Université de Paris

Guide d'aide à la rédaction d'un plan de gestion des données à Université de Paris

Modèle ANR

Ce guide a été conjointement réalisé par la Direction générale déléguée des bibliothèques et musées et le Département des archives de Université de Paris

V3 - Avril 2021

VOS INTERLOCUTEURS

Elise Lehoux est conservatrice de bibliothèques et référente données de la recherche pour Université de Paris

elise.lehoux@u-paris.fr

Benjamin Rullier est le responsable du département des archives d'Université de Paris

benjamin.rullier@u-paris.fr

Pour toute question sur les plans de gestion de données et les données de la recherche : donnees.recherche.dbm@listes.u-paris.fr

Introduction

Dans le contexte d'une science de plus en plus ouverte, les financeurs (ANR, Horizon 2020 puis Horizon Europe, ERC en 2021) demandent aux porteurs de projet de réaliser un **Plan de Gestion de Données*** (PGD) ou **Data Management Plan** (DMP) qui explicite la façon dont les données du projet seront produites, stockées, gérées, archivées, et éventuellement partagées, au cours et à la fin du projet.

Le PGD/DMP permet de se poser un certain nombre de questions de gestion en amont du projet. Une première version doit être rendue à l'ANR dans les 6 mois qui suivent le début du projet et la version finalisée à la fin de ce dernier. Une version intermédiaire est demandée à mi-parcours pour les projets de plus de 30 mois.

Se former sur la gestion des données, les PGD/DMP et la Science ouverte :

Pour s'auto-former sur les données de la recherche :

- Un parcours interactif a été réalisé par DoRANum : https://doranum.fr/enjeux-benefices/parcours-interactif-sur-la-gestion-des-donnees-de-la-recherche/
- Un guide d'autoformation est également disponible : https://zenodo.org/record/3920869
- Un guide réalisé par le CNRS sur les bonnes pratiques de gestion : https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html
- Sur les principes FAIR : https://www.force11.org/group/fairgroup/fairprinciples

Pour lire des exemples de PGD/DMP :

- Le site DMP OPIDoR en rassemble un certain nombre : https://dmp.opidor.fr/public_plans comme son équivalent anglo-saxon DMP Online : https://dmponline.dcc.ac.uk/public_plans

Pour mieux appréhender la place de la Science ouverte dans les projets :

- Un guide pour améliorer son projet ANR grâce à la Science ouverte dans les projets de recherche : https://zenodo.org/record/3769954

Des **formations** sont également proposées aux porteurs de projet plusieurs fois dans l'année, en collaboration entre la Direction générale déléguée des bibliothèques et musées et le Département des archives. **Pour toute question** : donnees.recherche.dbm@listes.u-paris.fr.

Le présent guide est destiné à faciliter la réalisation d'un PGD/DMP.
Les recommandations sont déclinées par facultés, avec des exemples pour chaque section seront et sont régulièrement révisées. Il se présente comme un complément aux recommandations Université de Paris disponibles sur le site DMP OPIDoR.

N.B. : dans le document, nous utilisons les sigles PGD et DMP pour qualifier les Plans de gestion de données/data management plan de manière indistincte.
Les mots suivis d'une (*) sont à retrouver dans le lexique.

RÉDIGER UN PGD

Etape 1 – Choisir un modèle de PGD

Pour créer un PGD, plusieurs options de « modèles » sont disponibles : le modèle de votre institution, d'un autre organisme ou du financeur.

Nous vous recommandons le **modèle ANR** sur lequel se fonde ce guide : https://anr.fr/fileadmin/documents/2019/ANR-modele-PGD.pdf.

Le modèle est accessible en anglais ou en français sur le site DMP OPIDoR : https://dmp.opidor.fr/

NB : le choix du modèle est totalement libre.

Etape 2 – Renseigner les informations sur le projet

À cette étape, il est possible d'ajouter des recommandations qui vous aideront tout au long de la saisie du PGD.

Informations générales

Renseigner les mêmes informations que celles fournies dans la réponse à l'AAP ANR/Convention de financement.

Contact pour les données

Ce peut-être le même que le chercheur responsable ou un autre membre de l'équipe qui aura en charge l'actualisation du PGD au fur et à mesure de l'avancement du projet.

Le PGD est un document évolutif, qui s'affine au fur et à mesure où le projet progresse.

Etape 3 – Indiquer les produits de recherche (optionnel)

Les **produits de recherche*** désignent des jeux de données qui vont être gérées, stockées, archivées ou partagées différemment au cours et à la fin du projet de recherche.

En effet, un ou plusieurs jeu(x) de données peu(ven)t être lié(s) au projet de recherche, et désigner :

a) un lot techniquement homogène,
b) un lot intellectuellement cohérent même si celui-ci est composé de lots techniquement hétérogènes.

L'onglet 'Produits de recherche' est à renseigner pour les produits de recherche nécessitant une gestion spécifique en fonction de leur nature ou discipline.
Définir différents produits de recherche permet de subdiviser le PGD en plusieurs sections indépendantes et facilite sa rédaction au fur et à mesure du projet.
Le nom abrégé permet de nommer ces différentes sections au sein du PGD.
L'identifiant pérenne (ex : DOI) pourra être renseigné en fin de projet.

Les produits de recherche peuvent être sélectionnés parmi la liste suivante :

audiovisuel, collection, jeu de données, image, ressource interactive, modèle, objet physique, service, logiciel, son, texte, workflow, autre.

Vous pouvez en ajouter autant que votre projet le nécessite.

Attention toutefois : le premier produit de recherche renseigné ne peut être déplacé ou supprimé.

Exemples de différents produits de recherche :

Exemple SH

1. Une base de données avec des données d'enquête anonymisées (*dataset*)
2. Un ensemble de vidéos (*audiovisual*) dont le traitement et la gestion sera spécifique tout au long du projet.

Exemple STM

1. Quack : QuAcK (Logiciel)
2. QP : Quantum Package (Logiciel)
3. Notebooks : Notebooks (Ressource interactive)
4. Publications : Publications (Texte)

Source : https://dmp.opidor.fr/plans/6397/export.pdf

Etape 4 – Rédiger

Six sections composent le PGD, qui se subdivisent en fonction du nombre de produits de recherche renseigné (ou non).

1. Description des données et collecte ou réutilisation de données existantes

a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Dans cette partie, il s'agit d'expliquer les données utilisées au cours du projet : celles qui seront **produites** au cours de celui-ci et/ou les données **réutilisées** dans le projet. Dans les deux cas, il convient de **justifier** la nécessité d'en produire de nouvelles, d'expliquer leurs modalités de production ou de réutilisation (méthodologie, logiciels, tests sur des échantillons…).
Pour la propriété des données réutilisées : se référer aux licences qui existent ou aux contrats passés pour connaître les droits de réutilisation.

Quelques exemples d'outils pour réutiliser des jeux de données existants :

- Depuis des portails comme Open AIRE Explore https://explore.openaire.eu/ ; Isidore (SHS) : https://isidore.science/ ; Datacite.org
- Depuis des entrepôts directement : consulter un wiki sur les entrepôts disciplinaires : http://oad.simmons.edu/oadwiki/Disciplinary_repositories ou un répertoire international d'entrepôt comme re3data.org
- Depuis les bases d'Open Data : le portail data.gouv.fr (pour les données de l'État et des collectivités).
- Pour les données de santé : https://www.health-data-hub.fr/

Exemple SH

La production des données textuelles est réalisée par les participants au projet. Elle pourra comprendre la mise en place de datavisualisations, ainsi que de contributions diverses recueillies et centralisées sur notre CMS (Omeka S). Le programme de recherche s'appuie sur des corpus d'archives dispersées (en ligne, au Mundaneum, présentes chez les partenaires, issues de numérisations, et/ou traitements automatisés, extractions manuelles de contenus à partir de textes OCRisés) afin de produire, à l'aide de méthodologies issues des humanités numériques (fouille de texte en python, web sémantique, indexation des entités nommées) des synthèses et des publications scientifiques. Parmi les outils utilisés figurent (…)
Source : https://dmp.opidor.fr/plans/4205/export.pdf

Exemple STM

Les données qui seront produites dans le cadre du projet IMPRINT sont de quatre types :
(i) des données microclimatiques recueillies in-situ sur les différents sites d'étude envisagés dans le projet ;
(ii) des données d'observations du milieu réalisées in-situ (p.ex. données de localisation géographique, relevés dendrométriques et botaniques, etc.) ;
(iii) des données issues de prélèvements réalisés in-situ puis analysés en laboratoire (p.ex. pièges Barber pour inventorier les communautés d'arthropodes du sol, échantillons de sols, etc.) ;
(iv) des données issues de la télédétection (imagerie LiDAR). (…)
Source: https://dmp.opidor.fr/plans/5082/export.pdf

Etape 4 – Rédiger

1. Description des données et collecte ou réutilisation de données existantes

b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données utilisées devront être documentées : **type de données**, **format** (en justifiant si un choix se présente – si le format est libre ou propriétaire par exemple), **volume de données produit** (approximatif en début de projet). Distinguer si possible le format pour la diffusion et celui pour la conservation.

Ressources externes :
- Pour décrire les données selon leur nature, voir sur DoRANum : https://doranum.fr/plan-gestion-donnees-dmp/origine-description-donnees-recherche/
- Pour les formats, voir le logiciel FACILE du CINES : https://facile.cines.fr/

Exemple SH

Les données produites seront les suivantes :
- Documents textuels des textes antiques non lemmatisés et lemmatisés en langue grec (alphabet grec ancien, format Unicode) en .txt regroupés en dossiers par auteurs, c'est-à-dire 6 dossiers de maximum 30 Mo, 30 Mo étant le volume de fichiers pour Libanios. À l'heure actuelle, nous disposons de 119 fichiers .txt (16+103) des lemmatisations de Denys d'Halicarnasse (16 doc .txt) et de Libanios (103 doc .txt).
- Documents textuels des passage bibliques liées à l'hospitalité en grec, en latin et en anglais en format .odt car libre, soit 73 documents car il y a 73 livres dans la Bible, fichier individuel d'environ 500 Ko
- Document xml des résultats Biblindex pour chaque passage étudié avec les références chez les auteurs antiques, soit entre 3000 et 4000 fichiers car la Bible compte 73 livres, pour chaque passage, nous pensons qu'il peut y avoir 50 passages (pour le moment environ 30 pour Genèse et pour Matthieu), 73x50=3650 fichiers.
- Bibliographie Zotero, pour gérer les références bibliographiques sur le sujet, maintenue par le porteur de projet
- Tableur en .csv contenant la liste des auteurs.

Source : https://dmp.opidor.fr/plans/5278/export.pdf

Exemple STM

Les données qui seront acquises au cours du projet sont pratiquement toutes de type numérique (tableurs de données et géodatabases). Quelques images (p.ex. photos hémisphériques ou photos traditionnelles) ou vidéos seront également acquises au cours du projet afin d'illustrer un site web ou un blog dédié au projet. En ce qui concerne les données textuelles, celles-ci concernent uniquement les protocoles d'acquisition des données ou bien les fichiers de métadonnées. Les données microclimatiques de T°C et d'humidité relatives seront stockées au format électronique uniquement (.txt, .csv, .xlsx). Les fichiers bruts issus des capteurs HOBO et TMS4 peuvent être enregistrés dans plusieurs formats puis traités par n'importe quel logiciel de traitement de données. Le volume des données microclimatiques qui seront générées pendant la durée du projet est estimé à plusieurs Go. Les données d'observations du milieu comme les données GPS ou bien les données issues des inventaires dendrométriques et floristiques seront enregistrées au format tableur électronique (.txt, .csv, .xlsx). De même, les données d'identification taxonomiques ou bien les données d'analyses de sol (pH, C/N, etc.), acquises en laboratoire, seront enregistrées au format tableur électronique (.txt, .csv, xlsx). Le volume de ces données est estimé à plusieurs Mo. Les données LiDAR seront stockées au format dédié (.las) à ce type de données 3D ou "point cloud" pour nuage de points 3D, avec un volume de données pouvant dépasser plusieurs To. Dans la mesure du possible, des formats standards (.txt) et ouverts seront privilégiés à des fins de partage et de réutilisation. Un projet de géodatabase ouverte à l'aide du logiciel ArcGIS ou de RShiny sera envisagé en fin de projet afin de mettre en libre accès ou en consultation libre l'ensemble des données acquises dans le cadre du projet IMPRINT.

Source : https://dmp.opidor.fr/plans/5082/export.pdf

Etape 4 – Rédiger

2. Documentation et qualité des données

a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Cette partie vise à documenter et justifier les métadonnées utilisées.
Les **métadonnées***, « données sur les données », sont des informations descriptives qui permettent de renseigner le contenu d'un jeu de données (ex : titre, date de création, format, etc.).
Il est préférable d'**utiliser des standards disciplinaires usuels** – afin que les métadonnées soient plus facilement repérables par les communautés et conformes aux **principes FAIR** (Findable, Accessible, Interoperable, Reusable).

Le standard dépend de la destination des données : dépôt, archivage, publication, etc.
Le plus connu est le Dublin Core, il peut être adapté en fonction des besoins. Veiller dans la mesure du possible à utiliser des formats et des standards ouverts, non propriétaires, adaptés à votre discipline et/ou à vos données.
Si un vocabulaire contrôlé ou une ontologie sont utilisés, ils peuvent être cités dans cette partie. Voir Loterre pour les thesaurus.
Plus la description est précise, plus les données pourront être visibles, citées et/ou réutilisées le cas échéant.

Principales métadonnées par grands champs disciplinaires

Général : CERIF, Data Package, DataCite Metadata Schema, DCAT, Dublin Core, OAI-ORE, Observations and Measurements, PREMIS, PROV, RDF Data Cube Vocabulary, Repository-Developed Metadata Schemas.

Sciences sociales & humanités : DDI, EAD, MIDAS-Heritage, OAI-ORE, QuDEx, SDMX

Sciences physiques : AVM, CIF, CSMD-CCLRC, FITS, International Virtual Observatory Alliance Technical Specifications, NeXus, Observations and Measurements, PDBx/mmCIF, SDAC, SPASE Data Model.

Sciences de la terre : AgMES, AVM, CF, CIM, DIF, FGDC/CSDGM, ISO 19115, Observations and Measurements, Repository-developed Metadata Schemas.

Biologie : Darwin Core, EML, Genome Metadata, ISA-Tab, MIBBI, Observ-OM, OME-XML, PDBx/mmCIF, Protocol Data Element definitions, Repository-Developed Metadata Schemas.

Source : https://www.dcc.ac.uk/guidance/standards/metadata

Ressources externes

Catalogue de standards de la RDA https://rdamsc.bath.ac.uk/

Digital curation standards : http://www.dcc.ac.uk/resources/metadata-standards

FAIRsharing : https://fairsharing.org/standards/exchange_format
et https://guides.lib.unc.edu/metadata/standards

Un générateur de métadonnées pour les données de recherche :
https://doranum.fr/wp-content/uploads/datacite_metadata_generator_4.0.html

Etape 4 – Rédiger

2. Documentation et qualité des données

a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Dans cette partie, il est aussi intéressant de préciser les règles de nommage des fichiers et de créer un fichier « ReadMe » précisant l'organisation des fichiers de données et le versionning, facilitant leur réutilisation et/ou la conservation à plus long terme.

Déterminer les règles de gestion, de classement, conservation, d'accès et de partage des données au cours du projet : arborescence, nommage, métadonnées, gestion des versions.

Ressources internes :
Fiches pratiques du Département des archives : « nommer ces documents », « organiser une arborescence »

Ressources externes :
DoRANum : https://doranum.fr/stockage-archivage/comment-nommer-fichiers/

Exemple SH

Concernant les conventions de nommage, les noms des livres bibliques sont indiqués en anglais. Le numéro du livre biblique sera séparé du numéro de verset par un underscore et pour les passages, le début et la fin du passage seront séparé par un tiret. Par exemple, Matthew26_7-13 pour Matthieu livre 26 passages du verset 7 au verset 1. Les noms de fichiers ne comporteront pas de caractères spéciaux, ni d'espaces. Chaque nom de fichier se termine par l'indication de la date et de l'heure de la sauvegarde au format américain, soit par exemple 01_13_2020_12:24, pour le 13 janvier 2020 à 12h24. Pour les dossiers, le dossier des six auteurs grecs sera séparé du dossier du travail en lien avec la Bible.
Le dossier « 6_auteurs_grecs » sera composé de 6 sous-dossiers, correspond à chacun des 6 auteurs, à savoir : Denys d'Halicarnasse, Plutarque, Apollonios de Tyane, Basile de Césarée, Jean Chrysostome et Libanios, soit Dionysos_of_Halicarnassus, Plutarch, Apollonius_of_Tyana, Basil_of_Caesarea, Chrysostomus and Libanios.
Source : https://dmp.opidor.fr/plans/5278/export.pdf

Exemple STM

Chaque fichier de données sera accompagné au minimum d'un fichier texte (.txt) de type "Read_me.txt" indiquant les métadonnées telles que la liste des noms (header) des variables enregistrées, avec pour chaque variable (header), le type d'appareillage utilisé pour la prise de mesure, les conditions dans lesquelles les données ont été collectées, l'unité de la variable mesurée ou toutes autres informations permettant une réutilisation des données. La langue utilisée sera l'anglais pour une plus grande réutilisation potentielle des données collectées. Le standard de métadonnées envisagé dans le cadre du projet est le standard "Ecological Matadata Language" (EML) avec une implémentation possible sous le logiciel R grâce au package EML. La convention de nommage des fichiers ou dossiers de données sera celle recommandée sur le site DoRANum en utilisant si possible la date, au format "AAAAMMJJ", de la dernière version du fichier de données, sans utiliser d'accents ou de caractères spéciaux mais le symbole "_" ou "underscore" en tant que séparateur de mots dans le nom du fichier et enfin en affichant l'élément important en premier dans le nom (p.ex. 200218_DMP_IMPRINT.docx"). Dans le cas de documents pouvant être amenés à évoluer, nous préciserons la version dans le nom du fichier (p.ex."20200218_DMP_IMPRINT_v01.docx").
Source : https://dmp.opidor.fr/plans/5082/export.pdf

Etape 4 – Rédiger

2. Documentation et qualité des données

b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Le CINES propose un outil pour vérifier que vos formats de fichiers sont correctement générés pour être intégrés dans sa plateforme d'archivage. Cet outil peut être utile pour vérifier des formats de fichiers dans d'autres buts : https://facile.cines.fr/

Ressources externes :
Un ensemble de guides pratiques du réseau Qualité en recherche (CNRS) :
http://qualite-en-recherche.cnrs.fr/spip.php?rubrique41 et
http://qualite-en-recherche.cnrs.fr/spip.php?article327

Exemple SH

La qualité des données sera contrôlée grâce à des validations, décidées lors de réunion de validation bimensuelle, deux fois par mois. Ces réunions regrouperont le porteur de projet, le post-doctorant en humanités numériques et l'ingénieur d'études. Nous utiliserons des outils permettant des vérifier les liens de la base de données qui renvoient vers les environnements hypertextes, comme par exemple LinkChecker, qui est un logiciel libre. Cet outil permet de vérifier s'il n'y a pas de liens cassés dans les documents HTML. LinkChecker est un outil python en ligne de commande qui permet de parcourir un site en suivant les liens. Il fournit un résumé (nombre de warning, nombre d'erreurs) et il est configurable pour correspondre à nos besoins. Il sera utilisé par l'ingénieur d'études du projet. Dans le tableur de la base de données, il y aura une colonne pour indiquer le statut d'une donnée : « en cours / validé le DATE ». Les données seront soumises à une validation intellectuelle, une validation experte par le porteur du projet. Concernant la gestion du versionning, nous utiliserons un logiciel de gestion des versions : Git. Git est un logiciel libre.
Source : https://dmp.opidor.fr/plans/5278/export.pdf

Exemple STM

De manière générale, les bonnes pratiques de terrain et de laboratoire seront suivies pour le contrôle et la qualité des données, que ce soit lors de la collecte de données in-situ (en forêt) ou bien en laboratoire sur des échantillons prélevés in-situ, sur le terrain. Le protocole d'échantillonnage des conditions microclimatiques le long d'un gradient d'ouverture-fermeture de la canopée forestière prévoit suffisamment de réplicats. Au total, trois Forêts Domaniales à dominante feuillue seront échantillonnées (FD de l'Aigoual, FD de Blois et FD de Mormal). Au sein de chaque FD, 60 placettes (un sous-ensemble de placettes de calibration et un sous-ensemble de placettes de validation) seront (i) équipées de capteurs pour enregistrer les conditions microclimatiques sous-couvert forestier et (ii) inventoriées sur le plan de la biodiversité végétale et animale (cf. communautés d'arthropodes du sol). Nous prévoyons également d'équiper une partie du réseau RENECOFOR (les placettes de forêts feuillues notamment) avec un jeu apparié des ondes TMS4 : (i) une sonde sous-couvert forestier au sein de la placette de suivi du réseau RENECOFOR et (ii) une autre sonde placée hors-couvert forestier et au sein d'une station météo située à proximité de la forêt, lorsque ce type de station météo est disponible. Ce jeu de données microclimatiques recueillies sur l'ensemble du territoire national sera très complémentaire du précédent qui est focalisé sur trois forêts domaniales équipées d'une plus grande densité de capteurs microclimatiques. Concernant la qualité et la conformité de la collecte des données microclimatiques, il est prévu une phase d'intercalibration des capteurs HOBO UA-001-08, HOBO UA-001-64 et TMS4 en conditions contrôlées. L'installation des capteurs de T°C et d'humidité relative du sol, in-situ, suivra un protocole standardisé pour l'ensemble des sites étudiés. Concernant les données saisies sur le terrain (inventaires dendrométriques et floristiques) ou en laboratoire (détermination des espèces de carabes et analyses de sol éventuellement), elles seront validées par l'ensemble des scientifiques impliqués dans le projet.
Source : https://dmp.opidor.fr/plans/5082/export.pdf

3. Stockage et sauvegarde pendant le processus de recherche

a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche

Dans cette partie, ne pas hésiter à être extrêmement **concret** (ex. nombre de disques durs stockés dans un bureau fermé à clé par exemple) et expliciter les procédures mises en place pour le stockage des données au cours du projet (lieu de stockage, nombre et fréquence de sauvegardes, supports de sauvegarde utilisés, chiffrement des données, etc.).

Le choix des outils dépend des usages et des besoins (forte disponibilité des données, partage, confidentialité, accès en situation de mobilité…)

Ressources internes

NextCloud le cloud institutionnel d'Université de Paris (https://cloud.parisdescartes.fr/) et à terme, une solution de cloud (fondée sur NextCloud) pour les laboratoires.

MyCore (pour le CNRS) et iBox (pour l'Inserm) sont deux autres solutions de stockage dans le cloud proposées par des institutions de recherche

Si le projet comporte des données sensibles (données à caractère personnel, données de santé notamment), cette partie peut être complétée à l'aide des recommandations du Délégué à la Protection des Données (dpo@u-paris.fr), en lien avec les services proposés par les institutions partenaires du projet.

Sur les questions de stockage des données, la DSIN d'Université de Paris peut être sollicitée : assistance.dsin@u-paris.fr

Ressources externes

Pour les SHS, Huma-Num propose des solutions de stockage sécurisé. Le Centre des humanités numériques d'UP (recherche.dbm@listes.u-paris.fr) pourra vous accompagner vers les services Huma-Num.

Pour les données de santé :
Cadre légal relatif au traitement de données de santé à des fins de recherche : https://www.cnil.fr/fr/recherche-medicale-quel-est-le-cadre-legal (et https://www.cnil.fr/fr/tag/Recherche+m%C3%A9dicale)

Liste des hébergeurs agréés de données de santé : https://esante.gouv.fr/labels-certifications/hds/liste-des-herbergeurs-agrees

Entrepôts de données de santé : https://www.cnil.fr/fr/traitements-de-donnees-de-sante-comment-faire-la-distinction-entre-un-entrepot-et-une-recherche-et

Rappel de la CNIL sur les traitements de données de santé à partir de la plateforme Health Data Hub : https://www.cnil.fr/fr/la-plateforme-des-donnees-de-sante-health-data-hub

3. Stockage et sauvegarde pendant le processus de recherche

b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Dans cette partie, ne pas hésiter à être extrêmement concret et précis (chiffrement des données, contrôle des accès via mot de passe, transfert sécurisé des données lors de partage, gestion des droits d'accès sur l'arborescence, etc.).
Cette partie peut aussi être complétée à l'aide des recommandations du délégué à la protection des données dont dépend votre structure de recherche, notamment concernant la politique institutionnelle de protection des données.

Pour l'analyse des risques, voir la Méthode EBIOS sur le site de l'ANSSI (https://www.ssi.gouv.fr/administration/management-du-risque/)

Ressources de la CNIL sur la sécurité des données : https://www.cnil.fr/fr/cybersecurite

Sur l'anonymisation des données : https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

Ressources internes : Contact Délégué à la Protection des Données (dpo@u-paris.fr) et DSIN (assistance.dsin@u-paris.fr) notamment le responsable de la sécurité du système d'information ou RSSI)

Etape 4 – Rédiger

4. Exigences légales et éthiques, code de conduite

a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Cette partie peut aussi être complétée à l'aide des recommandations du Délégué à la Protection des Données (DPD) : se rapprocher du DPD désigné pour l'UMR (DPD de l'une des tutelles).

Précisez ici :
- Si le traitement de données à caractère personnel a été déclaré dans le registre des activités de traitement tenu par le DPD
- La base légale du traitement (consentement ou autre)
- Les modalités de recueil et de gestion du consentement au cours du projet de recherche (possibilité de se rétracter et de sortir du champ de l'étude)
- Les moyens pour informer les personnes concernées du traitement de leurs données
- Les modalités d'exercice des droits des personnes concernées
- Les mesures techniques et organisationnelles pour garantir la confidentialité et la sécurité des données, techniques de pseudonymisation, chiffrement des données, etc.

Dans cette partie, l'accord d'un comité d'éthique (ou autre comité) peut être mentionné.

La réponse à cette question doit être articulée avec la demande formulée auprès du Comité d'éthique. Voir le formulaire du Comité d'éthique de la recherche d'Université de Paris : http://evenementiels.aphp.fr/wp-content/blogs.dir/229/files/2020/04/Formulaire-de-soumission-au-CER-Paris-Descartes-formulaire-unique-7-janvier-2020.pdf

Sur l'anonymisation des données :
https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

Pour la mise en conformité au RGPD, voir le site de la CNIL :
https://www.cnil.fr/fr/principes-cles/rgpd-se-preparer-en-6-etapes

Ressources internes :
Contact Déléguée à la Protection des Données d'Université de Paris : dpo@u-paris.fr

4. Exigences légales et éthiques, code de conduite

b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Voir le guide d'analyse juridique sur l'ouverture des données de recherche (réalisé par l'INRAE) : https://www.ouvrirlascience.fr/ouverture-des-donnees-de-recherche-guide-danalyse-du-cadre-juridique-en-france-v2/

Les bases de données relèvent d'un droit spécifique : l'architecture de la base de données est protégée par le droit d'auteur, tandis que son contenu bénéficie d'une protection distincte et indépendante, appelé « droit sui generis ». Voir : https://www.app.asso.fr/centre-information/base-de-connaissances/code-bases-de-donnees/introduction-au-droit-des-bases-de-donnees/definition-de-la-base-de-donnees

Ressources externes :

- Logigramme de l'ENCP
  https://espacechercheurs.enpc.fr/sites/default/files/logigramme_a_plat.pdf
ou sous une forme dynamique :
https://espacechercheurs.enpc.fr/fr/logigramme_dynamique

- Pour les SHS, voir le guide réalisé par le CNRS (V2 – février 2021) :
https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf

- Le carnet de recherche « Ethique & droit » : https://ethiquedroit.hypotheses.org/

4. Exigences légales et éthiques, code de conduite

c. Comment les éventuelles questions éthiques seront-elles prises en compte, et les codes déontologiques respectés ?

Préciser si des questions d'éthique pourront se poser avec le projet de recherche (conflit d'intérêt, traitement de données « sensibles » du point de vue sociétal, etc.) et indiquer les éventuelles incidences de ces questions d'éthique sur la façon dont les données seront stockées et transférées, qui pourra les voir ou les utiliser et quelles durées de conservation leur seront appliquées.

Définir les mesures qui seront mises en œuvre pour répondre à ces questions éthiques (application de codes de conduite, ou codes d'éthique et validation du projet de recherche par un comité d'éthique).

Etape 4 – Rédiger

5. Partage des données et conservation à long terme

a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Cette partie est dédiée au **partage – s'il est possible – des données**. Le partage de jeux de données peut se faire par différentes voies :

- le dépôt des jeux de données dans un entrepôt
- l'indexation des métadonnées dans un catalogue
- la publication de *data papers* associée au(x) jeu(x) de données, etc. Le *data paper** est une publication qui décrit un jeu de données scientifiques, à l'aide d'informations structurées (métadonnées). Ce sont des articles à part entière suivant le même processus éditorial que les articles scientifiques. Voir : https://doranum.fr/data-paper-data-journal/contenu-data-paper/

Les **entrepôts*** peuvent être de nature et de qualité différentes (institutionnels, thématiques, généralistes). Afin de trouver un entrepôt adapté aux types de données, consulter le catalogue international Re3data qui les recense : http://re3data.org/ ou Cat OPIDoR : https://cat.opidor.fr/index.php/ qui recense les services français de gestion des données de la recherche.

Pour choisir un entrepôt « de confiance », il peut être utile de vérifier la politique de l'entrepôt pour s'assurer d'un certain nombre de critère : présence d'identifiants pérennes (comme un DOI*), coûts éventuels, archivage pérenne, certification de l'entrepôt (CoreTrustSeal), etc. Les caractéristiques et fonctionnalités de l'entrepôt y sont également détaillées.

Pour une liste complète de critères, voir : https://doranum.fr/depot-entrepots/fiche-synthetique/

Exemples d'entrepôts de données :
Entrepôts généralistes pluridisciplinaires : Zenodo (moissonné par OpenAire)
Entrepôts institutionnels : Data INRAE, Data Sciences Po, DataSuds, Cirad dataverse
Entrepôts disciplinaires (STM) : CDS, EELS Data Base, RESIF
Entrepôts disciplinaires (SHS) : Nakala (HumaNum), Cocoon (Collections de Corpus Oraux Numériques), BeQuali (pour les données d'enquêtes)

Différents modes de consultation du jeu de données sont possibles au sein d'un entrepôt : accès ouvert, accès ouvert après demande d'une autorisation, la possibilité de déposer les données sous embargo, accès fermé. Les métadonnées restent accessibles. Quelle que soit la solution choisie, justifier votre choix.

Différentes **licences*** peuvent être utilisées pour la diffusion (et la réutilisation) des données : Licence Ouverte établie par le gouvernement (Etalab) - ou une licence compatible avec elle - et la licence ODbL pour les bases de données (ODC Open Database License). Pour plus de précisions, voir le site gouvernemental : https://www.data.gouv.fr/fr/licences.

Ressources internes :

Pour toute question sur le choix d'un entrepôt, vous pouvez contacter le service des bibliothèques d'UP (contacter : donnees.recherche.dbm@listes.u-paris.fr).

Exemple SH
Le moissonnage dans le moteur de recherche Isidore, par exemple, peut être indiqué ici

Etape 4 – Rédiger

5. Partage des données et conservation à long terme

b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (p. ex. un entrepôt de données ou une archive) ?

Toutes les données collectées, produites et réutilisées dans le cadre d'un projet de recherche ne sont pas à conserver indéfiniment. Un **tri est à réaliser** afin de distinguer les données à conserver car elles représentent un intérêt scientifique, juridique ou historique, des données pouvant être détruites car ne présentant pas d'intérêt sur le long terme.
Dans le cadre de ce tri, il est nécessaire de prendre en compte l'intérêt de réutilisation des données pour d'autres projets de recherche, la conservation des données pour valider des résultats, leur facilité à être reproduites ou non, leur coût etc.

Dans cette partie du PGD, il est nécessaire de préciser les **modalités d'archivage pérenne** des données à conserver à l'issue du projet. Il peut s'agir d'un entrepôt de données ou d'une plateforme d'archivage électronique comme la plateforme du CINES . Il est recommandé d'opter pour des solutions d'archivage pérenne dotées de l'accréditation Data Seal of Approval qui garantit que la plateforme assure bien une préservation des données à long terme (c'est-à-dire au-delà de 30 ans).

Pour déterminer les durées de conservation et les modalités de tri de vos données, vous pouvez vous aider du référentiel des durées de conservation des données disponible sur DoRANum (https://doranum.fr/stockage-archivage/referentiel-de-gestion-des-archives-de-la-recherche/).

Vous pouvez également vous appuyer sur le service archives de votre structure (le Département des Archives pour Université de Paris : archives.daj@u-paris.fr).

5. Partage des données et conservation à long terme

c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Préciser si des outils logiciels seront nécessaires pour l'accès et la réutilisation des données par les utilisateurs.

Indiquer les modalités d'accès aux données qui seraient déposées dans des entrepôts de données : demandes d'accès gérées en direct ou par le biais d'un autre mécanisme etc.

Le principe général guidant l'ouverte des données de la recherche est le suivant :
« Aussi ouvert que possible, aussi fermé que nécessaire »

Etape 4 – Rédiger

5. Partage des données et conservation à long terme

d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Les identifiants (ex. DOI*, Handle, ARK, etc.) permettent une identification pérenne d'un document (publications, jeux de données, etc.). Si l'URL d'un site peut varier au cours du temps, l'identifiant garantit de retrouver l'emplacement d'un document et ainsi, une citation fiable et pérenne. Il permet aussi de lier le jeu de données aux publications ou à d'autres produits de recherche* et assure une conformité avec les principes FAIR*. Un entrepôt pérenne délivrera un DOI.

Voir : https://opidor.fr/identifier/ et https://doranum.fr/identifiants-perennes-pid/identifiants-perennes-apercu/

Exemple SH

Les fichiers contenant les textes lemmatisés seront aussi librement accessibles afin de promouvoir l'Open Access. Les données obtenues grâce à l'outil Biblindex répondront aux mêmes principes que le site Biblindex, les données seront accessibles librement tant que le site Biblindex le sera librement. Le projet pourra bénéficier de la gamme d'outils et de service comme Nakala pour l'exposition des données. Le stockage sera délégué à Humanum et la préservation sera gérée par Humanum. Le but est de pouvoir conserver le document et l'information qu'il contient pour la durée du projet. Les données pourront être retrouvées et partagées par le dépôt dans un entrepôt de données de confiance, comme Nakala. Les données comme les textes lemmatisées pourront être réutilisé dans le cadre d'autres projets de recherche ou pour l'enseignement. L'accès aux données sera faisable via ISIDORE. ISIDORE consulte les données stockées dans Nakala. Nous utiliserons un schéma de métadonnées standard comme le Dublin Core, ceci permettra de rendre interopérables les métadonnées, c'est-à-dire la possibilité de pouvoir les connecter à d'autres entrepôts existants, et de les rendre moissonnables par des services spécialisés comme ISIDORE. Les données auront un identifiant unique : un Handle Nakala.

Source : https://dmp.opidor.fr/plans/5278/export.pdf

Exemple STM

Data will be shared and promoted as soon as their publication under CC-BY license (i.e. free re-use with reference to the publication). These datasets will be first hosted on its webpage as well as on the NOAA repository to guarantee data visibility and accessibility. COPERNICUS services will also be considered to ensure the easy access and use of these datasets to a broader community. Promotion of these datasets will be undertaken through the PAGES Floods WG network and by contributing to international (e.g. UNESCO Sustainable Developments Goals, Future Earth group on Extreme Events and Environments, IPCC Special Report on Extremes and WCRP Grand Challenge on Weather and Extreme Events), regional (e.g. Alpine Convention, Mountain Research Initiative, Interpraevent) and local (PARN) initiatives. At the end of the project, if some parts of datasets are not published yet, they will be gathered and made available publicly through one or several data papers in order to avoid their disappearance. Fieldwork information will be uploaded to the national scientific coring data repository. Paleoflood datasets will then be preserved through the PAGES Floods WG webpage as well as NOAA repository, which is one the favorite repositories of the paleo-science community. COPERNICUS services will also be considered to ensure the easy access and use of these datasets to a broader community such as researchers, engineers or stakeholders. Any specific tool is required to download and use most of data due to their preferred text (.txt / .csv) format. The netcdf format of the climate data can be read with many open software such as R.

Source : https://dmp.opidor.fr/plans/6265/export.pdf

Etape 4 – Rédiger

6. Responsabilités et ressources en matière de gestion des données

a. Qui (p. ex. rôle, position et institution de rattachement) sera responsable de la gestion des données (le gestionnaire des données) ?

Pour garantir **une gestion et un partage optimum des données** et veiller à la **mise à jour du PGD** par l'ensemble des acteurs, il est important que les responsables des données soient clairement désignés et identifiés par l'ensemble des partenaires.

Le ou les rédacteurs de PGD doivent rester impliqués dans le pilotage du projet de sa création à son achèvement, coordonner les actions nécessaires à la mise en œuvre du PGD, effectuer les modifications nécessaires à sa mise à jour et assurer sa transmission au financeur.

6. Responsabilités et ressources en matière de gestion des données

b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Trois sites peuvent être utiles pour évaluer les coûts :

UK Data Service Data management costing tool and checklist : https://www.ukdataservice.ac.uk/media/622368/costingtool.pdf

L'Université d'Utrecht détaille les coûts induits par la gestion des données à toutes les étapes du cycle de vie : https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management

L'Université de Lausanne a développé un calculateur de coût induit par la gestion, le stockage et la publication des données : https://costcalc.epfl.ch/

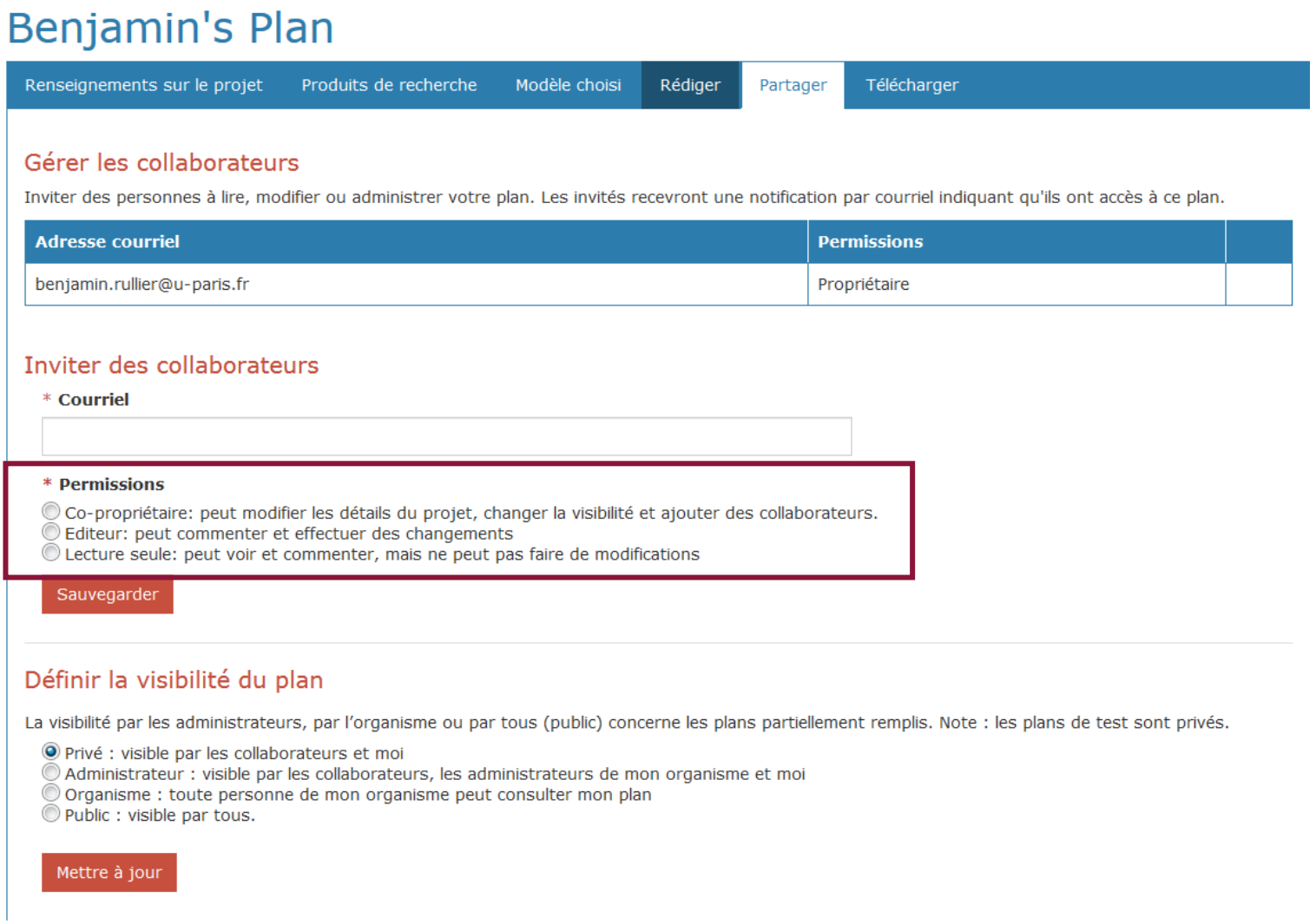
Ressources internes à UP : Contact DSIN (assistance.dsin@u-paris.fr).

Etape 5 – Partager

Il est possible **de partager le PGD avec vos partenaires et/ou des relecteurs extérieurs** en leur attribuant les droits de commenter le PGD.

Par défaut, le DMP est en mode « privé », c'est-à-dire uniquement visible par le créateur du DMP et ses collaborateurs.

Il y a 4 modes de visibilité possibles.



Etape 6 – Télécharger

Télécharger et mettre en forme le PGD autant de fois que nécessaire.

Plusieurs formats sont possibles pour le téléchargement (PDF, HTML, DOCX…).

Définitions

Archivage pérenne : l'archivage numérique pérenne de document numérique et de données se distingue de la sauvegarde et du stockage sécurisé, et répond à trois objectifs principaux :
- conserver le document,
- le rendre accessible,
- en préserver l'intelligibilité.

Ces trois services sont conçus sur le très long terme, c'est-à-dire plus de 30 ans.
Source : https://www.cines.fr/archivage/un-concept-des-problematiques/le-concept-darchivage-numerique-perenne/

Data paper : publication qui décrit un jeu de données scientifiques, notamment à l'aide d'informations structurées (métadonnées). Ce sont des articles à part entière suivant le même processus éditorial que les articles scientifiques. Ils ont pour but de rendre des jeux de données accessibles, interprétables et réutilisables.
Source : https://doranum.fr/data-paper-data-journal/contenu-data-paper/

DMP/PGD : est un outil de gestion. Il se présente sous forme d'un document structuré en rubriques. Il a pour objectif de synthétiser la description et l'évolution des jeux de données de votre projet de recherche. Il prépare le partage, la réutilisation et la pérennisation des données.
Source : https://doranum.fr/plan-gestion-donnees-dmp/

DOI : le DOI (Digital Object Identifier), identifiant pérenne et unique, permet de référencer, citer et fournir un lien stable vers un fichier en ligne. Le but des DOI est de faciliter la gestion sur le long terme de tout objet numérique en ligne à l'aide de métadonnées. Ces métadonnées peuvent évoluer au cours du temps, mais l'identifiant pointant sur la ressource reste invariant. La ressource (données, publications, etc.) peut toujours être citée.
Source partielle : IRD https://data.ird.fr/obtenir-un-doi/

Entrepôt : Un entrepôt permet de stocker des données de recherche, d'y accéder et de les réutiliser. Il existe des milliers d'entrepôts répartis en plusieurs types : disciplinaires, multidisciplinaires, propres à un éditeur, institutionnels, spécifiques d'un projet de recherche…
Source : https://doranum.fr/depot-entrepots/fiche-synthetique/

Licence : La licence choisie par l'auteur engage le ré-utilisateur à respecter l'intégrité des données, à faire mention de la source des données et à porter l'indication de la date de dernière mise à jour.
Source : https://doranum.fr/aspects-juridiques-ethiques/fiche-synthetique/
Afin d'éviter la prolifération des licences, la loi pour une République numérique a prévu la création d'une liste, fixée par décret, de licences qui peuvent être utilisées par les administrations pour la réutilisation à titre gratuit de leurs informations publiques, qu'il s'agisse de données ou de code source d'un logiciel.
Source : https://www.data.gouv.fr/fr/licences

Principe FAIR : Publiés dans Scientific Data en 2016, les principes FAIR (Findable, Accessible, Interoperable, Reusable) fournissent des lignes directrices pour améliorer la facilité de repérage, l'accessibilité, l'interopérabilité et la réutilisation des ressources numériques. Ces principes sont très accès sur la capacité des machines à gérer des données de façon automatique, avec le minimum d'interventions humaines.
Source : https://doranum.fr/enjeux-benefices/principes-fair/

Définitions

Métadonnées : les métadonnées, « données sur les données », sont des informations descriptives qui permettent de renseigner le contenu d'un jeu de données (ex : titre, date de création, format, etc.). Il existe plusieurs types de métadonnées :

- des métadonnées de gestion, permettant d'accéder au document (auteur, titre, date de création, date de modification, langue…)
- les métadonnées de description, pour en comprendre le contenu (sujet, description) ;
- les métadonnées de préservation, pour garantir la pérennité de l'accès et de la compréhension du document (droits, format du fichier, source, résolution, relation, couverture…).

Les métadonnées sont la carte d'identité d'un document. Elles permettent de l'identifier, de le décrire, d'expliquer l'origine de sa création, son utilité et ses destinataires. Au-delà de cette seule description, elles facilitent la recherche et le partage des ressources, la gestion de collections, leur préservation autant que la gestion des droits et l'authentification des documents.
Source : https://www.enssib.fr/le-dictionnaire/metadonnees

Produits de recherche : jeux de données qui vont être gérées, stockées, archivées ou partagées différemment au cours et à la fin du projet de recherche. En effet, un ou plusieurs jeu(x) de données peu(ven)t être lié(s) au projet de recherche, et désigner : a) un lot techniquement homogène, ou b) un lot intellectuellement cohérent même si celui-ci est composé de lots techniquement hétérogènes.

Liste des PGD publics partagés sur DMP OPIDoR ayant servis d'exemples dans ce guide :

SHS :

-https://dmp.opidor.fr/plans/5278/export.pdf

-https://dmp.opidor.fr/plans/4205/export.pdf

STM :

-https://dmp.opidor.fr/plans/6265/export.pdf

-https://dmp.opidor.fr/plans/5082/export.pdf

Merci aux équipes de les avoir partagés !

Contacts
Pour toute question sur les PGD (conseils, relecture, etc.), la gestion des données et l'utilisation de DMP OPIDoR : donnees.recherche.dbm@listes.u-paris.fr

Pour toute question sur la Science ouverte : recherche.dbm@listes.u-paris.fr ou sur HAL : hal.dbm@listes.u-paris.fr.

Université de Paris

Ce guide a été conjointement réalisé par la Direction générale déléguée des bibliothèques et musées et le Département des archives de Université de Paris

V3 - Avril 2021