Le Plan de Gestion de Données pas à pas

Pour accéder à la ressource sur DoRANum : https://doranum.fr/plan-gestion-donnees-dmp/le-plan-de-gestion-de-donnees-pas-a-pas_10_13143_t94g-9j96/
Pour accéder à la ressource sur Callisto : https://callisto-formation.fr/course/view.php?id=160

Date de publication : 06/03/2023
Date de dernière mise à jour : 17/07/2024

Sommaire

Bienvenue

Bienvenue dans ce parcours pédagogique sur le Plan de Gestion de Données (PGD) ou Data Management Plan (DMP).

Ce parcours a été conçu en pensant à toutes les questions qui peuvent se poser lors de la rédaction d'un PGD.
Pour chacune des questions, plusieurs exemples extraits de PGD rendus publics par leurs auteurs sont présentés dans un carrousel encadré de bleu. Vous pouvez vous inspirer de ces exemples pour rédiger votre PGD.

Les chapitres peuvent être suivis de façon linéaire et progressive, mais aussi de manière fragmentée. Vous pouvez consulter uniquement les parties qui vous intéressent.

1. Pourquoi est-ce important de bien gérer ses données ?

Cela vous est-il déjà arrivé de vouloir réutiliser les données d'un autre chercheur et de vous retrouver dans la même situation que ces deux personnages ?

(Pour cette vidéo de 5 min, vous pouvez accéder à une traduction automatique des sous-titres anglais en français en allant dans : paramètres (roue crantée) > sous-titres > traduire automatiquement puis choisir la langue "Français".)

Data Sharing and Management Snafu in 3 Short Acts

Pour éviter de vous retrouver dans une situation similaire et pour vous aider à mieux gérer vos données, il existe un outil : le Plan de Gestion de Données (PGD).

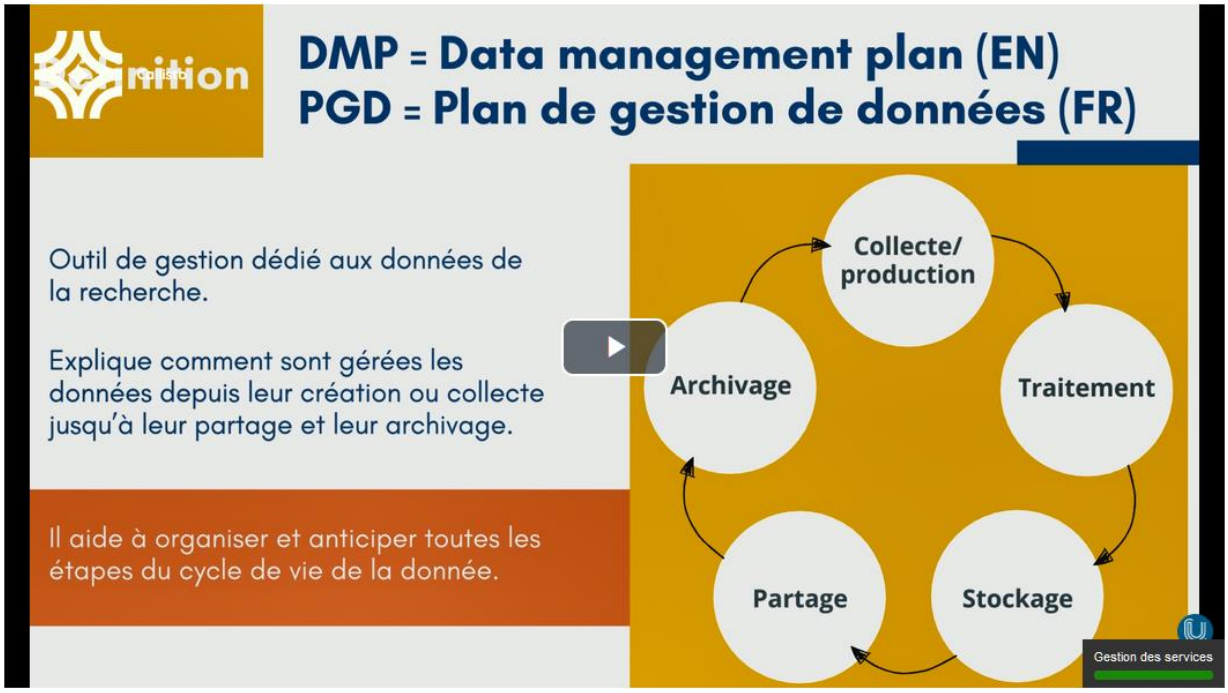
2. Premières questions à se poser sur le PGD

2.1. Qu'est-ce qu'un PGD ?

Le plan de gestion des données est un outil de gestion. Il se présente sous la 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.

Cette vidéo de 3 min vous permettra de mieux comprendre ce qu'est le PGD :



https://www.canal-u.tv/chaines/callisto/les-minutes-doranum/la-minute-plan-gestion-de-donnees

2.2. Qui recommande le PGD ?

La ressource ci-dessous vous permettra de savoir qui recommande le PGD, à différents niveaux :

https://view.genial.ly/62b1bbacdfcd54001097d23f

2.3. Dois-je rédiger un PGD pour les données préexistantes dans mon labo ?

Il est tout à fait possible de rédiger un PGD pour des données préexistantes :

- Pour lesquelles il n'y a pas eu de politique de gestion des données
- Qui ne se rapportent à aucun projet en cours.

Il est également possible de rédiger un PGD suivant un modèle spécifique pour son laboratoire ou son infrastructure. Ce modèle dit " d'entité " sera expliqué dans le chapitre 5.4 de ce cours sur les modèles de plan de gestion de données, dans le module " Comment choisir un modèle de PGD ? ".

2.4. Quelle est la différence entre un PGD de projet et un PGD d'entité ?

	PGD de projet	PGD d'entité
Stratégie de gestion des données de recherche	Sur toute la durée d'un projet	Sur toute la durée de vie d'un laboratoire, d'une plateforme…
Données concernées	Données de recherche préexistantes ou produites, utilisées pendant le projet	Toutes les données de recherche gérées par l'entité
Durée	Déterminée (pendant la durée du projet)	Indéterminée (vision à plus long terme)

Dufayard, J.-F., & Lieby, P. (2024, juin 21). Les Plans de Gestion des Données entités. Zenodo. https://doi.org/10.5281/zenodo.12205835

2.5. À quel moment dois-je rédiger un PGD et combien de temps cela prend ?

Comme mentionné dans la vidéo sur le PGD, c'est un document évolutif.
Il faut commencer à le rédiger dès le début du projet, avec les éléments déjà connus ou prévus. Ensuite, vous pouvez le compléter au fur et à mesure. Il faut prévoir 2 versions au minimum : au début et à la fin du projet.
Pour les projets de plus de 30 mois, une version intermédiaire est demandée. Le temps de rédaction d'un PGD dépend :

- De l'étendue du projet,
- Du type et de la diversité des données,
- Des ressources humaines et des compétences disponibles,
- De votre degré d'expérience en rédaction de PGD…

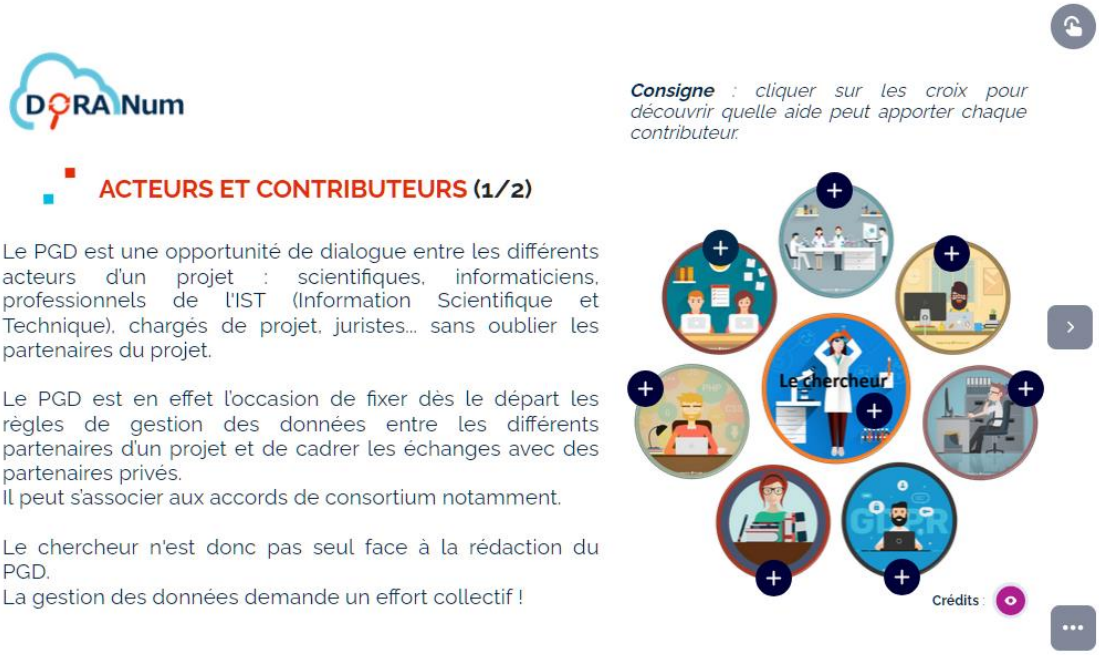
La mise en place de bonnes pratiques et la rédaction d'un PGD vous permettront de gagner du temps par la suite.

3. Avec qui rédiger un PGD ?

3.1. Les différents acteurs

Vous devez rédiger un PGD et le mettre en œuvre ? Vous ne vous sentez pas compétent pour le faire ? Vous pensez ne pas avoir le temps ? Vous pouvez solliciter de l'aide à différents niveaux.

Cette ressource vous donnera un aperçu des différents acteurs qui peuvent vous accompagner dans la rédaction de votre PGD :



https://view.genial.ly/62b1d4133f752e0017115538

Le répertoire des Services Opérationnels de Soutien à la rédaction des Plans de Gestion des Données (**SOS-PGD**) recense les services accompagnant la rédaction des plans de gestion des données au sein des établissements d'enseignement supérieur et de la recherche. Il vise à aider les chercheurs à identifier leurs interlocuteurs au sein de leur institution et à faciliter la mise en relation entre les services supports de différentes institutions pour les projets de recherche multi-partenariaux.

Dans le cadre de l'écosystème ***Recherche Data Gouv***, les **ateliers de la donnée** sont en proximité géographique des équipes de recherche pour leur apporter une première expertise dans la gestion raisonnée des données de recherche.

Vous trouverez également dans le catalogue **CatOPIDoR** un recensement des services dédiés aux données de la recherche en France.

3.2. Qui doit rédiger le PGD ? Doit-on le rédiger en collaboration avec les partenaires du projet ?

En général, c'est le coordinateur de projet qui rédige le PGD. Cependant, n'importe laquelle de ces personnes peut s'en charger :

- Le gestionnaire des données,
- Un chercheur de l'équipe,
- Un partenaire du projet,
- Le documentaliste du laboratoire,
- Toute autre personne désignée par les membres du projet.

Cette personne sera également chargée des mises à jour du PGD.
Il est tout à fait possible de le rédiger de manière collaborative (notamment avec les partenaires).

Voici deux exemples pour illustrer cette question :

- « Le gestionnaire de données sera l'ingénieur recruté. […] La saisie des données et la production des métadonnées sera assurée par les chercheurs du projet. L'ingénieur s'occupera de la qualité des données, du stockage et de la sauvegarde ainsi que de l'archivage et du partage des données. Il sera responsable de la mise en œuvre du plan de gestion de données. […]
  La personne référente à la fin de l'ERC concernant les gestions des données sera le porteur du projet. […] »
  Exemple extrait du PGD public Hospitam, Lauriane Locatelli (ENS Lyon)

- « Le responsable de la gestion des données est le Chef de projet RHU PrediMAP-APHP.
  Les personnes impliquées dans le traitement des données cliniques :

- Technicien de recherche clinique pour la saisie des données clinique
- Attaché de Recherche Clinique : gestion et traitement des queries
- Data management programmation de la base de données / eCRF, pour l'extraction, fusion des bases et le nettoyage de la base finale (contrôles de cohérence) vérification de la qualité des données
- Statisticien pour l'analyse et l'archivage de la base
- Responsable de l'unité de recherche clinique
- Chef de projet promotion APHP pour les demandes réglementaires (CNIL…) en lien avec la recherche clinique
- Délégué à la protection des données pour rédaction et supervision des contrats à mettre en place entre les partenaires pour le transfert des données.

Données biologiques :
- Chef de projet pour les données du laboratoire : analyse, archivage et partage des données
- Chercheurs, ingénieurs, doctorants pour la production, analyse
- Bioinformaticiens : gestion et traitement des données, vérification de la qualité des données, production des métadonnées.

Métadonnées : DAC (Data Access Consortium) : Chef de projet de laboratoire. »

Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

Vous n'êtes pas obligé d'assumer seul la rédaction : n'hésitez pas à vous faire aider.

4. Modèles de PGD

4.1. Existe-t-il une trame pour rédiger un PGD ?

Il n'existe pas de trame unique. Toutefois de nombreux modèles de PGD ont été établis par des organismes, instituts, financeurs à destination de leurs utilisateurs, afin de répondre aux spécificités propres à certains organismes de recherche, pour correspondre au contexte local des établissements, etc.

On y retrouve néanmoins les mêmes éléments, à savoir :

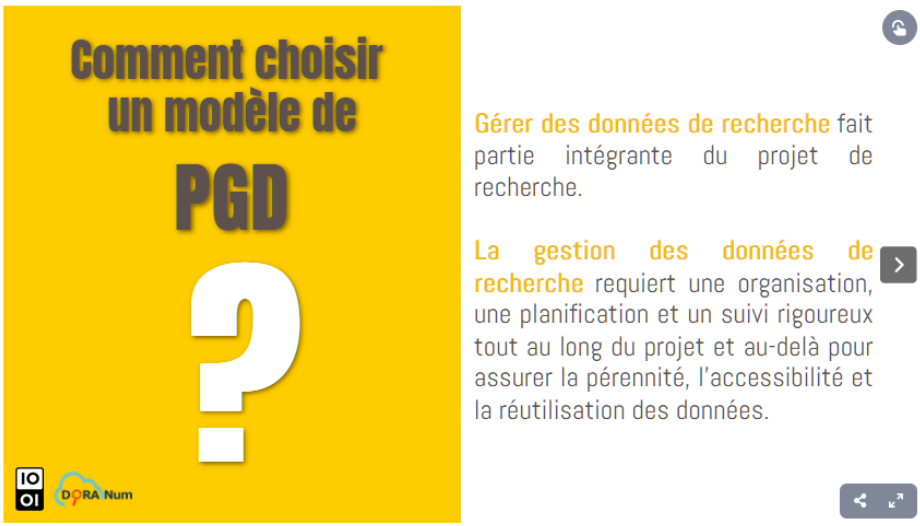
- Informations administratives
- Description des données
- Documentation, métadonnées, standards

- Aspects juridiques
- Sécurité des données
- Stockage des données durant le projet
- Partage des données après le dépôt dans un entrepôt
- Archivage pérenne
- Coûts
- Responsabilités

4.2. Les financeurs exigent-ils un modèle précis ?

Les financeurs peuvent proposer un modèle mais aucun modèle n'est obligatoire. Vous pouvez choisir celui qui vous convient le mieux, par exemple celui de votre établissement s'il en propose un.
La présentation suivante vous aide à choisir entre différents modèles :

- Modèle Science Europe
- Modèles de financeurs
- Modèles institutionnels
- Modèles d'entité
- Modèles disciplinaires
- Modèles pour les logiciels



https://doranum.fr/plan-gestion-donnees-dmp/comment-choisir-un-modele-de-plan-de-gestion-de-donnees_10_13143_2h9s-kt61/

Le choix du modèle est important. Il doit répondre à vos besoins.

5. Que contient un PGD ?

Ce chapitre est subdivisé en 8 parties qui suivent les grandes sections d'un Plan de Gestion de Données et répondent à de nombreuses questions que vous pouvez vous poser :

- Description des données
- Documentation / Métadonnées
- Exigences légales et éthiques
- Traitement et analyse des données
- Stockage, partage et archivage : quelles différences ?
- Stockage et organisation des données
- Partage des données
- Archivage pérenne des données

5.1. Description des données

5.1.1. Mes données, quelles sont-elles ?

Pour avoir des exemples concrets, visionnez cette vidéo d'EOSC (European Open Science Cloud) de 5 min 30 réalisée dans le contexte du projet EOSC-Pillar dans laquelle des chercheurs français parlent des données couramment utilisées dans leurs disciplines scientifiques :

What is data? - Interviews with French researchers :
https://www.youtube.com/watch?v=J7nkClygNng

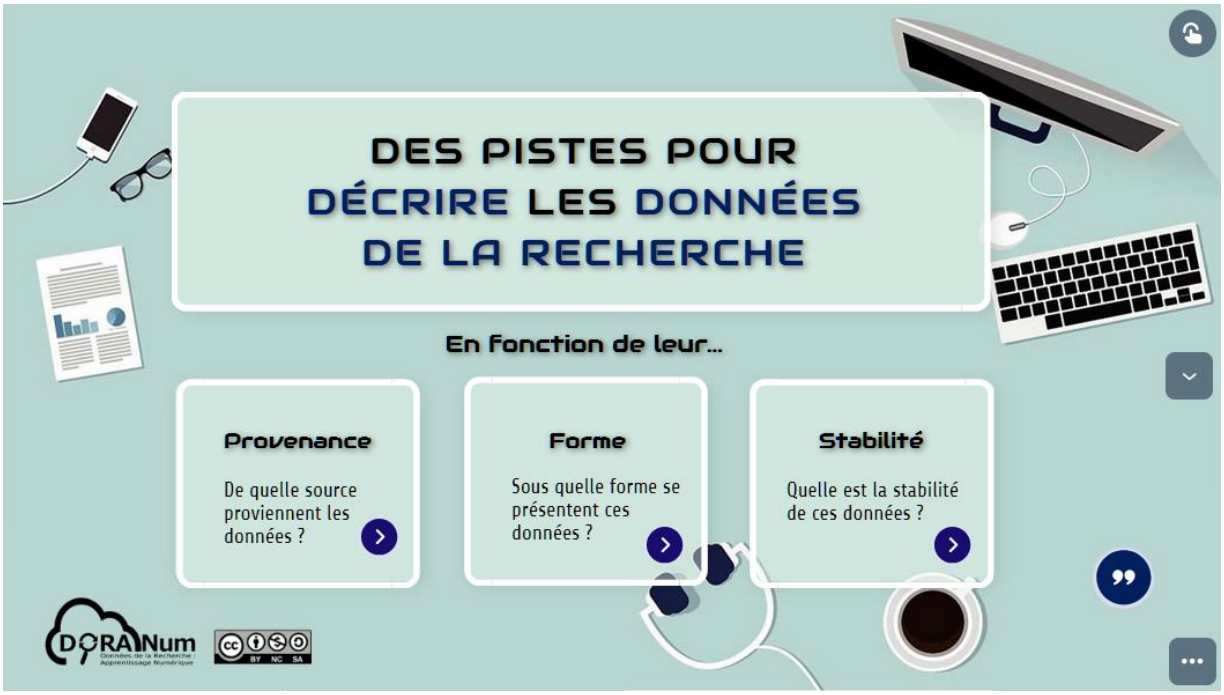
« Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images, son, vidéo…) utilisés comme sources primaires pour la recherche scientifique et qui sont habituellement acceptés par la communauté scientifique comme étant nécessaires pour valider les résultats de la recherche. »

- *OCDE, Organisation de Coopération et de Développement Économiques. Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. Paris, 2007. https://doi.org/10.1787/9789264034020-en-fr*
- *MESR, Ministère de l'Enseignement supérieur et de la Recherche. Deuxième Plan national pour la science ouverte. Généraliser la science ouverte en France 2021-2024. Juillet 2021. https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2021-09/2e-plan-national-pour-la-science-ouverte-12968.pdf*

Pour aller plus loin sur la définition des données de la recherche, consultez **cette page de DoRANum**.

5.1.2. Comment décrire mes données ?

Dans cette ressource vous trouverez des pistes pour décrire vos données de recherche :



https://view.genial.ly/620e2bf1829d7400113ad0c3

Voici plusieurs exemples pour illustrer différentes questions sur la description des données :

Comment décrire des données produites ?

- « Les données qui seront produites [...] sont de quatre types :
  - des données microclimatiques recueillies in-situ sur les différents sites d'étude envisagés dans le projet ;
  - des données d'observations du milieu réalisées in-situ (p.ex. données de localisation géographique, relevés dendrométriques et botaniques, etc.) ;
  - 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.) ;
  - des données issues de la télédétection (imagerie LiDAR). [...] »

  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

- « Le baluchon est un sac comportant un carnet de bord, un appareil photo (jetable) si la personne n'a pas de téléphone portable qui prend des photos, un enregistreur numérique de poche, ainsi que plusieurs enveloppes pour recueillir des « objets ». Confié et élaboré par les habitants sur une semaine. Il s'agit d'un objet multiforme composé selon les habitants de texte, photo, dessin, collage, objet, son... Cette méthode permet de connaître et comprendre les expériences sensorielles et sensibles des habitants. [...] »

  Exemple extrait du PGD public PROSECO, Théa Manola (CNRS)

- « The following data will be collected during the project for the WP-1:
  - Data on capture effort for semi-invasive and non-invasive sampling (trap surface/number, time...) (format excel, CSV) 10 Mo
  - Number of individuals captured per species, Capture-Mark-Recapture data (format excel, CSV) 10 Mo
  - Bat population counts (format excel, CSV) 10 Mo
  - Study site description (format excel, CSV) 10 Mo
  - Observations on human/animal interactions (format excel, CSV) 10 Mo
  - Fruit tree phenology (format excel, CSV) 10 Mo
  - Data on biological samples collected from captured bats (N=4000) and Rodents (N=2400): Rectal and salivary swabs, Blood (Dried Blood Spot), Hair, Skin, organs (=200) (format excel, CSV) 20 Mo

- Morphological measurements, traits and physiological data from captured bats and rodents: Age category, gender, reproductive status (format excel, CSV), 10 Mo
- Pictures (study site, Bats, Rodents, CYT B agarose gel, Format JPEG), 1To
- Data on bat feces collected by using non-invasive method (=1800) (format excel, CSV) 10 Mo
- Meteorological data: Temperature and humidity , weather at study sites (format excel, CSV), 10 Mo
- Nucleotidic and proteic acid alignements (fasta files), 100 Mo
- Phylogenetic trees (PDF, power point), 10 Mo
- General Database (REDCap-Research Electronic Data Capture), 100 Go
- Excel and CSV sheet: Victoria DB Bats, Victoria DB rodents, Victoria Seasonal field site data, Victoria daily field site data. […]
- Data on serological samples from human population (excel and CSV format) 100 Mo
- Serological tests and results from human samples (excel and CSV format), 100 Mo
- Questionaries, focus group and semi-structured interviews qualitative data (CSV, Excel, PDF and Doc format) 20 Go
- Astrovirus, Coronavirus and Paramyxovirus nucleotidic sequences (format Fasta, ABI, TXT, PDF), 50 Go »

Exemple extrait du PGD public VICTORIA, Hélène De Nys, Florian Liegeois (CIRAD et IRD)

Comment décrire des données réutilisées ?

- « Les modèles hydrologiques s'appuient sur des développements existants :
  - Le code de simulation CWATM : https://cwatm.iiasa.ac.at/
  - La version couplée avec le code de simulation hydrogéologique MODFLOW : https://gmd.copernicus.org/articles/15/7099/2022/gmd-15-7099-2022.html
  - Le code de projection de l'occupation des sols Foresight : https://journals.openedition.org/cybergeo/27397
  - À titre d'exemple, ces codes ont été déployés sur les bassins versants sur Scorff et du Blavet pour une action de modélisation prospective participative avec Lorient Agglomération, publiée dans la thèse d'Elias Ganivet : https://www.theses.fr/s239878
  Il s'agit de mettre en œuvre ces codes sur les 2 territoires non bretons (bassin versant associé au lac d'Annecy et bassin versant associé à l'agglomération Pays Basque). Sur

le territoire breton, des éléments issus des prospectives avec Lorient pourront être réutilisés.
Une documentation technique accompagne le code déployé sur les territoires et permettra de s'approprier et réutiliser les codes.
Les données d'entrées seront issues de différentes sources :
- Sorties du modèle SURFEX de Météo France pour les données de forçage historiques.
- Paramètres de sol (épaisseur, propriétés) issues de bases de données européennes - https://esdac.jrc.ec.europa.eu/ - ou plus précises à définir en interaction avec les territoires.
Les données d'évaluation suivantes seront mobilisées sur les bassins versants :
- Les données de débit des cours d'eau de la base nationale - https://hydro.eaufrance.fr/
- Les données piézométriques de la base nationale ADES - https://ades.eaufrance.fr/Recherche
- Éventuellement les données d'étiage de la base de données nationale ONDES - https://onde.eaufrance.fr/
Pour la prospective, les données suivantes seront utilisées :
- Ensemble de projections climatiques futures du portail DRIAS - https://www.drias-climat.fr/
- Scénarisations d'occupation des sols, d'usage des sols et d'accroissement de la population issues des ateliers prospectifs sur les territoires.
Les sorties des modèles seront stockées sous format natif et transformée en données scientifiquement exploitable sous forme d'images ou de données cartographiques. La zone de stockage de données est à identifier.
Les données élaborées des modèles (conditions hydriques, etc.) seront mobilisées par l'institut du Design pour crée des images de paysages futurs réalistes. »
Exemple extrait du PGD public PAGAIE, Laurent Longuevergne, Véronique Van Tilbeurgh (CNRS)

- Le projet réutilisera les données de la cohorte Constances https://www.constances.fr/.
Un questionnaire spécifique au projet sera réalisé auprès de 80 0000 volontaires de la cohorte ayant déjà renseigné leurs historiques résidentiels.
Un croisement sera par ailleurs réalisé avec des bases de données environnementales de l'Ineris (plateforme PLAINE https://www.ineris.fr/fr/dossiers-thematiques/tous-dossiers-thematiques/inegalites-environnementales/plateforme-analyse) et avec plusieurs bases de données sur l'exposition environnementale, médicale et professionnelles aux rayonnements ionisants de l'IRSN dont SISERI.
Toutes ces nouvelles données permettront d'estimer les expositions à plusieurs substances chimiques et aux rayonnements ionisants des membres de la cohorte

Constances. Ces données d'expositions seront ensuite intégrées à la base de données générale de la cohorte Constances. […]

Exemple extrait du PGD public COREXCA, Olivier Laurent (Inserm)

Comment décrire les protocoles / méthodes / logiciels utilisés ?

- « Toutes les expériences et leurs résultats seront rapportés sur le cahier de laboratoire électronique (CLE Inserm).

  1. Données inédites de quantité de molécules générées par dosage ou par immuno-PCR dans les sécrétions cervicovaginales générées par dosage ELISA et analyse spectrophotométrique, les données brutes seront importées du spectrophotomètre et stockées au laboratoire sous forme de fichiers excel.
  2. Images de tissus humains inédites qui seront obtenues après immunofluorescence multiplex pour 30 femmes. Ces données seront recueillies par un scanner de lames (Lamina slide scanner" haut débit, Akoya Perkin Elmer). Les lames virtuelles générées par le scanner seront importées et stockées dans la CID (Cochin Image Database) sur un site dédié et dont l'accès est protégé.
  3. Recueil de données génétiques inédites (séquences d'ARN) par transcriptomique spatiale et séquençage à haut débit (NextSeq™ 500 Illumina) pour 30 femmes. Les données brutes seront traitées avec Space Ranger pour réaliser les alignements et les comptages des UMI (Unique Molecule Identifier) et la génération des matrices associant coordonnées spatiales et données traitées. Les données cliniques et de mesures des biomarqueurs dans les sécrétions cervico-vaginales de la cohorte Inspire, seront réutilisées pour les analyses et la réalisation de l'algorithme. »

  Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

- « Enquête qualitative

  Plusieurs méthodes sont utilisées :

  1. Entretiens, avec des habitants et des professionnels,
  2. Entretiens auprès des professionnels peuvent inclure une partie marchante in situ,
  3. Parcours commentés auprès d'habitants,
  4. Ateliers (sur la base de la méthode des focus groups),

  - Ateliers réunissant des professionnels enquêtés et des membres de l'équipe scientifique. Environ 4 ateliers.
  - Ateliers habitants afin de mettre en débat des différentes expériences mais aussi aborder les nouveaux aménagements et les enjeux socio-politiques. 3 ateliers.

5. Entretiens post-baluchons avec les habitants : remise des matériaux et réflexivité sur le protocole. »

Exemple extrait du PGD public PROSECO, Théa Manola (CNRS)

- « Les données d'observation des poissons et de l'eau du lac (comptage, mesures et dosage) seront saisies sur un carnet de terrain puis retranscrites sous tableur le soir même (un classeur par campagne) avant de les valider et de les inclure dans la base de données Savoie.
  Ces données seront exploitées pour des analyses statistiques ultérieures (logiciel R), jusqu'à la fin du contrat ANR au 30 avril 2023.
  En dehors des campagnes de prélèvement, les échantillons d'eau et de nageoires recueillis sur le bateau, seront amenés en laboratoire pour analyses microscopiques (données d'observation) toxicologiques et moléculaires (données expérimentales. […]
  L'ensemble des expérimentations en laboratoire (protocoles, suivi du déroulement de l'expérience et résultats) et les étapes d'analyses bio-informatiques et statistiques seront consignés dans un cahier de laboratoire papier. »
  Exemple extrait du PGD public « PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

Grâce à ces exemples concrets, vous voyez qu'il est relativement facile de décrire ses données.

5.2. Documentation / Métadonnées

5.2.1. Comment faire en sorte que mes données soient compréhensibles pour les autres chercheurs ?

Ce sont principalement les métadonnées qui permettent de rendre compréhensibles les données.
Pour en savoir plus consultez ce cours introductif sur les métadonnées :

https://doranum.fr/metadonnees-standards-formats/cours-introductif-sur-les-metadonnees_10_13143_vwce-g965/

5.2.2. Comment documenter ses données de manière plus précise ?

Des standards de métadonnées ont été créés par les communautés scientifiques, afin de décrire plus précisément les données et surtout de les rendre interopérables.

Voici deux exemples pour illustrer cette question :

- « 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. [...]
Une documentation détaillée des données utilisées ainsi que des analyses effectuées à partir des données sera également produite au format RMarkdown (.rmd) à partir du logiciel libre RStudio. A partir du code brut (.rmd), plusieurs formats de sorties sont possibles (.html, .pdf, .docx) pour documenter les données en terme de :
(i)méthodologie utilisée pour acquérir ou analyser la donnée ;
(ii)nature des variables utilisées ;
(iii)d'unités de mesure utilisées ;
(iv)de distribution de la donnée.
Des sorties graphiques permettant une meilleure visualisation des données et de leurs distributions peuvent être intégrées directement dans le document final [...] »
Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

- « Glacier data
  Collected raw data:
  Metadata for raw data is meticulously documented in a notebook, accessible to all researchers within the lab. The recorded information encompasses date and time collection (in UTC format) and location. It also includes records of anomalies during data collection, such as measurement-related issues or problems with sensors. Additionally, any actions taken during data collection, such as cleanings or specific operations, are documented.
  Teletransmitted raw data:
  There is a file called "SAV" that is teletransmitted daily. It includes information about sensor programs, the temperature of the location, and details about the battery status. It also contains diagnostic variables, data transmission issues or information about sensors that need to be emptied or cleaned. »
  Exemple extrait du PGD public GLACIOCLIM, Daniel Arroyo, Isabelle Gouttevin (Météo-France), Delphine Six (OSUG)

5.2.3. Aucun standard de métadonnées ne répond à mes besoins : comment faire ?

Si aucun standard dans votre discipline ne répond à vos besoins, il est possible de créer votre propre schéma de métadonnées

- Soit à partir d'un standard généraliste comme *Dublin Core* ou *DataCite Metadata Schema*,
- Soit à partir d'un standard disciplinaire.

Il s'agit ensuite de le compléter par un fichier avec les métadonnées correspondant à vos besoins spécifiques. Concrètement, cela consiste à lister vos métadonnées sous forme de tableur, avec l'intitulé de la métadonnée d'un côté et le descriptif en face.

L'utilisation d'un standard de métadonnées, notamment disciplinaire, est un élément clé pour atteindre un haut degré de respect des principes FAIR. Cela concerne l'*Interopérabilité*, mais aussi les caractères *Facile à trouver* (une donnée n'est souvent trouvable que par les éléments de métadonnées indexés dans le moteur de recherche consulté), *Accessible* (notamment les métadonnées) et *Réutilisable* (provenance décrite dans les métadonnées, licence attribuée et surtout le côté standard disciplinaire).

5.2.4. Comment remplir les champs de métadonnées ?

Dans la plupart des disciplines, il est conseillé d'utiliser des référentiels ou des vocabulaires contrôlés (taxonomies, thésaurus, ontologies...). Ceux-ci facilitent l'intelligibilité et l'interopérabilité des données.

Voici un exemple pour illustrer cette question :

- « Référentiels
  - le référentiel taxonomique TAXREF v. 13 (INPN) sera utilisé pour référencer les espèces
  - la norme internationale ISO 80000-1:2009 sera utilisée pour les unités
  - les stations seront référencées en WGS84 et le référentiel de l'IGN sera utilisé pour les noms des communes

- les standards en biologie moléculaire seront utilisés, en particulier le code IUPAC pour le séquençage d'ADN. »

Exemple extrait du PGD public « PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

5.2.5. Comment m'assurer que les données auxquelles j'accède ou que je produis sont fiables ?

Il est important de mettre en place un contrôle qualité des données.
Cela concerne tout ce qui environne et permet de montrer la rigueur des méthodes et la qualité des données (par exemple, processus de calibration, mesures répétées, contrôles standards positifs/négatifs, contrôle des données en double aveugle ou par évaluateur externe, etc.).

La qualité des données est pour tous un élément capital en vue de leur diffusion et de leur réutilisation. Mais quels sont les éléments qui permettent de dire que des données sont de qualité ? Cette notion concerne à la fois leur qualité intrinsèque ainsi que la qualité des métadonnées associées.
Cette vidéo de 31 min présente les différents principes de la démarche qualité :

https://www.canal-u.tv/chaines/ad/qualite-des-donnees/quel-lien-entre-qualite-et-donnees-alain-rivet-cermav-et-henri

Voici trois exemples pour illustrer cette question :

- « 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. »
  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

- « Les Bonnes Pratiques de Laboratoire seront suivies pour le contrôle et la qualité des données. Au niveau des données de spectrométrie de masse, la qualité et la conformité des données seront contrôlées par les divers contrôles qualités et les calibrations des appareils réalisés régulièrement. De plus des réplicas biologiques et techniques seront réalisés pour des études de protéomique quantitative. Au niveau de la synthèse de nouvelles molécules, les spectromètres RMN sont calibrés et révisés par le Laboratoire de Mesure Physique de l'IBMM. Concernant les mesures de diffraction aux rayons X, les lignes de lumières sont calibrées au quotidien par les responsables de ligne. Le traitement des données à la volée, fournissant un grand nombre d'indicateurs de qualité, permet également de contrôler la qualité des données. »
  Exemple extrait du PGD public LipInTB, Jean-François Cavalier (CNRS)

- « Pour les données cliniques et biologiques :
  1. Monitoring : l'étude a été classée à risque minimal (risque A). L'attaché de recherche clinique n'effectuera pas de contrôle des données saisies sur l'eCRF par rapport aux données sources.
  2. Data management : 3 types de contrôles sont prévus dans le plan de data management :
  - Queries automatiques CleanWeb exécutés périodiquement (lot de queries sur l'eCRF)
  - Contrôle lors du saisie exécuté en temps réel dès la saisis
  - Contrôle à posteriori par le statisticien.
  À l'issue de chaque lot de queries, un rapport de data management est rédigé par le data manager. Ce rapport contient différents indicateurs sur le taux de saisie et l'évolution du statut des queries.

Pour les données biologiques :

Les analyses bioinformatiques et statistiques seront réalisées régulièrement. Les fichiers seront analysés par au moins deux investigateurs de façon indépendante. Des rapports réguliers seront produits qui contiendront différents indicateurs de traçabilité et degré de confiance des données et résultats obtenus. [...] »

Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-hadjadene (APHP)

Grâce à ces exemples concrets, vous comprenez l'importance de bien documenter les données.

5.3. Exigences légales et éthiques

Les aspects juridiques et éthiques accompagnent tout le cycle de vie des données. Dès le début du projet, il faut s'y intéresser et se poser les bonnes questions. Dans le PGD, il est très important de renseigner correctement cette partie.
Dans certains cas, en fonction du type de données, il est possible que vous ne soyez pas concernés par les aspects juridiques et éthiques. Si c'est le cas, il suffit tout simplement de l'expliquer dans le PGD.

Pour en savoir plus concernant ces aspects, consultez cette ressource :

https://doranum.fr/aspects-juridiques-ethiques/cours-introductif-sur-les-aspects-juridiques-et-ethiques_10_13143_eanf-pz90/

5.3.1. Que dois-je faire avec des données personnelles / sensibles ?

Cette vidéo d'1 min réalisée par la CNIL vous explique ce qu'est une donnée personnelle et précise ce que sont les données dites " sensibles " :

https://video.cnil.fr/w/qkLeFsgBYLWPRZiaXocn3N

Pour être éventuellement partagées, les données sensibles doivent être anonymisées et chiffrées. Les données personnelles doivent respecter le Règlement Général pour la Protection des Données (RGPD).
Durant le projet, il peut être nécessaire de limiter l'accès aux données aux seuls membres de l'équipe de recherche. Une fois le projet achevé, il peut être tout aussi important de limiter l'accès aux données.
Dans le cas d'un partenariat, l'accord de consortium prévoit les règles qui encadreront l'accès, l'utilisation, l'exploitation et la propriété intellectuelle des résultats générés par les partenaires pendant le projet ainsi que les conditions d'accès, d'utilisation et d'exploitation des données.

En cas de doute, il ne faut pas hésiter à prendre conseil auprès des juristes de votre institution.

Quelques questions/réponses de la ressource " Les données : Les questions à se poser pour leur diffusion " traitent des données sensibles.

Pour en savoir plus concernant le RGPD, consulter le webinaire Tuto@mate " Le RGPD appliqué aux SHS ", ainsi que la ressource ci-dessous sur la protection des données personnelles et le RGDP dans la recherche :

https://doranum.fr/aspects-juridiques-ethiques/protection-des-donnees-personnelles-et-rgpd-dans-la-recherche-consequences-obligations-implications_10_13143_34ef-n525/

Voici quatre exemples pour illustrer ces aspects :

- « Lors de la publication et de la valorisation de la recherche, un codage est utilisé pour référer à chaque habitant, les propos sont ainsi anonymisés et les photos sont retravaillés (floutage, bandeau noir). […]
  Avant le début de l'enquête de terrain, un travail en collaboration avec la DPD du CNRS a permis la validation des protocoles d'enquête. Le certificat d'inscription au registre tenu par la DPD du CNRS a été délivré le 29/09/2021 (n° 2-21205) permettant la mise en œuvre du traitement dans le respect du RGPD.
  Pour chaque enquête, une note d'information sera explicitée et transmise aux participants. Il s'agira ici de s'assurer que les personnes comprennent les engagements de leur participation, et qu'elles connaissent leur droit de retrait et de modification.
  Un formulaire de consentement et autorisation d'enregistrement de la voix et de l'image ainsi que de son exploitation sera signé et conservé par le responsable scientifique de la recherche pour la durée effective de la recherche, qui est de 3 ans après la fin du projet. […]

Un processus d'anonymisation sera mis en place pour tous les enquêtés sauf pour ceux (essentiellement pour les professionnels) qui auront demandé explicitement d'être nommés. »

Exemple extrait du PGD public PROSECO, Théa Manola (CNRS)

- « In order to comply with the General Data Protection Regulation (GDPR), we are going to implement the following:
  - Obtain consent from individuals regarding the use of their personal contact information.
  - An email will be sent once per year reminding them of the possibility to modify or delete their personal information. »

  Exemple extrait du PGD public GLACIOCLIM, Daniel Arroyo, Isabelle Gouttevin (Météo-France), Delphine Six (OSUG)

- « Le fichier informatique utilisé pour cette Investigation Clinique est mis en oeuvre conformément à la règlementation française (loi Informatique et Libertés modifiée) et européenne (Règlement Général sur la Protection des Données –RGPD).
  [...] Le consentement libre, éclairé et écrit de la personne est recueilli par l'investigateur principal ou par un collaborateur déclaré et formé à l'investigation clinique avant l'inclusion de la personne dans la recherche.
  [...] A l'issue de l'Investigation Clinique, les échantillons pourront être utilisés pour des analyses ultérieures non prévues dans le protocole pouvant se révéler intéressantes dans le cadre de la grossesse et de ses complications, en fonction de l'évolution des connaissances scientifiques, sous réserve que la patiente ne s'y soit pas opposée, après en avoir été informée, comme indiqué dans le formulaire d'information/consentement.
  Les patientes seront informées que les données génétiques feront l'objet d'un traitement informatique et pourront être transférées à une organisation européenne pour archivage dans l'intérêt public, à des fins de recherche scientifique, sous réserve qu'elles ne s'y opposent pas, comme indiqué dans le formulaire de consentement conformément aux articles 13 et 14 du RGPD. »

  Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

- « - Animal and human ethical clearances will be request from Zimbabwe and South Africa institutions (Ministries, Ethical committee etc..).
  - For the human part of the project (questionnaries and serological analyses), samples will be anonymized.
  - Management of Personal data will comply with the General Data Protection Regulation (GDPR)

- Nagoya protocol will be followed.
- Consent form (in English and vernacular languages) will document the voluntary approval of participants. »

Exemple extrait du PGD public VICTORIA, Hélène De Nys, Florian Liegeois (CIRAD et IRD)

5.3.2. Que dois-je faire avec des données confidentielles ?

Voici un exemple pour illustrer cette question :

- « Les jeux de données génotypiques et phénotypiques sur les 5 génotypes du partenaire privé IFV seront considérés comme confidentiels. […]
  Une clause de confidentialité sera incluse dans l'accord de consortium pour ce matériel. […]
  Les données seront sécurisées sur un serveur dédié, avec un mot de passe limitant l'accès et un cryptage de la communication. L'accès à ces données sera examiné sur demande selon les règles définies dans l'accord de consortium. »
  Exemple extrait du PGD public G2WAS, Cédric Goby et Laurent Torregrosa (INRAE)

5.3.3. Suis-je propriétaire de mes données ?

Dans une interview réalisée le 02 juillet 2019, Lionel Maurel s'appuie sur ses compétences de juriste pour répondre à une série de questions, portant sur les droits liés aux données de recherche.
Ses réponses sont présentées dans ces courtes vidéos.

Voici trois exemples pour illustrer cette question :

- « As the data produced by GLACIOCLIM is funded by public funders, in accordance with the Law for a French Digital Republic (LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique), they are obligated to be open and accessible, and they are made available as soon as the processing is done (usually on an annual basis). Data is freely reusable, with the condition of acknowledging its authors according to the sentences indicated here below:
  "Les auteurs remercient le Service National d'Observation GLACIOCLIM (programme du CNRS-INSU, OSUG, IRD, INRAE, Météo France, IPEV) pour les données fournies."

"The authors thank the GLACIOCLIM National Observation Service (CNRS-INSU program, OSUG, IRD, INRAE, Météo France, IPEV) for providing the data.»
Exemple extrait du PGD public GLACIOCLIM, Daniel Arroyo, Isabelle Gouttevin (Météo-France), Delphine Six (OSUG)

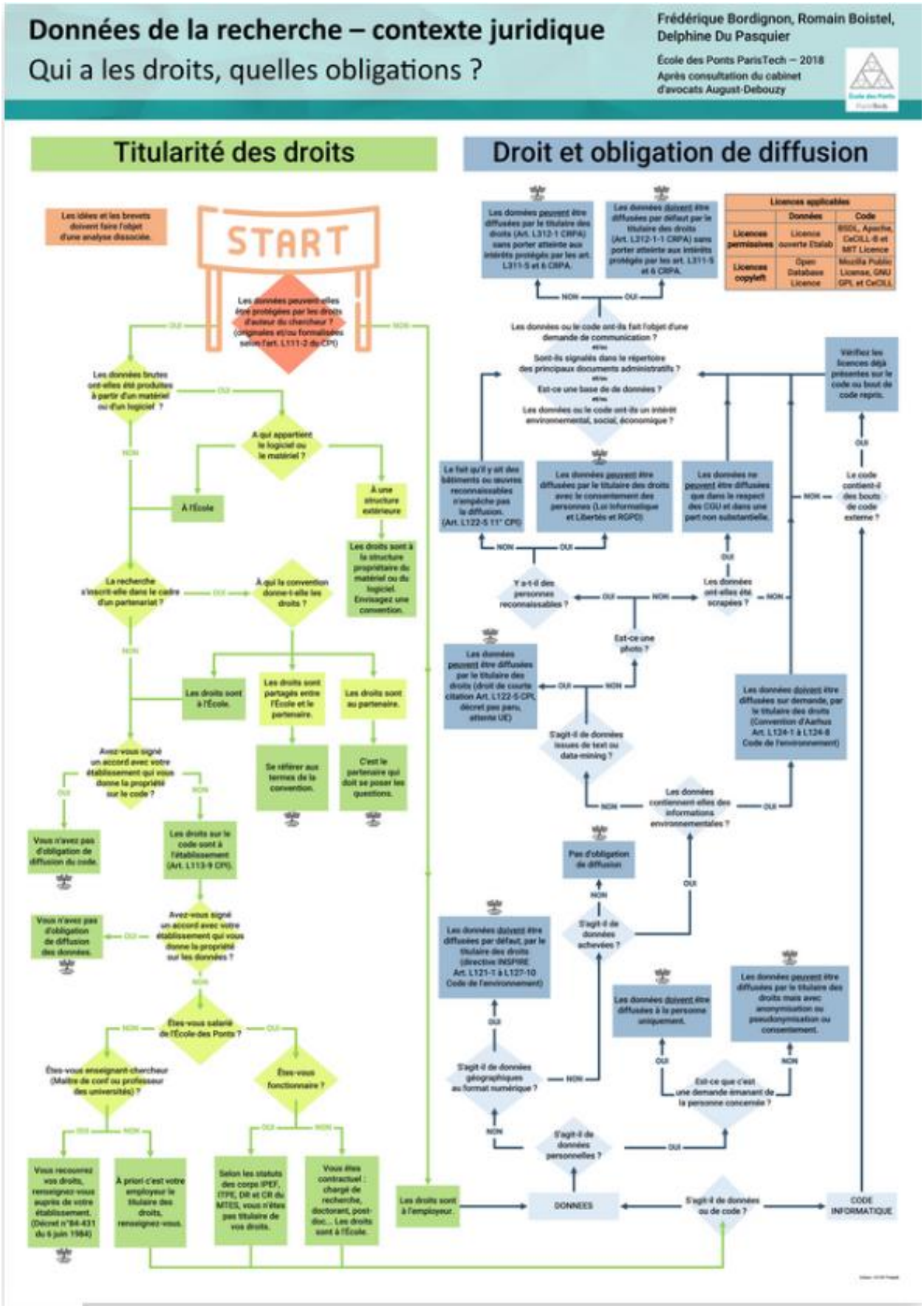
- « Chaque scientifique impliqué dans le projet respectera les règles de sa tutelle scientifique, en matière de propriété intellectuelle. Aucun brevet n'est envisagé sur la base du projet IMPRINT. Tous les scientifiques impliqués dans le projet IMPRINT seront associés aux publications en fonction de leurs contributions respectives. »
  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

- « Chaque partenaire est propriétaire de ses données (données fournies ou données produites). Les autres dispositions de propriété intellectuelle concernant les données produites dans le cadre du projet sont encore en cours de discussion pour l'établissement du contrat de consortium.
  L'AP-HP est propriétaire des données cliniques et aucune utilisation ou transmission à un tiers ne peut être effectuée sans son accord préalable. »
  Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

5.3.4. Dans le cadre d'un partenariat, à qui appartiennent les données ?

Cette infographie de l'École des Ponts ParisTech vous permettra de :

- Savoir à qui appartiennent les données de la recherche que vous traitez
- Qualifier si vos données sont diffusables
- Déterminer quand elles sont diffusables.



https://doranum.fr/aspects-juridiques-ethiques/qui-a-les-droits-quelles-obligations_10_13143_8dh5-d615/

Le 14 septembre 2020 les Tuto@Mate du Réseau Méthodes Analyses Terrains Enquêtes en SHS (InSHS) ont accueilli Lionel Maurel, juriste et Directeur Adjoint Scientifique de l'InSHS où il est en charge des questions d'information scientifique et technique.

Ce webinaire répond à une série d'interrogations telles que :

- Dans un projet de recherche avec plusieurs collaborateurs qu'ils soient chercheurs, ingénieurs, techniciens ou doctorants, membres d'un laboratoire ou d'un consortium **qui est propriétaire des données** ?
- Quand un prestataire met à la disposition d'une équipe de recherche un **ensemble de données** (entretiens, traitement d'archives, texte numérisé …), **qui en est propriétaire** ?
- Selon **quelles modalités** puis-je partager ou diffuser mes données ? …

Voici un exemple pour illustrer cette question :

> « Un accord de consortium sera conclu entre les partenaires pour définir les droits de propriété intellectuelle, les droits d'exploitation, sur les résultats du projet. Principes généraux : Les données antérieures restent la propriété du partenaire fournisseur (P1, P2, P3). Les données générées et les résultats obtenus par un seul partenaire appartiennent au partenaire qui les a générés. Les données générées et les résultats obtenus par plusieurs partenaires sont la propriété à parts égales des partenaires qui les ont générés. […]
>
> Le matériel végétal de l'IFV (5 variétés dans les WP2 et WP5) et de l'INRA (panel de 279 variétés du Centre de Ressources Biologiques de la Vigne de Vassal-Montpellier dans le WP3) sera utilisé dans le projet.
>
> L'accord de consortium précisera les droits d'utilisation du matériel végétal et les formalités à accomplir le cas échéant. »
>
> *Exemple extrait du PGD public G2WAS, Cédric Goby & Laurent Torregrosa (INRAE)*

5.3.5. Mes données sont-elles concernées par des problèmes d'éthique ou d'intégrité scientifique ?

Ces deux ressources contenant des extraits de MOOCs répondent à ces questions :

- Éthique de la recherche - Extraits d'un MOOC
- Intégrité scientifique - extraits d'un MOOC

Voici trois exemples pour illustrer ces aspects :

- « Le code d'éthique institutionnel de l'établissement dont relève le porteur de projet (EPHE) est appliqué. L'étude envisagée est conforme à l'agrément préfectoral de l'établissement pour ce qui concerne les espèces animales prélevées. Un comité d'éthique de scientifiques du CEFE a validé la nécessité scientifique du recours aux poissons vivants qui seront tous remis en liberté, ainsi que le choix de 5 espèces bio-indicatrices pour l'étude biométrique et le prélèvement de tissus, afin de mesurer l'impact des politiques mises en place sur le lac du Bourget pour réduire la pollution aquatique. Le comité a constaté que les principes de la bientraitance sont respectés et que les conditions d'utilisation des animaux sont optimisées (principe des 3R « Replace, reduce and refine ») compte tenu des nécessités expérimentales. Le responsable de campagne et le technicien en charge de la manipulation des poissons et du prélèvement de tissu sont habilités en expérimentation animale. Aucune étude préliminaire n'a été exigée puisque les protocoles expérimentaux proviennent du réseau des observatoires des Lacs de Montagnes Savoyards (OLMS) et sont couramment utilisés. »
  Exemple extrait du PGD public « PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

- « A propos de l'enquête 1, au niveau des ateliers, le groupe de recherche est vigilant sur le fait que le monde professionnel par région est relativement petit. Les professionnels risquent de se connaître.
  Nous suivons le code de conduite stipulé par la Charte nationale de déontologie des métiers de la recherche (2019) : https://comite-ethique.cnrs.fr/charte/ »
  Exemples extrait du PGD public PROSECO, Théa Manola (CNRS)

- « According to ethical rules defined by local (Zimbabwe and South Africa) and french regulations:
  - National Animal Research Ethics Council (NAREC) – Zimbabwe for animal handling and sampling, testing
  - Medical Research Council of Zimbabwe (MRCZ) – Zimbabwe for human sampling, testing and social study
  - University of Pretoria Research Ethics Council – South Africa
  France: the project will be submitted to the INRAE-CIRAD-IFREMER-IRD Consultative ethical committee. »
  Exemple extrait du PGD public VICTORIA, Hélène De Nys, Florian Liegeois (CIRAD et IRD)

Grâce à ces exemples concrets, vous voyez que chaque problématique qui relève des aspects juridiques et éthiques trouve réponse. N'hésitez pas à solliciter le DPO/DPD et le service juridique de votre institution si vous avez besoin d'aide.

5.4. Traitement et analyse des données

" La phase de **traitement des données** correspond au prétraitement des données brutes issues des acquisitions et des collectes. Il s'agit souvent de regrouper, choisir, qualifier les données pertinentes parmi celles qui ont été collectées, puis les reformater dans des formats standards interopérables, et les préparer en vue de leur analyse ultérieure. [...]
Derrière le terme "**analyser**" s'entend l'extraction de l'information des données le plus souvent par l'utilisation de puissance de calcul. Cela recouvre de nombreux types de techniques (calcul intensif, traitement statistique, machine learning, visualisation ...), et nécessite également des plateformes adaptées.
Cette étape du cycle de vie de nombreuses données impose que ces données soient exploitables, c'est-à-dire bien organisées, dans des formats adaptés à l'analyse envisagée, de façon à pouvoir leur appliquer des traitements automatisés. "

Source : Hadrossek Christine, Janik Joanna, Libes Maurice, Louvet Violaine, Quidoz Marie-Claude, Rivet Alain, Romier Geneviève. Guide de bonnes pratiques sur la gestion des données de la recherche. Version 2.0. 8 janvier 2023. https://mi-gt-donnees.pages.math.unistra.fr/guide/

C'est durant la phase d'analyse que l'on met souvent en place ce qu'on appelle un workflow, notamment dans les traitements informatiques.
Un workflow est une série structurée d'étapes pouvant être exécutées pour produire un résultat final, offrant aux utilisateurs un moyen de mettre en œuvre leur travail de manière plus reproductible.

Voici quatre exemples pour illustrer cette question :

- « Les workflows des principales données du SI sont ci-dessous :
  - pour la physico-chimie : Workflow-traitement-Physico-Chimie-OLA.png
  - pour la détermination et le comptage du zooplancton : Workflow-analyse-du-zooplancton.png

- pour l'insertion de ces données vers le SI OLA : Workflow-insertion-des-donnees-vers-le-SIOLA.png. »

Exemple extrait du PGD d'entité public OLA-Infrastructure, Ghislaine Monet (INRAE)

- « Les données microclimatiques seront déchargées au moins une fois par an (2021 et 2022), lors des campagnes de terrain, à l'aide des logiciels dédiés HOBOware et Lolly Manager puis analysées sous le logiciel libre R de traitements statistiques. »

  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

- « Les données d'origine instrumentale nécessitent des étapes de traitement pour être interprétables. Ceci couvre un ensemble vaste de processus qui sont sollicités en fonction de l'appareil utilisé pour générer la donnée initiale mais aussi des choix de stratégies de traitements en lien avec la question scientifique relié à l'échantillonnage. Ces processus génèrent des données intermédiaires et finalisées, qui nécessitent parfois la prise en compte de données préexistantes tiers caractérisant les échantillons analysés (exemple : détails de protocoles cliniques). »

  Exemple extrait du PGD public EQUIPEX MetaboHUB-METEX+, Franck Giacomoni et Fabien Jourdan (INRAE)

- « Pour les données biologiques

  Les analyses bioinformatiques et statistiques seront réalisées régulièrement. Les fichiers seront analysés par au moins deux investigateurs de façon indépendante. Des rapports réguliers seront produits qui contiendront différents indicateurs de traçabilité et degré de confiance des données et résultats obtenus.

  Du côté de BForCure, les algorithmes seront développés par le pôle Intelligence Artificielle (IA), et la mise à disposition sous forme de produit logiciel sera assuré par le pôle Software. Des analyses de données plus poussées pourront être faites par le pôle IA. »

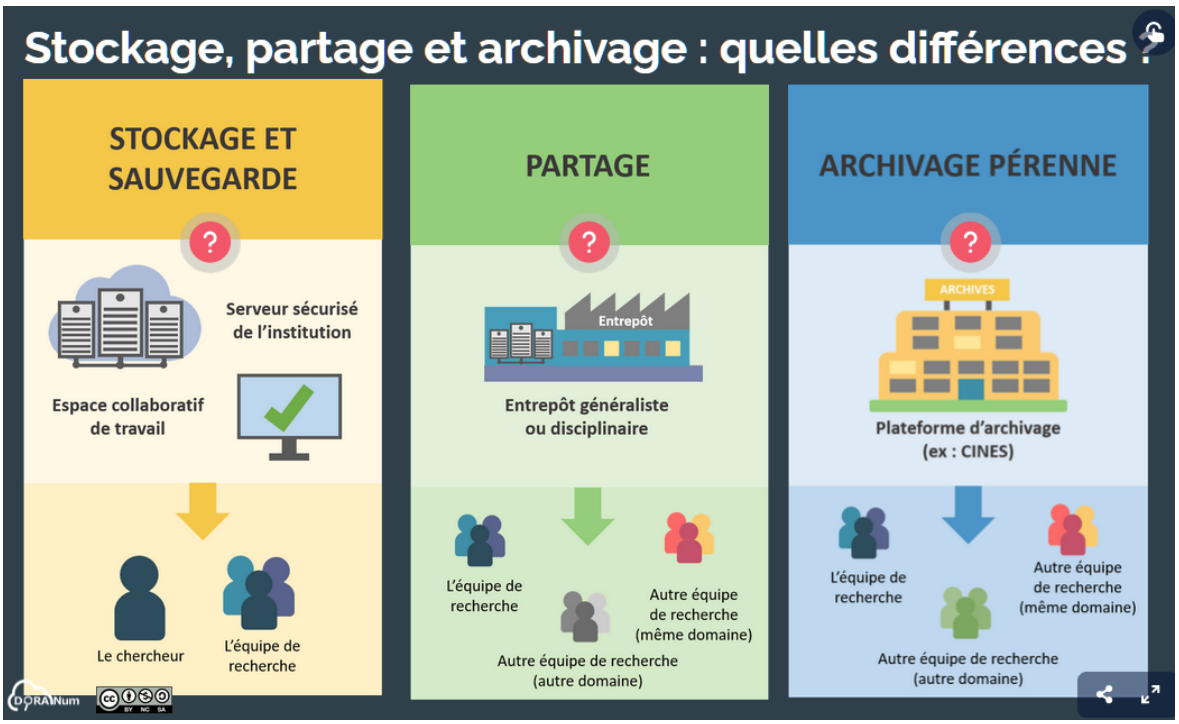
  Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

Grâce à ces exemples concrets, vous voyez différentes manières d'expliquer comment sont traitées et analysées les données.

5.5. Stockage, partage et archivage : quelles différences ?

Le stockage et la sauvegarde, le partage et l'archivage pérenne interviennent à différentes étapes du cycle de vie des données, et ont des fonctions distinctes.

Voici une ressource qui vous permettra de mieux comprendre la différence entre ces 3 étapes et de connaître les actions à effectuer afin de les mettre en pratique :



https://doranum.fr/stockage-archivage/stockage-partage-archivage-quelles-differences_10_13143_5dax-gp58/

5.6. Stockage et organisation des données

5.6.1. Histoires vécues

Visionnez ces vidéos d'1 min réalisées par l'EPFL :

- « RDM horror stories | Episode 1 - Lost Data »

https://www.youtube.com/watch?v=t_rEXpfCTrg&list=PLPkfOHxsjx2hH-QmfYp_ZHZI2WmE6pXLv&index=1

Traduction :

HISTOIRES D'ÉPOUVANTE SUR LA GESTION DES DONNÉES DE RECHERCHE

Épisode 1 - Données perdues

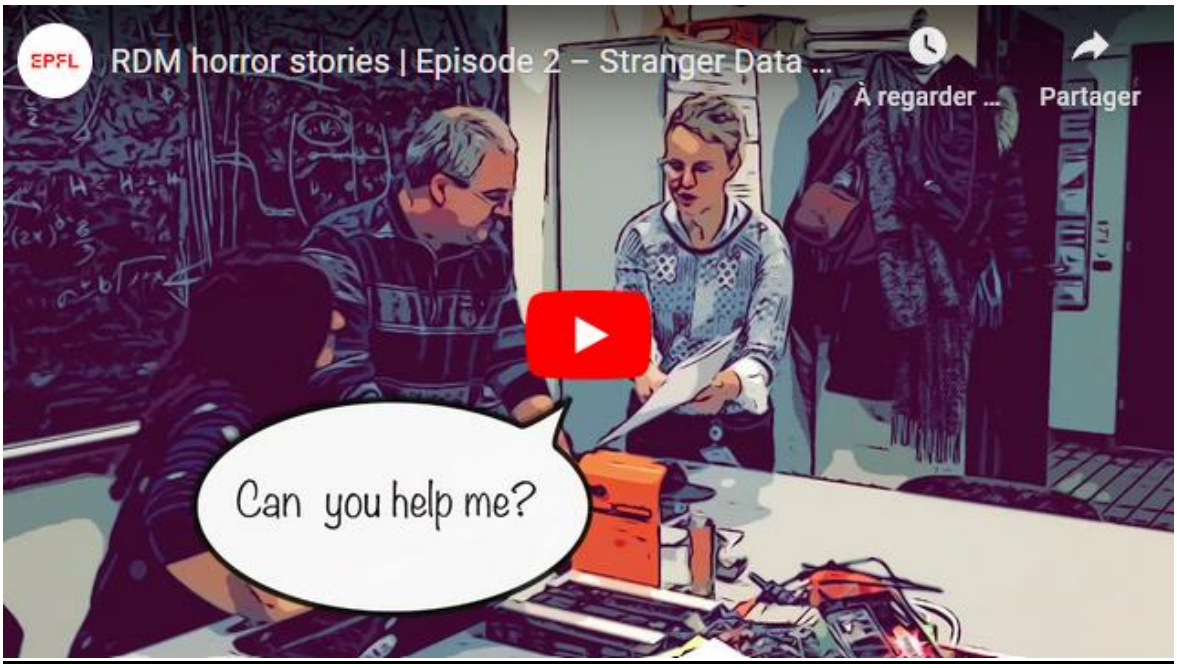
Il y a deux types de personnes dans le monde, celles qui font des sauvegardes et celles qui n'en font pas.

À quelle catégorie de personnes appartenez-vous ?

- Mon travail !!! J'ai perdu des mois de travail !!!

- C'est pas possible !!

- Vidéo RDM horror stories | Episode 2 – Stranger Data Things :



https://www.youtube.com/watch?v=tFWd2M2OXwQ&list=PLPkfOHxsjx2hH-QmfYp_ZHZI2WmE6pXLv&index=2

Traduction :

HISTOIRES D'ÉPOUVANTE SUR LA GESTION DES DONNÉES DE RECHERCHE

Épisode 2 - Le mystère des données

Il y a deux types de personnes dans le monde, celles qui ont rédigé un PGD et celles qui ne l'ont pas fait.

À quelle catégorie de personnes appartenez-vous ?

- Je recherche les données associées à ce schéma. Peux-tu m'aider ?

- Bien sûr

Sans PGD :

- Tu ne les as toujours pas trouvées ?

- Non, non. Où les as-tu sauvegardées ? C'est toi qui les avais.

- C'était toi !! Pas moi !!!

- C'était toi !!

Ces vidéos montrent l'intérêt de bien gérer ses données.

Des solutions sont proposées au niveau local :

- Un séminaire consacré au stockage des données de recherche a été organisé par la Cellule Data du site Grenoble Alpes, l'Inist-CNRS et l'URFIST de Lyon le 25 mai 2021.
  Cet évènement a été complété par des sessions locales ayant pour but de présenter les solutions de stockage mises en place au niveau local pour les communautés concernées. Les supports de présentation des sessions locales sont disponibles à la fin de la ressource.
- Dans le cadre de l'écosystème *Recherche Data Gouv*, les ateliers de la donnée sont en proximité géographique des équipes de recherche pour leur apporter une première expertise dans la gestion raisonnée des données de recherche.

5.6.2. Où dois-je stocker mes données ?

Pour ne pas perdre vos données, un des principes est de dupliquer et stocker les données à différents endroits sur différents supports selon une temporalité pertinente pour le projet.

L'idéal est d'appliquer la règle du 3-2-1, ce qui veut dire :

- Garder 3 exemplaires des données,
- Sur 2 supports ou technologies différents (les tester régulièrement et migrer les fichiers sur un autre support si nécessaire),
- Dont 1 se trouve hors site.

Cette vidéo de 2 min détaille la règle du 3-2-1 :

https://www.canal-u.tv/chaines/callisto/la-sauvegarde-3-2-1

Voici trois exemples pour illustrer cette question :

- « Les données seront sauvegardés sous les trois supports :
  - Carnet de terrain papier : sauvegarde dans des armoires ignifugées du CEFE
  - Tableurs : sauvegarde sur l'ordinateur portable terrain puis transmission par 4G d'une copie sur le serveur de stockage du CEFE (sauvegardes automatiques quotidiennes et contrôle des sauvegardes assuré par le service informatique du CEFE)
  - Base de données : sauvegarde selon le principe 3-2-1 (3 sauvegardes sur 2 supports différents (SQL/ binaire) avec 1 lieu déporté (sur le campus de l'université de Montpellier). »

  Exemple extrait du PGD public « PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

- « En cas d'incident, les données et métadonnées les moins volumineuses pourront être récupérées facilement à partir de la plateforme de sauvegarde et de partage sécurisé du CNRS (myCoRe). Pour ce qui est des données LiDAR, plus volumineuses, elles pourront être récupérées à partir de la sauvegarde effectuée sur le NAS de l'unité de recherche EDYSAN (UMR 7058 CNRS). Nous envisageons également l'enregistrement des données LiDAR sur un cloud avec un espace de stockage conséquent et en accord avec la plateforme MATRICS de l'Université de Picardie Jules Verne (UPJV). »

  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

- « Les eCRF Cleanweb sont stockés dans un serveur de l'APHP certifié hébergeur de données de santé.

  Les données biologiques seront stockées dans trois disques durs externes conservés dans des endroits distincts à accès contrôlé et dans un serveur dédié à la sauvegarde de l'Institut Cochin. Les données génétiques seront stockées sous forme de fichier .txt (Bam et fastq) dans trois disques durs conservés dans des lieux différents et protégés par un code d'accès et pourront être transférées en fin d'étude à une organisation européenne (European Genome Archive) pour archivage dans l'intérêt public, à des fins de recherche scientifique, sous réserve que les patientes ne s'y soient pas opposées après en avoir été informées.

  L'hébergement de GAIA est assuré par un hébergeur certifié « hébergeur de données de santé », avec une localisation des données assurées en Union Européenne, à travers un entrepôt de données de type SQL. Une sauvegarde est effectuée toutes les nuits, avec un format agnostique du moteur de base de données (JSON). La méthode utilisée pour

le stockage permet de revenir en arrière à tout moment et de lister les actions effectuées. »
Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

5.6.3. Comment optimiser la sauvegarde de mes données ?

Voici deux exemples pour illustrer cette question :

- « Des sauvegardes (backup) régulières (tous les 6 mois au moins) de l'ensemble des données et métadonnées du projet (inclues les données LiDAR) seront également effectuées et enregistrées sur le serveur NAS de l'unité de recherche EDYSAN (UMR 7058 CNRS). A noter également que les données LiDAR seront partagées avec chaque agence territoriale de l'ONF concernée par le projet (FD de l'Aigoual, FD de Blois et FD de Mormal). Cela assurera une sauvegarde supplémentaire. »
  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

- « Concernant la gestion du versioning, nous utiliserons un logiciel de gestion des versions : Git. Git est un logiciel libre. »
  Exemple extrait du PGD public Hospitam, Lauriane Locatelli (ENS Lyon)

5.6.4. Comment stocker de manière sécurisée les données personnelles / sensibles ?

Voici un exemple pour illustrer cette question :

- « Données cliniques :
  Dans l'interface CleanWeb standard, les données sensibles peuvent être recueillies au niveau de variables spéciales de type " DONNÉE PERSONNELLE " ; elles sont alors :
  - stockées de manière cryptée dans une base séparée de la base étude accessibles de manière non-cryptée seulement sous certaines conditions (pour les utilisateurs du centre auquel est rattaché le patient)
  - non-exportables
  - non-utilisables comme élément de la référence patient.
  Les données personnelles de type " Nom " et " Prénom " seront affichées sur la liste des patients par défaut et le champ de recherche filtrera sur la référence ainsi que sur les colonnes de type " Nom " et " Prénom " si l'utilisateur peut les visualiser.

Dans l'interface CleanWeb ePRO, les données de contact (email ou téléphone) du patient :
- ne font pas partie du cahier d'observation, mais ont un accès assujetti à un droit spécifique
- sont stockées dans la base étude dans deux colonnes séparées
- sont cryptées en base, et lors de la connexion
- sont non-exportables, non-imprimables, visibles et éditables uniquement au niveau du calendrier patient
- sont supprimées de manière définitive à la fin de l'étude, lorsque le statut de l'étude est coché " Terminée "
- peuvent être effacées par patient au fur et à mesure de l'avancement de l'étude si besoin. »

Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

5.6.5. Comment partager les données avec les partenaires pour faciliter la collaboration ?

Vous pouvez les partager dans un espace collaboratif de travail institutionnel avec accès sécurisé. Il est recommandé d'éviter au maximum les outils du type One Drive, Google Drive, Dropbox, etc. N'hésitez pas à vous rapprocher de votre établissement afin de connaitre les espaces de stockage sécurisés mis à disposition.

Voici deux exemples pour illustrer cette question :

- « Les données destinées à être partagées entre membres du programme sont stockées sur la AMUbox. AMUbox est une solution de stockage de type cloud, qui s'appuie sur la solution technique Nextcloud. Celui-ci en assure la sécurité numérique, l'optimisation et la fiabilité dans le temps. Les données sont stockées et sécurisées au sein du datacenter de AMU. L'accès à la Box se fait sur invitation. Il est strictement réservé aux membres du programme. Ce service est mis à disposition de ses chercheurs par l'université d'Aix-Marseille. »

  Exemple extrait du PGD public Transfunéraire, Clara Duterme et Elisabeth Anstett (CNRS et Unistra)

- « Pour les données les moins volumineuses (données microclimatiques et les données issues des inventaires dendrométriques, floristiques et faunistiques), ainsi que toutes les

informations relatives au projet, et ce pour faciliter l'échange d'information entre les membres du projet, il est prévu d'utiliser un dossier partagé et intitulé "IMPRINT" sur la plateforme de sauvegarde et de partage sécurisé du CNRS : myCoRe. Ce dossier partagé entre les scientifiques impliqués dans le projet a déjà été créé. »
Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

5.6.6. Pendant combien de temps dois-je stocker mes données ?

En général, les données doivent être stockées au moins durant toute la durée du projet.

Voici un exemple pour illustrer cette question :

- « Les données de planification et d'organisation des campagnes, d'observations et de mesures et d'analyses en laboratoire seront sauvegardées sur les serveurs du CEFE pour la durée du projet et pendant 3 années supplémentaires. A l'expiration de cette échéance, seuls les carnets de terrain papier, les cahiers de laboratoire papier et la base de données seront conservés. La pérennisation de la base de données sera assurée dans le cadre des missions de la plateforme SIE du CEFE. Les données génétiques déposées dans Genbank seront préservées dans le cadre des missions du NCBI (https://www.ncbi.nlm.nih.gov/genbank/). »
  Exemple extrait du PGD public « PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

5.6.7. J'ai beaucoup de données. Comment faire pour m'y retrouver ?

Bien organiser ses données est essentiel pour pouvoir les retrouver, les partager et les réutiliser.

Ce guide peut vous aider à mieux organiser vos données et nommer vos fichiers :

https://doranum.fr/stockage-archivage/conseils-pour-lorganisation-des-donnees_10_13143_h6gx-e249/

Voici un exemple pour illustrer cette question :

- « Un répertoire spécifique nommé "Terrain" recensera tous les protocoles de collecte des données, les notices d'utilisation et de calibrage des appareils de mesure et les images d'aide à l'identification des espèces de poissons. Il contiendra aussi le principe établi d'organisation et de codage des répertoires et des fichiers.
  Un répertoire spécifique nommé "Laboratoire" recensera tous les protocoles, les notices des appareils, les résultats expérimentaux et d'analyses bio-informatiques et statistiques. Un répertoire "Données Terrain" sera créé pour chaque campagne. Il sera organisé conformément aux éléments mis dans le document d'organisation et de codage des répertoires et des fichiers.
  Un répertoire "Données Laboratoire" sera créé pour chaque campagne. Il sera organisé conformément aux éléments mis dans le document d'organisation et de codage des répertoires et des fichiers. »
  Exemple extrait du PGD public « PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

5.6.8. Que faire des données sous différents formats ?

Il est recommandé de privilégier les formats ouverts mais saurez-vous les reconnaître dans ce jeu ?

https://view.genial.ly/5f7f1acda093260da6f58b8c/game-ouvert-ou-ferme

Voici deux exemples pour illustrer cette question :

- « […] Les protocoles expérimentaux sont conservés sous forme papier (dans les cahiers de laboratoire) et version électroniques (.docx, .pdf). Les données RMN seront stockées sous forme électronique (fichiers FID bruts et format PDF) et en laboratoire sous forme de copies papier des spectres. […] Les données biochimiques générées seront : des données expérimentales (formats csv, xlsx), des données textuelles (format pdf, docx, pptx), des données de protéomique, des données cristallographiques, des images (microscopie confocale et électronique – format png). Les données protéomiques seront de plusieurs types, les fichiers bruts en sortie des spectromètres de masse (.raw), les données traitées (.txt, .xml, .mztab) et les résultats (.txt, .xlsx, .docx et images). Le volume des données protéomiques qui seront générés pendant ce projet est estimé à plusieurs Go. Les données de diffraction au rayons X seront stockées sous forme électronique (format .cbf et .h5) à la fois par les sites de rayonnement synchrotron (Soleil et ESRF), et au sein du laboratoire du partenaire du projet. […]
  Dans la mesure du possible des formats standards et ouverts seront privilégiés à des fins de partage et de réutilisation. »
  Exemple extrait du PGD public LipInTB, Jean-François Cavalier (CNRS)

- « […] Les données relatives à un échantillon analysé vont être collectées sous un fichier (ou un ensemble de fichiers) au format défini par le constructeur de l'instrument (format propriétaire). Une fois la donnée instrumentale obtenue, des solutions logicielles existent pour la convertir en un format standardisé ouvert (exemple : nmrML ou MzML). »
  Exemple extrait du PGD public EQUIPEX MetaboHUB-METEX+, Franck Giacomoni et Fabien Jourdan (INRAE)

Grâce à ces exemples concrets, vous voyez qu'il est relativement facile d'expliquer comment sont stockées, sauvegardées, organisées les données et sous quels formats.

5.7. Partage des données

5.7.1. Histoire vécue

Visionnez cette vidéo d'1 min réalisée par l'EPFL :
« RDM horror stories | Episode 5 – Data PubliCaution »

https://www.youtube.com/watch?v=NdkIWkRi-ZQ&list=PLPkfOHxsjx2hH-QmfYp_ZHZI2WmE6pXLv&index=5

Traduction :
HISTOIRES D'ÉPOUVANTE SUR LA GESTION DES DONNÉES DE RECHERCHE
Épisode 5 - Public-attention aux données

Il y a deux types de personnes dans le monde, celles qui savent comment publier leurs données et celles qui ne le savent pas.
À quelle catégorie de personnes appartenez-vous ?
- Super, mon papier est accepté !!!
- FAIR ? Gestion des Données de Recherche ? Entrepôt de données ?
- Je connais pas et je m'en moque !!!
- Je vais publier mes données TRÈS FACILEMENT !!!
- FAIR… Comment on fait ?
- Gestion des Données de Recherche… Quoi ???
- Entrepôt de données… Lequel ?

Pour y voir plus clair, consultez cette vidéo de 4 min sur le dépôt des données :

https://www.canal-u.tv/chaines/callisto/deposer-ses-donnees-de-recherche-pourquoi-quoi-quand-ou-et-comment

5.7.2. Suis-je obligé de partager mes données ?

- Dans certaines disciplines, le partage est obligatoire : Environnement, Géographie
- Dans d'autres cas, il y a des exceptions au partage : données industrielles, secret défense, secret professionnel, confidentielles, données personnelles ou protégées par le droit d'auteur…

Pour plus d'informations, consultez la ressource " Les données de la recherche et les codes sources obligatoirement diffusables ".

Vous pouvez également utiliser l'outil d'aide à la décision sur la diffusion des données de recherche du Cirad.
Voici deux exemples pour illustrer cette question :

- « Pour une éventuelle exploitation commerciale par l'IFV partenaire, les données de génotypage et de phénotypage pour les 5 variétés de l'IFV ne seront pas partagées à la fin du projet, sauf accord contraire de l'IFV. Les autres données seront partagées selon les règles définies dans l'accord de consortium. »
  Exemple extrait du PGD public G2WAS, Cédric Goby et Laurent Torregrosa (INRAE)

- « Les données produites par les observatoires du SNO KARST étant financées par des subventions publiques, selon la Loi pour une République Numérique, elles ont l'obligation d'être ouvertes et accessibles. Ces données sont donc ouvertes dès que possible, et aucune authentification n'est nécessaire pour accéder aux données publiques. »
  Exemple extrait du PGD public SNO KARST, Juliette Fabre et Hervé Jourde (CNRS)

5.7.3. Où et comment partager mes données ?

Pour choisir le bon entrepôt, appuyez-vous sur les **pratiques de votre communauté scientifique** !

1) Entrepôt disciplinaire

Certaines disciplines sont bien organisées pour la gestion des données, et proposent des entrepôts disciplinaires spécifiques. Vous pourrez ainsi vous appuyer sur un ensemble de bonnes pratiques et de standards bien définis, ce qui facilitera grandement la préparation, la documentation et le dépôt des données.

Si aucun entrepôt n'est recommandé par votre communauté, vous pouvez identifier celui qui pourrait convenir à vos besoins grâce aux annuaires dédiés comme re3data, OAD, OpenDOAR, FAIRsharing...

2) Entrepôt institutionnel

Si aucun entrepôt disciplinaire ne convient, il est conseillé de déposer vos données dans l'entrepôt de votre institution, s'il existe.

3) Recherche Data Gouv

Si aucun entrepôt disciplinaire ou institutionnel ne correspond à vos besoins, il est recommandé de déposer dans le nouvel entrepôt national pluridisciplinaire *Recherche Data Gouv*. Il permet à la communauté scientifique française de déposer et d'ouvrir ses données de recherche.
Développé à partir de l'application web open source Dataverse, l'entrepôt *Recherche Data Gouv* est organisé en espaces institutionnels de publication et de signalement des données des établissements qui souhaitent participer.

Vous pouvez vous aider des critères suivants pour choisir votre entrepôt :

https://doranum.fr/depot-entrepots/criteres-pour-choisir-entrepot_de_donnees_10_13143_zgpb-9449/

Voici trois exemples pour illustrer cette question :

- « Les données seront partagées à la fin du programme de recherche sur la plateforme Didomena, l'entrepôt de données de recherche de l'EHESS. (https://didomena.ehess.fr). Cette plateforme est dédiée aux sciences sociales et permet de partager et valoriser les

données de recherche. Il n'existe pas de raison motivant un embargo, les données partagées seront immédiatement accessibles. [...] »
Exemple extrait du PGD public Transfunéraire, Clara Duterme et Elisabeth Anstett (CNRS et Unistra)

- « Toutes les structures protéiques seront librement disponibles car déposées dans RCSB Protein Data Bank (https://www.rcsb.org/).
  Les données de protéomique seront également accessibles via le Consortium PRIDE : PRoteomics IDEntifications database (http://www.proteomexchange.org/). »
  Exemple extrait du PGD public LipInTB, Jean-François Cavalier (CNRS)

- « Next state
  Data will be accessible on the Theia/OZCAR web portal: https://in-situ.theia-land.fr/
  The data download service will be available in the near future. Users will have the option to download data in CSV and NetCDF formats. Data downloading will require user authentication through Data Terra Single Sign-On authentication, ensuring adherence to embargoes and access restrictions for certain data. Authentication will also grant access to authenticated data producers for statistics on data downloads.
  The data will also be indexed in a Geonetwork metadata catalog, allowing for automatic harvesting of the data catalog through the CSW webservice. »
  Exemple extrait du PGD public GLACIOCLIM, Daniel Arroyo, Isabelle Gouttevin (Météo-France), Delphine Six (OSUG)

5.7.4. Comment faire en sorte que d'autres chercheurs puissent lire et traiter mes données sans soucis ?

Si besoin, l'idéal est de déposer les codes sources utiles à l'analyse et au traitement de vos données de recherche dans un entrepôt dédié. Il existe notamment l'archive universelle Software Heritage qui permet de partager et conserver les codes sources de manière pérenne.

Pour en savoir plus, consultez la ressource ci-dessous :

https://doranum.fr/stockage-archivage/les-codes-sources-definition-enjeux-et-preservation_10_13143_7tj2-gw58/

Voici un exemple pour illustrer cette question :

- « Les données utilisées pour la publication des résultats seront également publiées en libre accès au moment de la publication de l'article, ainsi que les scripts (lignes de codes) utilisés pour générer les résultats à partir des données. Ceci dans le but d'assurer la reproductibilité des résultats publiés. »
  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

5.7.5. Comment faire en sorte que mes données soient localisables dans le temps ?

Afin de garantir l'Accessibilité à vos données (selon le " A " des principes FAIR), l'un des éléments importants est de leur attribuer un identifiant unique et pérenne.

Pour en savoir plus sur les identifiants pérennes, consultez ces deux ressources :

https://doranum.fr/identifiants-perennes-pid/identifiants-perennes-apercu_10_13143_t427-f432/

https://doranum.fr/identifiants-perennes-pid/zoom-doi_10_13143_j5xt-6j41/

En ce qui concerne les logiciels (codes sources), il existe un identifiant spécifique généré par l'archive universelle Software Heritage :

https://doranum.fr/identifiants-perennes-pid/zoom-swhid_10_13143_3qqg-yx41/

Voici un exemple pour illustrer cette question :

- « For glaciological data, field reports are produced annually for each site, and is currently shared with the community through the GLACIOCLIM website and access to data files is available through the Nextcloud platform of OSUG. Part of the data is also sent annually to the World Glacier Monitoring Service, with verifications performed by national correspondents. Mass balance (Terre Adelie) data, is integrated into the QGIS Quantarctica distribution, an Geographical Information System (GIS) visualization software.
  At the snow sites, validated data are produced once a year, leading to an update of the distributed files under the dois, and can be downloaded at:
  Col de Porte: https://doi.org/10.17178/CRYOBSCLIM.CDP.2018
  Col du Lac Blanc: https://doi.org/10.17178/CRYOBSCLIM.CLB.all. »
  Exemple extrait du PGD public GLACIOCLIM, Daniel Arroyo, Isabelle Gouttevin (Météo-France), Delphine Six (OSUG)

5.7.6. Comment associer mon identité de chercheur aux données ?

Dans le PGD, les contributeurs peuvent indiquer leur ORCID dans la partie administrative.

Cette vidéo de 2 min 30 montre l'intérêt de l'identifiant ORCID :

https://www.youtube.com/watch?v=LBh-D9Wx0k8&t=7s

Cette deuxième ressource explique ce qu'est un ORCID et comment l'utiliser :

https://doranum.fr/identifiants-perennes-pid/zoom-orcid_10_13143_c6rx-9w77/

5.7.7. Comment bien définir les conditions pour que d'autres chercheurs puissent réutiliser mes données ?

Attribuer une licence à vos données est très important car cela permet de bien définir les modalités de réutilisation et de les afficher clairement.

Pour en savoir plus, consultez le guide ci-dessous, ainsi que cette ressource : Questions juridiques liées aux données de recherche.

https://doranum.fr/aspects-juridiques-ethiques/guide-des-licences-ouvertes_10_13143_tv6f-sv31/

Voici trois exemples pour illustrer cette question :

- « Les données sont diffusées sous Licence Ouverte 2.0, compatible avec la Licence Creative Commons Attribution 4.0 (CC BY 4.0) : les données sont librement réutilisables, sous condition de citer leurs auteurs.
  - Par ailleurs, en cas d'utilisation des données du SNO KARST, il est demandé :
  - de contacter la personne ressource du site d'observation, de préférence dans le cadre d'une collaboration de recherche ;
  - de fournir à la personne ressource du site d'observation l'opportunité de donner leur expertise sur les résultats obtenus à partir de l'utilisation du jeu de données ;
  - si l'utilisation de ces données donne lieu à une publication, de communiquer le DOI et de remercier le SNO KARST via une phrase type. »

  Exemple extrait du PGD public SNO KARST, Juliette Fabre et Hervé Jourde (CNRS)

- « Les données d'observation, traitées et achevées seront librement accessibles (licence CC-BY libre d'utilisation à la condition de l'attribuer à l'auteur en citant son nom). […] L'ensemble des données collectées dans le cadre de cette ANR (01/05/2020 - 30/04/2023) seront disponibles en libre accès (licence CC-BY) après l'embargo de publication finale des données traitées. »

  Exemple extrait du PGD public « PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

- « Licenses
  Glacier data licence: CC-BY-NC
  Snow data licence: CC-BY
  Meteorological data at snow sites: CC-BY. »

  Exemple extrait du PGD public GLACIOCLIM, Daniel Arroyo, Isabelle Gouttevin (Météo-France), Delphine Six (OSUG)

Grâce à ces exemples concrets, vous voyez qu'il est relativement facile d'indiquer comment et où vous allez partager vos données.

5.8. Archivage pérenne des données

Cette vidéo de 2 min vous aidera à mieux comprendre ce qu'est l'archivage pérenne des données et quels sont les acteurs susceptibles d'intervenir à cette étape :

https://www.canal-u.tv/chaines/callisto/les-minutes-doranum/la-minute-archivage-perenne-des-donnees

Pour aller plus loin, vous pouvez également visionner les vidéos des interventions captées lors de la journée « Archivage Numérique des Données de Recherche » organisée le 20 novembre 2019 par l'UMS GRICAD (Grenoble Alpes Recherche – Infrastructure de Calcul Intensif et de Données) et le SARI (Réseau des Informaticiens du Sillon Alpin) : https://videos.univ-grenoble-alpes.fr/recherche/archivage-numerique-des-donnees-de-recherche/

5.8.1. Si je partage mes données dans un entrepôt, cela veut-il dire qu'elles sont archivées ?

La plupart des entrepôts de données ne permettent pas un archivage pérenne.

Dans l'**annuaire re3data** vous pouvez trouver des entrepôts certifiés qui s'engagent à proposer un archivage à long terme. C'est notamment le cas de 4TU.ResearchData.

Si l'entrepôt choisi ne permet pas l'archivage à long terme, il faut d'abord réfléchir à quelles données vous allez sélectionner pour un archivage pérenne. Le Ministère de

l'Enseignement Supérieur et de la Recherche a mandaté le CINES pour vous accompagner dans cette étape d'archivage.

Si vous êtes dans un domaine des SHS, l'infrastructure Huma-Num peut également vous accompagner pour l'archivage de vos données.

Voici deux exemples pour illustrer cette question :

- « L'archivage pérenne des données sera envisagé sur le long terme, en s'appuyant sur les centres de données et les offres de service du moment (exemple : offre d'archivage du CINES). Seules les données produites par les observatoires et les partenaires du SNO KARST seront préservées. Les séries calculées simplement à partir d'autres séries (ex : moyenne journalière à partir de données horaires) pourront ne pas être archivées. On pourra également sélectionner seulement certains niveaux de données (ex : Raw data, ou Quality-controlled data). »
  Exemple extrait du PGD public SNO KARST, Juliette Fabre et Hervé Jourde (CNRS)

- « L'archivage de données au Cines est prévu à condition d'obtenir des financements permettant de couvrir les frais de l'archivage pérenne. »
  Exemple extrait du PGD public " Prospection du territoire d'Amathonte ", Anna Cannavò (CNRS)

5.8.2. Cela a-t-il un intérêt d'archiver toutes mes données ?

L'archivage pérenne ne concerne en général qu'une partie des données produites par un projet. Pour certains projets, il n'est d'ailleurs pas nécessaire de prévoir d'archivage pérenne.
En effet, la question de l'archivage pérenne se pose uniquement pour les données présentant une valeur scientifique reconnue par la communauté d'où elles proviennent et qui nécessitent une conservation pour au moins 30 ans.

C'est une opération coûteuse qui nécessite un budget alloué. Elle se décide à l'échelle du laboratoire ou de l'institution et non pas à l'échelle du chercheur.

Concrètement, l'archivage numérique pérenne consiste à conserver le document et l'information qu'il contient :

- Dans son aspect physique comme dans son aspect intellectuel,
- Sur le très long terme,
- De manière à ce qu'il soit en permanence accessible et compréhensible.

Voici trois exemples pour illustrer cette question :

- « Le choix des données à conserver sur le long terme se fera en concertation commune entre les membres du programme ; la décision finale appartiendra aux deux coordinatrices du Programme, en fonction du caractère inédit des données, de leur valeur et de leur pertinence pour des démarches comparatives futures. […] »
  Exemple extrait du PGD public Transfunéraire, Clara Duterme et Elisabeth Anstett (CNRS et Unistra)

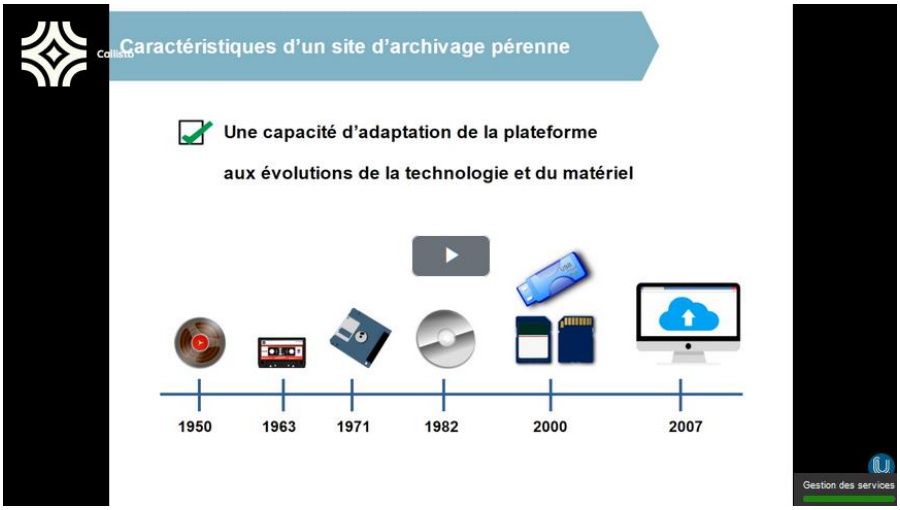
- « La plupart des données générées par l'UEFP sont des données brutes non reproductibles et sont donc conservées sur le long terme sans date limite.
  Nous avons besoin de conserver nos données sur des dizaines d'années, c'est-à-dire bien au-delà de la révolution des écosystèmes forestiers sur lesquels l'UEFP travaille. »
  Exemple extrait du PGD d'entité public UEFP, Laetitia Deyris et Frédéric Bernier (INRAE)

- « Due to the climatic purposes of GLACIOCLIM, all data must be stored for the long term. These data serve as relevant climate indicators in high-latitude, mid-altitude and high-altitude areas. All observations are considered as Essential Climate Variables by international organizations such as World Meteorological Organization, that means variables or a group of linked variables that critically contributes to the characterization of Earth's climate. The preservation of data is planned on platforms such as OSUG data center, INREA servers and CNRM servers, which serve as data repository.
  At the current time, there is no long-term preservation solution. However, all necessary technological guarantees and measures are in place to ensure at least medium-term storage. »
  Exemple extrait du PGD public GLACIOCLIM, Daniel Arroyo, Isabelle Gouttevin (Météo-France), Delphine Six (OSUG)

5.8.3. Où puis-je archiver mes données pour qu'elles soient encore consultables dans 30 ans ?

C'est le rôle d'une plateforme d'archivage pérenne de garantir la conservation de vos données sur le très long terme.

Cette vidéo d'1 min explique ce qu'est un site d'archivage pérenne et ses caractéristiques :



https://www.canal-u.tv/chaines/callisto/caracteristiques-d-un-site-d-archivage-perenne

Le CINES (Centre Informatique National de l'Enseignement Supérieur) est l'opérateur mandaté par le Ministère pour opérer la mission d'archivage pérenne pour l'Enseignement Supérieur et la Recherche. Il vous guide dans votre démarche d'archivage pérenne.
Selon son institution, sa discipline ou l'entrepôt choisi, il existe déjà des partenariats avec le CINES, proposant un accompagnement pour l'archivage (notamment Huma-Num).

Pour vérifier la compatibilité et la pérennité des formats de fichiers, le CINES propose l'outil FACILE.

Voici deux exemples pour illustrer cette question :

- « L'archivage pérenne des données sera envisagé sur le long terme, en s'appuyant sur les centres de données et les offres de service du moment (exemple : offre d'archivage du CINES). Seules les données produites par les observatoires et les partenaires du SNO KARST seront préservées. Les séries calculées simplement à partir d'autres séries (ex : moyenne journalière à partir de données horaires) pourront ne pas être archivées. On pourra également sélectionner seulement certains niveaux de données (ex : Raw data, ou Quality-controlled data). »
  Exemple extrait du PGD public SNO KARST, Juliette Fabre et Hervé Jourde (CNRS)

- « Pour l'archivage à long terme, la TGIR [Huma-Num] propose un service en lien direct avec la Plateforme d'Archivage au CINES (PAC). Cette démarche sera automatisée au maximum par l'ensemble des normes de création, de nommage et de validation des données produites par le projet. »
  Exemple extrait du PGD public « Techniques artisanales au premier âge du Fer en Italie. Repenser les interactions culturelles », Veronica Cicolani (CNRS)

Grâce à ces exemples concrets, vous voyez que seuls certains jeux de données nécessitent un archivage à long terme.

6. Coûts de gestion des données

6.1. Histoire vécue

Visionnez cette vidéo d'1 min réalisée par l'EPFL :
« RDM horror stories | Episode 4 – Data Money Freaks »

https://www.youtube.com/watch?v=LCZijZP916o

Traduction :
HISTOIRES D'ÉPOUVANTE SUR LA GESTION DES DONNÉES DE RECHERCHE
Épisode 4 -Le fric des données rend fou
Il y a deux types de personnes dans le monde, celles qui gèrent un budget et celles qui ne le font pas.
À quelle catégorie de personnes appartenez-vous ?
- C'est la fin de notre projet. Ça se fête !!!
- C'est tout ce qui reste ?
- Comment as-tu dépensé l'argent ?

6.2. Le stockage des données a-t-il un coût ?

Voici des exemples de ce qui peut générer un coût lors de l'étape du stockage :
Les serveurs
Les différentes sauvegardes
La sécurité des installations
Les espaces collaboratifs de travail
...
Voici un exemple pour illustrer cette question :

- « Storage funding and associated costs are assured for 7 years (2020-2026) if the data volume remains within the expected bound at the time of the present DMP. Later, or if the data volume explodes, additional funding will have to be found. Since the infrastructure is included in an OSU and the project belongs to a national initiative, it should be possible to obtain this funding [...]. »
  Exemple extrait du PGD public "Theia/OZCAR Information System", Sylvie Galle (IRD)

6.3. Le partage des données dans un entrepôt a-t-il un coût ?

Le partage des données dans un entrepôt est souvent gratuit. Cependant, le coût peut varier en fonction de la volumétrie. Par exemple pour 4TU Research data, les dépôts sont gratuits jusqu'à 10 Go par an, payant au-delà (€ 4.50 per GB).

6.4. L’archivage pérenne des données a-t-il un coût ?

L'archivage pérenne a un coût non négligeable. Les services d'archivage à long terme proposés par le CINES sont facturés en fonction de la nature de l'établissement déposant et du volume des données à archiver.
C'est la raison pour laquelle il est important de bien sélectionner les données à archiver.

6.5. Comment puis-je évaluer les ressources humaines nécessaires à mon projet ?

Voici deux exemples pour illustrer cette question :

- « La plateforme SI de l'OSU OREME est responsable de la gestion FAIR des données du SNO KARST. Elle y consacre le temps nécessaire, en fonction des autres contraintes de la plateforme. Elle dispose de 2 ingénieurs dédiés à la gestion de l’ensemble des données d’observation dont elle a la charge. Elle possède sa propre infrastructure matérielle hébergée à Géosciences Montpellier et sur le campus CNRS de la Délégation Régionale Occitanie-Est. »
  Exemple extrait du PGD public SNO KARST, Juliette Fabre et Hervé Jourde (CNRS)

- « En termes de ressources humaines, le recrutement de la doctorante à 100% sur le projet IMPRINT […] permettra d'assurer que les données générées par le projet IMPRINT seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable). »
  Exemple extrait du PGD public IMPRINT, Jonathan Lenoir (CNRS)

6.6. Exemples synthétiques de coûts liés à la gestion des données de recherche

Type de coût (Titre)	Montant	Etape du cycle de vie
Matériel informatique (Matériel de prelevements)	2000 EUR	Coûts liés à la collecte/production des données - **Carnet et tableur**
Frais de personnel (Recrutement des personnelles)	10000 EUR	Coûts liés à la collecte/production des données - **Carnet et tableur**
Stockage (Cloud)	5000 EUR	Coûts liés au stockage et à la sauvegarde des données - **Carnet et tableur**
Stockage (Archivage à longue terme)	3000 EUR	Coûts liés à la conservation à long terme des données - **Carnet et tableur**
Stockage (Hébergement)	5000 EUR	Coûts liés au stockage et à la sauvegarde des données - **BDD**

Exemple extrait du PGD public « PGD1 : Suivi (fictif) de population de poissons dans le lac du Bourget », Arlette Sauvé (CNRS)

- « Ressources (budget et temps alloués) dédiées à la gestion des données :
  - 390 000 euros pour chef de projet RHU 60 mois ETP.
  - 120 000 euros pour la supervision des aspects techniques de la gestion des données de la recherche clinique.
  - 280 000 euros pour les chefs de projet en charge de la supervision des aspects réglementaires de la recherche (chef de projet promotion, délégués à la protection des données.
  - 1 034 000 euros pour les techniciens de recherche clinique chargés de la saisie des données cliniques : 235 mois ETP.
  - 250 800 euros pour l'attaché de recherche clinique (gestion et traitement des queries) : 57 mois ETP.
  - 71 000 euros pour le data management : programmation de la base de données / eCRF, pour l'extraction, fusion des bases et le nettoyage de la base finale (contrôles de cohérence) vérification de la qualité des données; 15 mois ETP.
  - 161 076 euros pour l'analyse et l'archivage de la base : PhD étudiant en épidémiologie (EPOPé), Ingénieur Statisticien (expérience 3-5 ans) et 2 stagiaires, M2 (EPOPé), 60 mois ETP.

  Soit un budget total de 2 306 876 euros pour l'ensemble de la gestion des données. »

  Exemple extrait du PGD public PrediMAP, François Goffinet et Karima Mesbahi-Ihadjadene (APHP)

6.7. Comment évaluer les coûts liés à la gestion des données ?

Voici quelques pistes pour vous aider à évaluer les coûts de la gestion des données :

- IST@INRAE. Evaluation des coûts éligibles au financement pour la gestion des données. 21 mars 2018. https://ist.blogs.inrae.fr/questionreponses/2018/03/21/evaluation-des-couts-eligibles-au-financement-pour-la-gestion-des-donnees/
- Antoine Masson, Eliane Blumer. Le data Management Plan Cost Calculator : l'automatisation au service des chercheurs de l'EPFL. Bulletin des bibliothèques de France (BBF), 2021-2. https://bbf.enssib.fr/consulter/bbf-2021-00-0000-057
- UK Data Service. Data management costing tool and checklist. https://dam.ukdataservice.ac.uk/media/622368/costingtool.pdf
- Ghent University. Costing RDM. https://www.ugent.be/en/research/datamanagement/before-research/costs.htm
- Guides OpenAIRE :
- OpenAIRE. How to identify and assess Research Data Management (RDM) costs. https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs
- O'Connor Ryan, Delipalta Alexandra, Jones, Sarah. What will it cost to manage and share my data? 21 mai 2020. https://zenodo.org/record/4548344

6.8. Qui peut financer la gestion des données ?

Pour les projets financés par l'ANR, toutes les dépenses liées à la gestion des données sont éligibles durant le projet : acquisition, collecte, stockage, personnel dédié à la gestion des données…
En cas de recours à un tiers, le coût de stockage des données est admissible jusqu'à 5 ans après la date de fin scientifique du projet, sous réserve que le contrat avec ce tiers soit conclu avant la fin scientifique du projet.

Source : Gala Garcia Reategui. La politique science ouverte de l'Agence Nationale de la Recherche et le DMP. https://octaviana.fr/document/VUN0041_02#

" Une bonne gestion des données de la recherche nécessite un investissement financier (infrastructures) et humain (temps de travail). La viabilité des projets

retenus par l'ANR impliquera donc une bonne estimation de ces dépenses par le porteur. "

Comité pour la Science ouverte. Plan de gestion de données – Recommandations à l'ANR. Juin 2019. https://www.ouvrirlascience.fr/plan-de-gestion-de-donnees-recommandations-a-lanr/

Voici un exemple pour illustrer cette question :

« La plateforme SI possède un budget propre pour la gestion des données dont elle est responsable, qu'il s'agisse du SNO KARST ou d'autres services d'observation. Ce budget couvre :

- La maintenance de l'infrastructure matérielle (serveurs, robot de sauvegarde, baie de stockage, onduleurs) ;
- Les licences (système d'exploitation, logiciel de sauvegarde);
- La cotisation auprès de DataCite (attribution de DOI).

Mutualisé à l'ensemble des données d'observation de l'OREME, il couvre intégralement les besoins budgétaires de la gestion des données du SNO KARST.

L'infrastructure matérielle est garantie jusqu'à fin 2022. Au-delà, trois solutions sont à l'étude pour chaque équipement :

- Le prolongement de garantie de l'équipement sur des crédits de fonctionnement ;
- La jouvence de l'équipement sur des crédits d'équipement ;
- Le non remplacement de l'équipement via la location de ressources sur des Cloud recherche, via des crédits de fonctionnement. »

Exemple extrait du PGD public SNO KARST, Juliette Fabre et Hervé Jourde (CNRS)

Cette ressource synthétise tous les éléments à prendre en compte dans l'estimation des coûts de gestion des données :

https://doranum.fr/enjeux-benefices/le-cout-de-la-gestion-des-donnees_10_13143_hch2-h207/

6.9. Recommandations pour la rédaction du PGD

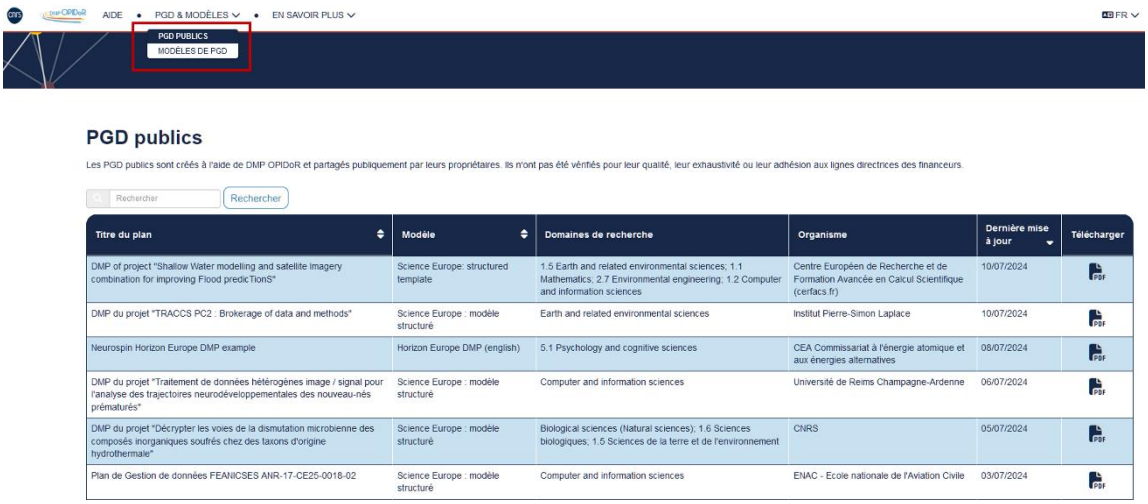
- Expliquer comment les ressources nécessaires (par exemple le temps) à la préparation des données pour le partage/préservation (curation des données) ont été chiffrées.
- Examiner et justifier soigneusement toutes les ressources nécessaires pour diffuser les données. Il peut s'agir de frais de stockage, de coût matériel, de temps de personnel, de coûts de préparation des données pour le dépôt, de frais d'entrepôt et d'archivage.
- Indiquer si des ressources supplémentaires sont nécessaires pour préparer les données en vue de leur dépôt ou pour payer tous les frais demandés par les entrepôts de données. Si oui, précisez le montant et comment ces coûts seront couverts.

Il est important de bien spécifier dans le PGD les coûts engendrés par la gestion des données.

7. PGD publics

7.1. Existe-t-il des exemples de PGD dont je peux m'inspirer ?

Outre ceux mentionnés tout au long de ce cours, vous trouverez des exemples sur DMP OPIDoR, dans la rubrique " **PGD publics** " :



https://dmp.opidor.fr/public_plans

Dans la **sélection de sources d'accès à des PGD** du site CoopIST du CIRAD, vous trouverez différentes plateformes de création de PGD qui proposent des PGD publics.

Il y en a également dans des bases de données bibliographiques, archives ouvertes de publications, des entrepôts de données de recherche, des moteurs de recherche académiques et des revues scientifiques.

Vous inspirer de PGD rendus publics vous permet d'avoir une vue d'ensemble d'un PGD et peut vous aider à compléter le vôtre.

8. Choix de l'outil de rédaction du PGD

8.1. Existe-t-il des outils facilitant la rédaction du PGD ?

Voici une liste d'outils de rédaction de PGD, non exhaustive :

- DMP OPIDoR : https://dmp.opidor.fr/
- ARGOS : https://argos.openaire.eu/home
- DSW (Data Stewardship Wizard) : https://ds-wizard.org/

Outils spécifiques à une discipline :

- RDMO (CC-IN2P3) : https://dmp.in2p3.fr/

Vous avez aussi la possibilité de rédiger votre PGD à partir d'un outil bureautique.

Le guide ci-dessous propose un comparatif des trois outils DMP OPIDoR, ARGOS et DSW :

https://view.genial.ly/62334341cc78d2001855ef43/guide-guidecomparatifoutilspgdgtsodr2022

8.2. Focus sur DMP OPIDoR

En France, DMP OPIDoR vous accompagne à travers l'élaboration et la mise en pratique de plans de gestion de données et de logiciels. Il suffit de créer un compte pour rédiger son (ses) plan(s) de gestion des données. Vous pouvez créer, exporter et partager votre PGD. C'est un outil collaboratif qui facilite les échanges entre les partenaires d'un même projet et les services d'accompagnement. DMP OPIDoR est également personnalisable par tout organisme de recherche pour la mise en place de sa politique de données. Il est possible d'ajouter des modèles et des recommandations de PGD, des exemples ou des réponses par défaut.

Le choix de l'outil de rédaction est important, notamment si votre projet implique des partenaires étrangers.

9. DMP OPIDoR

9.1. Comment utiliser DMP OPIDoR ?

L'outil DMP OPIDoR est mis à la disposition de la communauté de l'Enseignement Supérieur et de la Recherche.

C'est un outil gratuit qui permet de rédiger un PGD de manière collaborative et progressive.

Suivez le tutoriel « L'outil de rédaction DMP OPIDoR » qui vous guidera pas à pas : https://doranum.fr/plan-gestion-donnees-dmp/le-pgd-et-loutil-de-redaction-dmp-opidor_10_13143_es2g-0f16/

Vous avez fait le plus gros du travail en rédigeant votre PGD. Félicitations !

10. Grille de relecture

10.1. J'ai besoin de savoir s'il manque quelque chose dans mon PGD. Existe-t-il une checklist ?

La grille de relecture de PGD ci-dessous a été conçue avec 2 objectifs :

- Être utilisée comme outil d'accompagnement pour l'animation d'ateliers/formations sur le Plan de Gestion de Données ;
- Être utilisée comme checklist par les équipes de recherche pour vérifier que leur PGD est correctement rempli.

Grille de relecture de Plans de Gestion de Données : https://doranum.fr/plan-gestion-donnees-dmp/grilles-de-relecture-de-plans-de-gestion-de-donnees_10_13143_r7gm-6c38/

11. Le mot de la fin

L'essentiel n'est pas de répondre dès le début à toutes les questions du Plan de Gestion de Données. Vous avez la possibilité de le compléter au fur et à mesure de votre projet. Certaines questions ne nécessiteront peut-être pas de réponses.

L'important est de montrer que vous avez réfléchi à toutes ces questions.

Le PGD est vraiment un outil qui vous est utile pour la gestion de votre projet.

Retour d'expérience

Un chercheur vous parle de sa [première expérience de rédaction d'un PGD](#) :

factuel
l'info de l'université de lorraine

NOTRE ÉTABLISSEMENT | NOS FORMATIONS | NOS LABORATOIRES | NO

EN CE MOMENT | Journées portes ouvertes | L'UL by Léonor | Solidarité Ukraine | Égalité, diversité

NOTRE ETABLISSEMENT |

Un modèle de Plan de gestion des données pour vous inspirer

Publié le 31/01/2023 - Mis à jour le 1/02/2023

Vous avez du mal avec la rédaction des Plan de gestion de données (PGD) ? Pas de panique, nous vous proposons une aide pour démarrer le travail en douceur ! En 2022, Alain Celzard, chercheur à l'Institut Jean Lamour (IJL), a obtenu au côté de Groupe BORDET un financement ANR (ANR-21-LCV3-0001) pour son projet LabCom « Laboratoire de Carbones Biosourcés IJL-Bordet – CarbioLab » et rédigé son premier Plan de Gestion des Données. Il a accepté de le partager pour qu'il puisse servir d'exemple.

Le document est librement accessible sur DMP Opidor à tous enseignants-chercheurs de l'UL. N'hésitez pas à le consulter et à vous en inspirer largement pour vos propres plans. Pour y accéder, il vous suffit de vous connecter à DMP Opidor via votre compte Université de Lorraine.

Dans la rédaction de son PGD, Alain a bénéficié des conseils de l'ambassadrice des données de son laboratoire : Sophie Legeai. Dans chaque laboratoire volontaire, un enseignant-chercheur formé aux bonnes pratiques de gestion des données de la recherche vous accompagne. Le réseau continue de recruter !

L'atelier de la donnée ADOC Lorraine peut aussi vous accompagner dans la rédaction de votre Plan de Gestion des Données. Contactez-nous.

L'avis du chercheur

Quel est votre avis sur le PGD ?

Dans la pratique, je n'ai pas encore d'avis très clair sur la question comme, je pense, nombre de mes collègues, car c'était mon tout premier PGD ! Néanmoins, le fait d'avoir dû s'y pencher et y réfléchir m'a permis de m'interroger sur ce que deviennent les données de la recherche. De ce point de vue, il faut reconnaître que si l'apparition des cahiers de laboratoire, auxquels chacun est maintenant habitué depuis longtemps, a été un réel progrès, ce système est encore insatisfaisant. On a tous essayé de rouvrir des cahiers issus de thèses déjà anciennes, et en dépit du soin apporté par les doctorants à leur rédaction, il est bien souvent difficile de retrouver l'information recherchée, le petit détail oublié ...

https://factuel.univ-lorraine.fr/node/22353

12. FAQ

D'autres questions peuvent se poser, n'hésitez pas à consulter la [FAQ de DoRANum](#) en complément de ce cours. Vous y trouverez peut-être des réponses à vos questions.

Dans le cas contraire, vous pouvez contribuer au développement de cette FAQ en nous [envoyant votre question](#).

13. Testez vos connaissances

Consigne : cochez la bonne réponse

1/12. Qu'est-ce qu'un PGD ?

☐ Un document évolutif qui prépare le partage, la réutilisation et la pérennisation des données

- [ ] Un dirigeant de société
- [ ] Un outil de gestion des données
- [ ] Un délégué à la protection des données

Solution :

Le PGD est un document évolutif qui peut être complété tout au long du projet. C'est également un outil de gestion de projet.

Non, le PGD n'est pas un PDG ! Ce n'est pas non plus le DPD/DPO !

2/12. À quelle étape de votre activité de recherche intégrez-vous la rédaction du PGD ?

- [ ] À la fin de mes recherches
- [ ] Quand je le souhaite
- [ ] Au début de mon projet

Solution :

Le fait de démarrer la rédaction du PGD **au tout début du projet** aide à organiser et à anticiper toutes les étapes du cycle de vie des données.

Ce n'est pas quand vous le souhaitez, car pour être utile, il doit être pensé dès le début du projet. À la fin des recherches, il ne sert plus à rien !

3/12. Dans le cas d'un projet financé par l'ANR (Agence nationale de la recherche)...

- [ ] La rédaction d'un PGD est obligatoire
- [ ] La rédaction d'un PGD est fortement recommandée

Solution :

La rédaction d'un PGD est obligatoire.

4/12. Qui peut m’aider à rédiger un PGD ?

- [ ] Un informaticien
- [ ] Un professionnel de l’IST
- [ ] Un assistant administratif du laboratoire
- [ ] Un délégué à la protection des données

Solution :

- Un informaticien peut vous aider sur tout ce qui concerne les formats, nommages de fichiers, le dépôt des codes sources...
- Un professionnel de l'IST peut vous accompagner tout au long du cycle de vie des données.
- Un délégué à la protection des données vous conseillera notamment sur la protection des données personnelles et le respect du RGPD.

La rédaction du PGD n'est pas du ressort de l'assistant administratif.

5/12. Vous voulez décrire le contenu de votre jeu de données. Quelles informations renseignez-vous dans les métadonnées ?

- ☐ Des informations sur le contexte de création ou de collecte des données
- ☐ Des informations sur les protocoles expérimentaux utilisés
- ☐ L'adresse du laboratoire
- ☐ Des mots-clés issus d'un vocabulaire contrôlé (thésaurus, ontologie...)
- ☐ L'âge de l'auteur

Solution :

- Les informations sur le contexte de création ou de collecte des données ainsi que sur les protocoles expérimentaux utilisés sont très importantes pour faciliter la compréhension des données.
- Les mots-clés favorisent l'Interopérabilité des données et les rendent plus Faciles à trouver.

Ce n'est pas l'adresse du laboratoire qui importe. C'est le lieu de l'étude qu'il est important de décrire.

L'âge de l'auteur n'a pas sa place ici !

6/12. En règle générale, l'attribution de la propriété intellectuelle des données revient...

- ☐ Au producteur de ces données
- ☐ À l'établissement de tutelle des producteurs de données

Solution :

Il ne s'agit pas du même droit que pour les publications.

Les données relèvent d'un **régime lié au droit des bases de données**. Dans ce cas, le droit de propriété appartient légalement au « producteur » de la base de données, compris au sens de la personne qui réalise l'investissement financier et matériel nécessaire à la constitution de la base. Il s'agira donc en général de **l'établissement de tutelle des chercheurs** qui **sera considéré comme le titulaire effectif du droit de propriété**.

Mais si ce droit existe formellement, il ne peut plus être opposé aux droits des ré-utilisateurs des données (principe d'ouverture des données). En effet, la **loi pour une République numérique** a explicitement « neutralisé » le droit des bases de données des administrations pour faire primer le **principe de libre réutilisation**. Il en résulte que les données produites par les chercheurs sont bien comprises dans le principe d'ouverture par défaut.

7/12. Stockage, partage, archivage : chacune de ces étapes a son rôle à jouer par rapport aux données…

Voir l'original sur LearningApps.org : https://learningapps.org/watch?v=p32z51huk22
Glissez et déposez chaque carte sur une des trois zones grises. Dans ces tableaux, les réponses sont dans le désordre.

• Faciliter l'accès aux données aux collaborateurs du projet • Rendre les données accessibles sur le long terme • Conserver les données à court et moyen terme • Conserver les données à long terme (plus de 30 ans)	Stockage et sauvegarde
• Déposer dans un entrepôt de données • Conserver des données présentant une valeur scientifique reconnue par la communauté d'où elles proviennent • Permettre la réutilisation des données	Partage
• Garantir la sécurité des données en utilisant un serveur sécurisé de l'institution • Faciliter l'accès aux données à des personnes extérieures au projet	Archivage pérenne

Solution :

Stockage et sauvegarde • Conserver les données à court et moyen terme • Faciliter l'accès aux données aux collaborateurs du projet • Garantir la sécurité des données en utilisant un serveur sécurisé de l'institution
Partage • Déposer dans un entrepôt de données • Permettre la réutilisation des données • Faciliter l'accès aux données à des personnes extérieures au projet
Archivage pérenne • Rendre les données accessibles sur le long terme • Conserver les données à long terme (plus de 30 ans) • Conserver des données présentant une valeur scientifique reconnue par la communauté d'où elles proviennent

8/12. Comment choisir une licence pour ses données ?

Consigne : glissez et accolez chaque élément comportant le nom d'une licence (à gauche) avec la modalité de réutilisation de la licence (à droite).

Licence ouverte (Etalab)	Respect du droit français et des principes du logiciel libre
CC-BY	Liberté de partager, créer, adapter une base de données et obligation de citation
CC-By-NC-SA	Attribuée aux données publiques françaises et obligation de citer les auteurs
CeCILL	Licence ouverte, obligation de citer les auteurs
ODBL	Obligation de citation, pas d'utilisation commerciale, partage dans les mêmes conditions

Solution :

Licence ouverte (Etalab)	Attribuée aux données publiques françaises et obligation de citer les auteurs
CC-By-NC-SA	Obligation de citation, pas d'utilisation commerciale, partage dans les mêmes conditions
CC-By	Licence ouverte, obligation de citer les auteurs
CeCILL	Respect du droit français et des principes du logiciel libre
ODBL	Liberté de partager, créer, adapter une base de données et obligation de citation

9/12. S'agit-il d'un identifiant pérenne ou d'un standard de métadonnées ?

Consigne : glissez et déposez chaque carte sur une des deux zones grises.

SWIHD DDI ORCID	Identifiant pérenne
EML DOI Dublin Core	Standard de métadonnées

Solution :

Identifiant pérenne SWHID ORCID DOI
Standard de métadonnées DDI EML Dublin Core

10/12. S’agit-il d’un format ouvert ou d’un format fermé ?

Consigne : glissez et déposez chaque carte sur une des deux zones grises.

.pdf .doc .csv	Format ouvert
.xls .txt .avi	Format fermé

Solution :

Format ouvert .pdf csv .txt
Format fermé .doc .xls .avi

11/12. À quoi servent ces outils/services ?

Consigne : glissez et accolez chaque élément comportant le nom d'un outil (à gauche) avec sa définition (à droite).

Software Heritage	Vérifier la validité des formats de fichiers de données pour l'archivage pérenne
re3data	Recense et décrit les services français dédiés aux données scientifiques
Data Management Plan Cost Calculator	Annuaire d'entrepôts
DMP OPIDoR	Archive universelle de codes sources logiciels
Cat OPIDoR	Pour évaluer les coûts liés à la gestion des données
FACILE	Pour savoir à qui appartiennent les données, qualifier si vos données sont diffusables
Infographie de l'École des Ponts ParisTech	Outil de rédaction des plans de gestion de données

Solution :

re3data	Annuaire d'entrepôts
Software Heritage	Archive universelle de codes sources logiciels
Cat OPIDoR	Recense et décrit les services français dédiés aux données scientifiques
FACILE	Pour la validité des formats de fichiers de données pour l'archivage pérenne
Data Management Plan Cost Calculator	Pour évaluer les coûts liés à la gestion des données
DMP OPIDoR	Outil de rédaction des plans de gestion de données
Infographie de l'École des Ponts ParisTech	Pour savoir à qui appartiennent les données, qualifier si vos données sont diffusables

12/12. Cherchez les erreurs dans ce texte original extrait d'un PGD public

Texte original :

" **How will you manage access and security?**

One risk to data security is the loss of de-identification logs which contain personally identifying information on study participants. De-identification logs will be kept in a security case with a key lock. Only the principal investigator and a trusted third party (the principal investigator's mother) will hold keys to the security case. "

Extrait du PGD de Nicholas Bell. Why do so few workers take trade adjustment assistance? https://osf.io/4zjwb

Traduction :

Comment allez-vous gérer l'accès et la sécurité ?

Un risque pour la sécurité des données est la perte des journaux de dépersonnalisation qui contiennent des informations d'identification personnelle sur les participants à l'étude. Les journaux de dépersonnalisation seront conservés dans une mallette sécurisée avec une serrure à clé. Seuls l'investigateur principal et un tiers de confiance (la mère de l'investigateur principal) détiendront les clés de la mallette sécurisée.

Solution :

Une mallette peut être perdue ou volée.

Le fait de confier les clés à sa mère n'est pas une solution sécurisée :

- Elle peut perdre les clés !
- Elle n'est pas habilitée à détenir un accès à des données personnelles pour un projet dans lequel elle n'est pas impliquée.

14. Webographie

- *ANR. L'ANR met en place un plan de gestion des données pour les projets financés dès 2019. 5 septembre 2019. https://anr.fr/fr/actualites-de-lanr/details/news/lanr-met-en-place-un-plan-de-gestion-des-donnees-pour-les-projets-finances-des-2019/*
- *Arroyo Daniel, Gouttevin Isabelle, Six Delphine. PGD GLACIOCLIM. 6 février 2024. https://dmp.opidor.fr/plans/23294/export.pdf*

- *Cannavò Anna. PGD du projet " Prospection du territoire d'Amathonte ". 21 janvier 2022. https://dmp.opidor.fr/plans/12314/export.pdf*
- *Cavalier Jean-François. PGD du projet " LipInTB ". 2 août 2021. https://dmp.opidor.fr/plans/4624/export.pdf*
- *Cicolani Veronica. "Techniques artisanales au premier âge du Fer en Italie. Repenser les interactions culturelles." project DMP. https://dmp.opidor.fr/plans/12645/export.pdf*
- *CINES (Centre Informatique National de l'Enseignement Supérieur). https://www.cines.fr/*
- *Cirad. CoopIST. Gérer les données de la recherche. https://coop-ist.cirad.fr/gerer-des-donnees*
- *Comité pour la Science ouverte. Plan de gestion de données – Recommandations à l'ANR. Juin 2019. https://www.ouvrirlascience.fr/plan-de-gestion-de-donnees-recommandations-a-lanr/*
- *Couperin. Groupe de travail science ouverte. SOS-PGD. https://gtso.couperin.org/gtdata/sos-pgd/*
- *Couperin. Groupe de travail science ouverte. Groupe données. https://gtso.couperin.org/groupe-donnees/*
- *Deboin Marie-Claude. Trouver des plans de gestion de données (PGD) pour s'en inspirer. 11 février 2022. https://coop-ist.cirad.fr/gerer-des-donnees/trouver-des-plans-de-gestion-de-donnees-pgd-pour-s-en-inspirer/*
- *Deyris Laetitia, Bernier Frédéric. PGD UEFP. 31 août 2022. https://dmp.opidor.fr/plans/5206/export.pdf*
- *Duterme Clara, Anstett Elisabeth. PGD du projet " TRANSFUNERAIRE ". 22 juillet 2020. https://dmp.opidor.fr/plans/6619/export.pdf*
- *EOSC Pillar. What is data? Interviews with French researchers. 17 janvier 2022. https://www.youtube.com/watch?v=J7nkClygNng*
- *Bibliothèque de l'EPFL. Research Data Management horror stories. https://www.youtube.com/watch?v=t_rEXpfCTrg&list=PLPkfOHxsjx2hH-QmfYp_ZHZI2WmE6pXLv&index=2*
- *Fabre Juliette, Jourde Hervé. PGD du projet " SNO KARST ". 14 juillet 2021. https://dmp.opidor.fr/plans/9351/export.pdf*
- *Galle Sylvie. PGD du projet "Theia/OZCAR Information System". 12 décembre 2022. https://dmp.opidor.fr/plans/7725/export.pdf*

- *Garcia Reategui Gala. La politique science ouverte de l'Agence Nationale de la Recherche (ANR) et le DMP. 2021. https://octaviana.fr/document/VUN0041_02#*
- *Giacomoni Franck, Jourdan Fabien. PGD du projet " EQUIPEX MetaboHUB-METEX+ ". 20 mai 2022. https://dmp.opidor.fr/plans/13271/export.pdf*
- *Goby Cédric, Torregrosa Laurent. PGD du projet " Grape Genes for WAter Scarcity ". 20 mai 2022. https://dmp.opidor.fr/plans/2486/export.pdf*
- *Goffinet François, Mesbahi-Ihadjadene Karima. PGD du projet "PrediMAP : Développement et évaluation clinique d'un dispositif médical innovant pour prédire l'accouchement prématuré - De la recherche fondamentale aux urgences obstétricales". 5 octobre 2022. https://dmp.opidor.fr/plans/16455/export.pdf*
- *Hadrossek Christine, Janik Joanna, Libes Maurice, Louvet Violaine, Quidoz Marie-Claude, Rivet Alain, Romier Geneviève. Guide de bonnes pratiques sur la gestion des données de la recherche. Version 2.0. 8 janvier 2023. https://mi-gt-donnees.pages.math.unistra.fr/guide/*
- *Huma-Num. https://www.huma-num.fr/*
- *Inist-CNRS. Cat OPIDoR, wiki des services dédiés aux données de la recherche. https://cat.opidor.fr/*
- *Inist-CNRS. DMP OPIDoR. https://dmp.opidor.fr/*
- *Inist-CNRS, GIS Urfist. DoRANum. https://doranum.fr/*
- *Lenoir Jonathan. PGD du projet " IMPRINT ". 12 mars 2020. https://dmp.opidor.fr/plans/5082/export.pdf*
- *Locatelli Lauriane. PGD du projet " Hospitam ". 24 avril 2020. https://dmp.opidor.fr/plans/5278/export.pdf*
- *Manola Théa. PGD du projet " PROduction SEnsible des projets urbains*
- *COntemporains ". 7 décembre 2021. https://dmp.opidor.fr/plans/11722/export.pdf*
- *Maurel Lionel. À qui appartiennent les données ? 14 septembre 2020. https://mate-shs.cnrs.fr/actions/tutomate/tuto25-propriete-donnees-lionel-maurel/*
- *Menot Lenaick. PGD du projet " ARDECO ". 1 octobre 2021. https://dmp.opidor.fr/plans/8349/export.pdf*

- *Ministère de l'Enseignement Supérieur et de la Recherche. Présentation du programme Horizon Europe. 9 décembre 2020. https://www.horizon-europe.gouv.fr/presentation-du-programme-horizon-europe-24104*
- *Ministère de l'Enseignement supérieur et de la Recherche. Deuxième Plan national pour la science ouverte. Généraliser la science ouverte en France 2021-2024. Juillet 2021. https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/*
- *Ministère de l'Enseignement supérieur et de la Recherche. Recherche Data Gouv. https://recherche.data.gouv.fr/*
- *Mission pour les Initiatives Transverses et Interdisciplinaires du CNRS. Atelier " Données ". https://mi-gt-donnees.pages.math.unistra.fr/site/*
- *Monet Ghislaine. PGD du projet " OLA-Infrastructure ". 21 août 2021. https://dmp.opidor.fr/plans/4781/export.pdf*
- *OCDE : Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. [archive] [PDF], Paris, 2007. http://www.oecd.org/fr/science/inno/38500823.pdf*
- *Sauvé Arlette. PGD du projet " PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget ". 17 novembre 2022. https://dmp.opidor.fr/plans/12770/export.pdf*
- *Software Heritage. https://www.softwareheritage.org/?lang=fr*
- *Stérin Anne-Laure. Diffuser des données de la recherche dans le respect du droit et de l'éthique – Comment faire lorsqu'on n'est pas juriste ? octobre 2018. https://hal.science/hal-02050510*
- *Université Grenoble Alpes, UMS GRICAD, SARI. Archivage Numérique des Données de Recherche. 20 novembre 2019. https://videos.univ-grenoble-alpes.fr/recherche/archivage-numerique-des-donnees-de-recherche/*