Affero.

- Je trouve un jeu de données en ligne, comment savoir s'il est réutilisable ?

Regarder à quelle licence est soumise mon jeu de données pour savoir si l'utilisation est possible ou s'il y a des restrictions de partage. (cf. question sur les différents types de licences)

- Anonymisation (si données sensibles) : Comment anonymiser mes données ?

Possibilité d'utiliser une table de correspondance avec une clé.
Outil aide à l'anonymisation des données : Amnesia, anonymisation des données en 5 étapes, avec possibilité d'enregistrer les données localement sur Zenodo.
Page de documentation

- Entrepôt : Sur quel entrepôt déposer mes données de manière pérenne, assurant l'attribution d'un DOI ?

Privilégier un entrepôt disciplinaire plutôt qu'institutionnel le dépôt des données en code source et proposer un choix de l'entrepôt adapté : Open Science Framework, OpenNeuro, UCI Machine Learning Repository, avec attribution d'DOI pour une indentification pérenne.

Sinon utiliser un annuaire d'entrepôt comme re3data, possibilité de rechercher par type de contenu, discipline, pays.

Lien DORANum : « critères pour choisir son entrepôt. »
Lien CoopIST : « comment choisir un entrepôt de données ? »

- Métadonnées : Standard des métadonnées le plus utilisé en IA ?

L'IA est une communauté pluridisciplinaire, une norme généralisée n'est pas fixée. Différentes normes sont possibles.

Digital Curation Center :
« *Pour les disciplines qui ne se sont pas encore fixées sur une norme de métadonnées et pour les référentiels qui fonctionnent avec des données interdisciplinaires, la section Données générales de recherche renvoie à des informations sur des normes de métadonnées plus larges qui ont été adaptées pour répondre aux besoins des données de recherche*. »

Liens de documentation normes et outils associés :
https://www.dcc.ac.uk/resources/subject-areas/general-research-data
https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html

- *Data paper* : Je souhaite valoriser mes données produites en tant que chercheur, je voudrais savoir ce qu'est un data paper et comment procéder ?

Béchec, M. L. (s. d.). Le data paper, un nouvel outil de communication scientifique? [Billet]. *Territoires numériques de marques*. Consulté 26 août 2021, à l'adresse https://ternumeric.hypotheses.org/374

Data journal en IA :
SpringerOpen : Applied informatics
https://applied-informatics-j.springeropen.com/articles

« Applied Informatics couvre la théorie et l'application de l'informatique dans divers domaines scientifiques, technologiques, techniques et sociaux. Dans le but d'inspirer de nouvelles recherches multidisciplinaires, le journal agit comme un lieu d'intégration qui recueille des articles de recherche originaux de haute qualité et des revues sur divers aspects de l'informatique appliquée, avec les fondements de l'informatique (théorie de l'information, modélisation statistique, apprentissage automatique, etc.) comme noyau moteur et les interactions entre les domaines essentiels comme centres de promotion ; Les interactions entre (a) les sciences de la vie (bioinformatique, informatique médicale, bio-ingénierie, etc.),

(b) les sciences de l'intelligence (informatique neuronale et cognitive, multimédia, reconnaissance des formes, etc.) et (c) les sciences communautaires (réseaux sociaux, informatique affective, analyse de données massives, etc.) »
https://ternumeric.hypotheses.org/374

Deux liens utiles à la publication d'un data paper :
Doranum « coomment le faire, le publier ? » :
CoopIST : « comment structurer un data paper ? »