Atelier
Plan de Gestion de Données

RÉSEAU NATIONAL
DES BIBLIOTHÈQUES
DE MATHÉMATIQUES

ANF RNBM

5 octobre 2021

- Lucie Albaret, Bibliothèques universitaires, Université Grenoble Alpes
- Brigitte Bidegaray, LJK, CNRS
- Paolo Lai, INIST, CNRS
- Violaine Louvet, GRICAD , CNRS

Enjeux liés aux données de la recherche

- En premier lieu **faciliter le travail de recherche** :
  - Assurer la **conservation** des données à moyen terme
  - Pouvoir **reproduire** les résultats scientifiques, y compris par l'équipe qui les a obtenus !
  - Pouvoir facilement les **réutiliser** pour produire de nouvelles recherches
  - **Valoriser** les résultats scientifiques, mais aussi les données et les codes qui ont permis de les obtenir et en augmenter la visibilité
  - Favoriser de **nouvelles collaborations**, de nouvelles approches
- Enjeux d'ordre sociétaux :
  - Assurer une **souveraineté** sur les données produites
  - Assurer l'**intégrité scientifique**
  - Garantir la **transparence**, et assurer la confiance des citoyens en la recherche

Dans un cadre légal de plus en plus structuré

- **RGPD** (2016) : règlement européen sur la protection des données personnelles
- **Obligation de Plan de Gestion de Données** exigée par de plus en plus de financeurs de la recherche (H2020, ANR, Horizon Europe, ...)
- Loi pour une république numérique (2016), Plan National pour la Science Ouverte (2018, 2021), H2020 (depuis 2017) et Horizon Europe (à partir de 2021), Feuille de route de la science ouverte du CNRS (2018), Plan données de la recherche CNRS (2020), Feuille de route Politique des données, des algorithmes et des codes sources du MESRI (2021) : l'**ouverture des données de recherche financées sur fonds publics devient la norme**

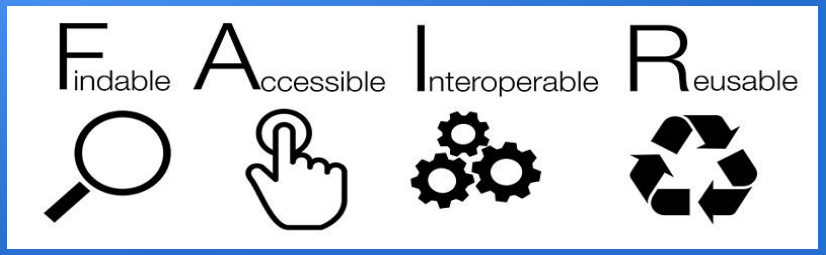
Ouverture des données de la recherche

Par défaut, l'**ouverture des données de la recherche est obligatoire selon le principe "aussi ouvert que possible, aussi fermé que nécessaire"**.

Exceptions :

- Données **personnelles** non anonymisées ou pseudonymisées
- Données **non achevées** (sauf données géographiques qui peuvent être inachevées, directive INSPIRE), certaines données environnementales (cf protection des espèces)
- **Photos** quand il y a une personne reconnaissable
- Données **scrapées** (car souvent interdit par les conditions d'utilisation) sauf :
  - si le préjudice du producteur est nul (risque encouru nul également)
  - le scraping d'une partie non substantielle des données reste possible

Principes FAIR



Findable / *Trouvable*

Données **faciles à trouver**.

- possédant un identifiant unique et pérenne
- décrites par des métadonnées riches
- enregistrées ou indexées dans une source interrogeable

Interoperable / *Interopérable*

Facile à combiner avec d'autres jeux de données, par les humains **et** les systèmes informatiques

- formats libres et ouverts
- mise à disposition du code source si le logiciel de traitement existe
- standards de métadonnées et vocabulaire standardisés

Accessible / *Accessible*

Données ou au moins méta-données facilement accessibles.

- entrepôt de confiance, pérenne, certifié
- définir les conditions d'accès et la licence de diffusion
- si embargo ou accès restreint : méta-données accessibles

Reusable / *Réutilisable*

Prêtes à être **réutilisables** pour une future recherche y compris via des méthodes informatiques

Plan de Gestion de Données (PGD) ou Data Management Plan (DMP)

Ce n'est pas :

- un nième document administratif …

C'est :

- une **aide concrète à la gestion des données** durant et après la phase de recherche
- un outil normalisé et évolutif tout au long du projet
- un livrable du projet

Objectif : anticiper la gestion des données dans tous les aspects de la recherche en se posant les bonnes questions

Exigé par des financeurs :

- Horizon 2020 depuis 2016, Horizon Europe
- ERC depuis 2017
- ANR depuis 2019

A quoi ça ressemble ?

- Document formel qui décrit la façon dont les données seront :

  - **obtenues,**
  - **traitées,**
  - **organisées,**
  - **stockées,**
  - **sécurisées, préservées,**
  - **partagées**, ...

  au cours et à l'issue d'un projet.

- Incite à la mise en place de **bonnes pratiques de gestion** à toutes les étapes du cycle de vie des données

Pour répondre aux principes FAIR de l'open data

De façon plus précise

Le PGD (ou DMP, Data Management Plan) **permet de réfléchir à la gestion des données d'un projet en amont** :

- **Quelles données vont être collectées :** quels types de données, comment sont-elles collectées, où les stocker, comment on sécurise le stockage, quelle volumétrie, quels formats, quelle organisation ...
- **Comment vont-elles être utilisées :** comment on les partage, comment on les traite, où on les traite, ...
- **Comment elles vont être préservées :** à quel terme, quelles données, où, comment ...
- **Comment elles vont être valorisées :** comment les diffuser, sous quel format, sous quelle licence, quelles données, comment associer les codes, …
- **Comment assurer le financement des ressources nécessaires ?**

Calendrier

Le calendrier pour les ANR et les projets européens :

- 1ere version (obligatoire) : 6 mois après le démarrage du projet
- Mises à jour (recommandées) à chaque évaluation du projet
- Version finale à la fin du projet (évaluation)

A noter :

la première version du PGD n'implique **pas des réponses complètes et précises** à l'ensemble des questions. Elle offre surtout la possibilité pour les porteurs du projet de réfléchir aux différentes problématiques liées à la gestion des données.

En pratique : modèle ANR

6 parties :

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

2. Documentation et qualité des données

3. Stockage et sauvegarde pendant le processus de recherche

4. Exigences légales et éthiques, codes de conduite

5. Partage des données et conservation à long terme

6. Responsabilités et ressources en matière de gestion des données

En pratique : modèle européen

6 parties :

1. Présentation des données

2. Données FAIR
   - Rendre les données trouvable, y compris les métadonnées
   - Rendre les données librement accessibles : diffusion des données
   - Rendre les données interopérables
   - Augmenter la réutilisation des données (licences)

3. Répartition des ressources

4. Sécurité des données

5. Aspects éthiques

6. Autres aspects

Suite de l’atelier

- Travail en deux groupes : présentiel et distanciel
  - Prévoir un rapporteur par groupe pour faire remonter les points importants des discussions dans chaque groupe
- Objectif : « entrer » dans des PGDs existants, assez différents, afin d’appréhender les contenus de chaque partie
- Travail collectif : ne pas hésiter à intervenir, à demander des précisions, à faire des remarques ...
- Fil conducteur : grille d’analyse et de relecture de PDG sur le modèle ANR
- Sur plusieurs PGD issus de projets en mathématiques, et dans d’autres disciplines afin d’aborder l’ensemble des points
- 45 mn d’échanges, puis une pause de 10 mn et enfin un temps d’échanges sur le travail réalisé

Séparation en 2 groupes

- Personnes en présentiel : personne ne bouge !!
- En distanciel : connexion à la salle zoom

1ère partie : informations administratives du projet

Informations sur le plan de gestion

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Auteur du PGD (si différent du responsable du projet) : nom, courriel, affiliation	☐	☐	☐	☐
Date de création du PGD	☐	☐	☐	☐
Date de la dernière version	☐	☐	☐	☐

Informations sur le projet de recherche

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Acronyme du projet	☐	☐	☐	☐
Nom/Titre du projet de recherche	☐	☐	☐	☐
Financeur(s) du projet	☐	☐	☐	☐
Numéro de subvention et/ou référence de la convention de financement	☐	☐	☐	☐
Résumé du projet	☐	☐	☐	☐
Dates et durée du projet	☐	☐	☐	☐
Personne responsable du projet (chercheur principal) : nom, identifiant (ORCID ou autre), courriel, téléphone	☐	☐	☐	☐
Personne responsable des données (si différente du responsable du projet) : nom, courriel, téléphone	☐	☐	☐	☐
Institution leader du projet, coordinateur bénéficiaire (nom, pays)	☐	☐	☐	☐
Autres partenaires (nom, pays, rôle de chacun des partenaires en dehors de l'institution leader)	☐	☐	☐	☐

PALAGRAM - PArticle-LAden GRAvity currents Modelling

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateurs du PGD : Laurent Lacaze, Hervé Bellot, florence.naaim@inrae.fr, cyrille.bonamy@univ-grenoble-alpes.fr, sébastien kiesgen de richter, julien chauchat, Marie Rastello, Yvan Dossmann, Matthieu Mercier

Affiliation du créateur principal : CNRS

Modèle du PGD : ANR - Modèle de PGD (français)

Dernière modification du PGD : 21/11/2020

Financeur : Agence nationale de la recherche (ANR)

Numéro de subvention : ANR-19-CE30-0041

Résumé du projet :

Les courants de gravité chargés en particules solides sont des situations récurrentes rencontrées dans la nature. La récurrence des ces écoulements peut être attribuée à la diversité de leur origine (rupture de terrain, apport fluvial, ...), l'impact climatique sur ces évènements restant alors incertain. L'érosion du sol ainsi que la dissipation locale à la surface du courant par mélange sont des points clés de leur un impact. Le projet PALAGRAM vise à caractériser la dynamique des courants chargés en particules de l'échelle des processus locaux (érosion/déposition, turbulence, flux à l'interface) à l'échelle macroscopique (rhéologie, modélisation de la turbulence) par des approches expérimentales de laboratoire et numériques. La modélisation apportée à ces écoulements sera confrontée à des mesures in-situ disponibles par des interactions avec la communauté internationale. Le consortium du projet PALAGRAM permet de regrouper des expertises variées sur les écoulements fluide-particules et leur modélisation, sur un même objet d'étude afin d'améliorer la connaissance et l'impact de ces écoulements complexes.

Chercheur Principal : Laurent Lacaze

Identifiant ORCID : 0000-0002-4945-7445

Contact pour les Données : Laurent Lacaze

Produits de recherche :

1. Ombro_IMFT : Ombroscopie Courant Stationnaire (IMFT) (Image)
2. PIV_IMFT : PIV Courant Stationnaire (IMFT) (Jeu de données)
3. Acoustique_IMFT : Mesure Acoustique Courant Stationnaire (IMFT) (Jeu de données)
4. openFOAM_LEGI : Simulations openFOAM Courant Stationnaire et Courant Instationnaire (LEGI) (Jeu de données)
5. Imag_LEGI/INRAE : Imagerie Courant Instationnaire Forte Pente (LEGI/INRAE) (Image)
6. PIV_LEGI/INRAE : PIV recouvrement d'indice optique Courant Instationnaire Forte Pente (LEGI/INRAE) (Jeu de données)

Selon vous, a-t-on toutes les informations pertinentes ?

1ère partie : informatic administratives du pro

Informations sur le plan de gestion

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseigr
Auteur du PGD (si différent du responsable du projet) : nom, courriel, affiliation	☐	☐	☐
Date de création du PGD	☐	☐	☐
Date de la dernière version	☐	☐	☐

Informations sur le projet de recherche

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseigr	
Acronyme du projet	☐	☐	☐	
Nom/Titre du projet de recherche	☐	☐	☐	
Financeur(s) du projet	☐	☐	☐	
Numéro de subvention et/ou référence de la convention de financement	☐	☐	☐	
Résumé du projet	☐	☐	☐	
Dates et durée du projet	☐	☐	☐	
Personne responsable du projet (chercheur principal) : nom, identifiant (ORCID ou autre), courriel, téléphone	☐	☐		
Personne responsable des données (si différente du responsable du projet) : nom, courriel, téléphone	☐	☐		
Institution leader du projet, coordinateur bénéficiaire (nom, pays)	☐	☐	☐	☐
Autres partenaires (nom, pays, rôle de chacun des partenaires en dehors de l'institution leader)	☐	☐	☐	☐

1. PROJECT INFORMATION

- **Action to be implemented:**
  Open Digital Research Environment Toolkit for the Advancement of Mathematics - OpenDreamKit, project number 676541
- **Funding Programme:**
  OpenDreamKit is a Horizon 2020 European Research Infrastructure project
- **Objective:**
  It will provide substantial funding to the open source computational mathematics ecosystem, and in particular popular tools such as LINBOX, MPIR, SageMath, GAP, PARI/GP, LMFDB, SINGULAR, MathHub, and the IPython/Jupyter interactive computing environment. From this ecosystem, OpenDreamKit will deliver a flexible toolkit enabling research groups to set up Virtual Research Environments, customised to meet the varied needs of research projects in pure mathematics and applications, and supporting the full research life-cycle from exploration, through proof and publication, to archival and sharing of data and code.
- **Partners:**
  1- Université Paris-Sud (UPSud)
  2- Centre National de la Recherche Scientifique (CNRS)
  3- Jacobs University Bremen GGMBH (JacobsUni - Terminated)
  4- Université Grenoble-Alpes (UGA)
  5- Technische Universitaet Kaiserslautern (UNIKL)
  6- The Chancellor, Masters and Scholars of the University of Oxford (UOXF)
  7- Uniwersytet Slaski (USlaski)
  8- The University of Sheffield (USFD - Terminated)
  9- University of Southampton (Southampton - Terminated)
  10- The University Court of the University of St Andrews (USTAN)
  11- Université de Versailles-Saint-Quentin-en-Yvelines (UVSQ)
  12- The University of Warwick (UWarwick)
  13- Universitaet Zuerich (UZH - Terminated)
  14- Logilab (Logilab)
  15- Simula Research Laboratory AS (Simula)
  16- Universiteit Ghent (UGent)
  17- European XFEL (XFEL)
  18- Universität Erlangen-Nürnberg (FAU)
  19- University of Leeds (ULeeds)
- **Timeframe:**
  The project lasts 48 months, from the 1st of September 2015 until the 31st of August 2019.

Autre exemple sur un projet européen

Description des données

Collecter et produire ses données

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

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

En cas de réutilisation de données, provenance et mode de collecte
En cas de production, méthodes, logiciels, outils utilisés

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Nature des données produites ou réutilisées (par ex : RMN) ?
Type des données (par exemple image, audio, vidéo, texte, numérique, etc.) ?
Format (ouvert ou fermé)
Volumétrie prévisionnelle

Description des données et collecte ou réutilisation de données existantes

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Quelle est la provenance des données ? (collecte de nouvelles données ; conversion / transformation de données ; partage / échange de données ; achat de données)	☐	☐	☐	☐
Quels méthodes, logiciels et outils seront utilisés pour produire et traiter les données ?	☐	☐	☐	☐
Quels types de données seront collectés ou réutilisés (données numériques, textuelles, images, audio, vidéos, médias divers…) ?	☐	☐	☐	☐
Y a-t-il des restrictions à la réutilisation des données ?	☐	☐	☐	☐

Description des données : exemples

Projet Hitrac, High-resolution traction force microscopy

Physique

1. DESCRIPTION DES DONNEES ET COLLECTE OU REUTILISATION DE DONNEES EXISTANTES

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les données produites dans le cadre du projet seront de trois types :
- protocoles de fabrication d'échantillons et de mesures expérimentales
- données numériques (texte ou image)
- fichiers ou programmes d'analyse et donneés analysées

Les protocoles sont documentés in situ à l'aide de cahier de laboratoire, et, dans les cas pertinents, retranscrits numériquement et synthétisés.
Les données numériques sont acquises sur plusieurs appareils : AFM, microscope confocal, microscope classique. La provenance des données sera documentée numériquement dans le dossier de sauvegarde de chaque série d'expériences.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les types de données produites seront :
- cahiers de laboratoire
- fichiers .doc, .pdf et .ppt (protocoles et compte-rendu d'expérience documentés dans des cahiers de laboratoire)
- images en format .lif ou .tif, accompagnées de fichiers de métadonnées (.xml,.txt)
- fichiers numériques .txt et .dat contenant des données expérimentales brutes (AFM notamment)
- fichiers d'analyse de données en format .nb, .mat, .xls

Des volumes importants sont prévus pour les fichiers image (typiquement plusieurs à quelques dizaines de To)

Description des données : exemples

Projet Palagram, PArticle-LAden GRAvity currents Modelling

Mécanique des fluides

Ombro_IMFT : Ombroscopie Courant Stationnaire (IMFT)

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les données seront recueillies par une caméra pco edge scmos (résolution 2400x2000) à fréquence contrôlable (0-100Hz), disponible à l'IMFT. La résolution sera de 1 mm/pixel
Aucune donnée préexistante sur cette configuration.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données sont des images au format .tiff avec un fichier de métadonnée par séquence (au format texte .rec). Le volume estimé est de 6To.

PIV_IMFT : PIV Courant Stationnaire (IMFT)

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les données seront recueillies par une caméra rapide pco edge (résolution 2000x2000) à fréquence 1000Hz, disponible à
ement PIV sera réalisé par le logiciel LAVISION. La sortie des fichiers
7DaVis Image Data.

s par ex.) seront collectées ou produites ?

champ de vitesse 2D dans le plan imagé. Le volume estimé est de

openFOAM_LEGI : Simulations openFOAM Courant Stationnaire et Courant Instationnaire (LEGI)

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

les données seront obtenues par des simulations numériques effectuées sur les plateformes GENCI à l'aide du logiciel opensource OpenFOAM.
Il n'existe pas de données existantes sur cette configuration.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données extraites sont au format txt et NetCDF.
Le volume estimé est de quelques 100aine de Go

Description des données : exemples

Projet LipInTB, Utilisation de nouveaux inhibiteurs sélectifs pour décrypter le métabolisme lipidique et de la virulence chez M. tb

Biologie

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

Les données recueillies dans le cadre de l'étude proposée comprendront la synthèse de composés chimiques (protocoles expérimentaux et caractérisation chimique) et leur utilisation en tant que sondes moléculaires. Chaque composé synthétisé sera purifié et entièrement caractérisé pour confirmer sa structure et sa pureté chimique à l'aide de méthodes spectroscopiques et chromatographiques. Les données spectroscopiques classiques comprennent la RMN 1H, la RMN 13C, la spectrométrie IR et la spectrométrie de masse ; les techniques chromatographiques comprennent la chromatographie en phase liquide couplée à la spectrométrie de masse (HPLC-MS).

Les composés seront ensuite utilisés en tant que sonde moléculaire pour étudier le métabolisme des lipides chez *M. tb*.

Les données préexistantes et déjà publiées seront également utilisées pour la validation des protéines cibles de nos composés. Une fois les protéines validées, elles seront cristallisées et leurs structures 3-D respectives seront déterminées par la méthode de diffraction au rayons-X.

Les données seront collectées par chaque scientifique impliqué dans le projet à l'aide de logiciel adapté à chaque instrument. Les protocoles expérimentaux sont conservés sous forme papier (cahiers de laboratoire) et version électronique (.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 fichiers FID bruts collectés sur les appareils Bruker peuvent être traités par n'importe quel logiciel de RMN, et seront disponibles sur demande. Les fichiers PDF contiendront à la fois des images des spectres originaux et des tableaux numériques détaillant le déplacement chimique, les constantes de couplage, les valeurs de masse et autres interprétations numériques des données.

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 partenaires du projet. Le volume des données de diffraction qui seront générées pendant ce projet est estimé à plusieurs To. Selon les directives H2020 les sites de rayonnement synchrotron doivent donner accès sur demande aux données de diffraction trois ans après leur acquisition. Les fichiers avec les coordonnées atomique des structure 3-D des protéines seront ensuite déposés à la Protein Data Bank (https://www.rcsb.org/) au format PDB.

Dans la mesure du possible des formats standards et ouverts seront privilégiés à des fins de partage et de réutilisation.

Description des données : exemples

Projet Atlas, Décrire, analyser et comprendre les effets de l'introduction de l'autodépistage du VIH en Afrique de l'Ouest à travers l'exemple du programme ATLAS
en Côte d'Ivoire, au Mali et au Sénégal
SHS

Collecte des données

Les différents volets de recherche et les collectes associées à chacun d'eux sont détaillés dans le protocole du projet consultable sur https://atlas.solthis.org/.

La nature des données collectées est de quatre types : (i) données d'enquêtes qualitatives (enregistrements audios ou prise de notes d'entretiens individuels et de groupe, en face à face ou par téléphone) ; (ii) données d'enquêtes téléphoniques quantitatives (enquête coupons) ; (iii) données d'enquêtes des temps et mouvements ; (iv) données financières (rapports financiers, collecte spécifique auprès de sites dispensateurs).

Données d'enquêtes qualitatives

Il s'agit d'entretiens individuels et de groupe. Les entretiens seront enregistrés (audio, format MP3 ou équivalent) avec le consentement des participants. Pour les entretiens individuels, si l'enquêté·e refuse l'enregistrement audio, l'enquêteur·rice procédera à une prise de notes manuelles.

Chaque entretien sera numéroté en fonction de l'enquête ou de la sous-enquête à laquelle il correspond et du pays de collecte. Le numéro d'entretien sera reporté sur le formulaire de consentement.

Les entretiens seront menés par des anthropologues seniors avec l'appui d'assistant·e·s de recherche formées aux sciences sociales et à la conduite d'entretien. L'équipe qualitative se réunira à intervalles réguliers pour coordonner les activités de terrain et vérifier la qualité des entretiens.

Les enregistrements audio et les prises de note feront l'objet d'une retranscription (fichier texte avec mise en forme) en français (avec traduction le cas échéant pour les entretiens menés dans une autre langue). Lors de la retranscription, les entretiens seront codés (changement des noms et prénoms par exemple) afin de les pseudonymiser.

Données d'enquêtes quantitatives téléphoniques

L'enquête téléphonique (WP Enquête Coupons) sera réalisée par la société Ipsos Côte d'Ivoire disposant d'une plateforme CATI (Collecte Assistée par Téléphone et Informatique) à Abidjan. Trois numéros verts seront mis en place (Côte d'Ivoire, Mali, Sénégal). Tous les appels seront routés vers le plateau CATI via des passerelles mises en place dans chaque pays (voix sur IP). Lors des phases de réception d'appels (phase 1 de l'enquête), les appels entrants seront automatiquement orientés vers un enquêteur ivoirien, malien ou sénégalais en fonction de la provenance de l'appel. Avant le démarrage de la collecte, les enquêteurs auront suivis une formation spécifique à l'enquête et à son questionnaire, menée à la fois par les superviseurs d'Ipsos Côte d'Ivoire et l'équipe de recherche.

Le programme PACCI, institution de recherche ivoirienne basée à Abidjan et partenaire du projet ATLAS, développera une plateforme dédiée à la capture des données, la gestion

Documentation et qualité des données

2. Documentation et qualité des données

2a. Quelles métadonnées et quelle documentation ?

Organisation des données

Structuration des dossiers)

Nommage des fichiers (ex : Doranum)

Gestion de versions

Description des données

Champs de description (ex : variables, les unités de mesure…)

Standards utilisés (ex DataCite, DublinCore)

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Par exemple, validation de données par l'ensemble des scientifiques impliqués dans le projet.

Documentation et qualité des données

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Une documentation est-elle nécessaire pour permettre la réutilisation des données (méthodologie de collecte, procédures et méthodes d'analyse, définition des variables et des unités de mesure…) ?	☐	☐	☐	☐
Métadonnées fournies pour aider à la recherche et à l'identification des données : comment ces métadonnées seront-elles produites ?	☐	☐	☐	☐
Quels seront les standards ou schémas de métadonnées utilisés pour renseigner les métadonnées ?	☐	☐	☐	☐
Un vocabulaire ou une taxonomie seront-t-ils utilisés pour décrire les données ?	☐	☐	☐	☐
Une procédure de contrôle qualité des données est-elle prévue ? ((joindre éventuellement le plan d'assurance qualité)	☐	☐	☐	☐

Documentation et qualité des données : exemples

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Les données seront systématiquement classées par dossiers, un dossier correspondant à une série d'expérience et nommé selon le format : AAAA-MM-DD.description succinte

Dans chaque dossier et dès que ce sera pertinent, un fichier "readme.txt" chapeautera le reste de l'arborescence pour donner les détails des acquisitions effectuées non enregistrés dans les fichiers de métadonnées automatiquement créés par les instruments utilisés.

Au sein de chaque dossier, chaque expérience sera rangée dans un sous-dossier avec une dénomination claire (à adapter en fonction de l'expérience)

Les fichiers d'analyse de données, les données analysées et les comptes-rendus d'expérience seront classés à part dans le même dossier, au sein d'un sous dossier "analyse" permettant de le séparer des données brutes.

Dans le cas de séries d'expérience, un tableur sera également mis à jour à chaque expérience avec les données expérimentales essentielles et des commentaires afin de permettre une navigation plus aisée entre les différents jeux de données.

Projet Hitrac, High-resolution traction force microscopy
Physique

2b. Quelles mesures de contrôle de la qualité des données seront mises en oeuvre ?

La fiabilité des expérience sera contrôlée par la mise en oeuvre systématique d'expériences identiques ou redondantes en accord avec les pratique de la communauté de microbiologie/mécanique/biophysique (selon le type d'expériences).

Documentation et qualité des données : exemples

Projet du laboratoire IAB, données de séquençage
Biologie

Les données sont accompagnées des **métadonnées** attendues par l'entrepôt de données NCBI/GEO en vue de leur publication.

Les métadonnées attendues sont décrites ici :
https://www.ncbi.nlm.nih.gov/geo/info/seq.html#metadata

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Recommandations:

- Expliquer comment la qualité et la conformité de la collecte des données seront contrôlées et documentées. Il s'agit là de préciser les processus comme la calibration, la répétition des échantillons ou des mesures, la capture standardisée des données, la validation de saisie des données, la revue par les pairs, ou la représentation basée sur des vocabulaires contrôlés.

Les données brutes (fichiers *fastq)* sont accompagnées de fichiers *md5* permettant de vérifier **l'intégrité** des fichiers transférés.

Par ailleurs, les pipelines utilisés intègrent à chaque niveau d'analyse des étapes de contrôle qualité.
Ce contrôle qualité se fonde principalement sur les outils *fastqc* et *multiqc*.

Documentation et qualité des données : exemples

Projet Atlas, Décrire, analyser et comprendre les effets de l'introduction de l'autodépistage du VIH en Afrique de l'Ouest à travers l'exemple du programme ATLAS
en Côte d'Ivoire, au Mali et au Sénégal
SHS

Données d'enquêtes qualitatives

Chaque retranscription d'entretien sera accompagnée d'un rapport d'entretien incluant : le contexte de passation de l'entretien, les caractéristiques sociodémographiques et une présentation de la ou des personnes interrogées et d'un résumé.

Données d'enquêtes quantitatives téléphoniques

Les fichiers analytiques générés à partir de la plateforme de collecte seront documentés selon la spécification DDI (*Data Documentation Initiative* http://www.ddialliance.org/) dédiée à la documentation de données d'enquêtes quantitatives en sciences humaines et sociales. La documentation des variables sera effectuée avec le logiciel *Nesstar Publisher* du *Norwegian Centre for Research Data*. Ce logiciel permet notamment de produire un dictionnaire détaillé des variables et un export des fichiers de données labellisées au format Stata et SPSS. Une copie des données au format CSV (fichiers texte) sera systématiquement conservée.

Données de l'enquête des temps et mouvements

Créé avec DMP OPIDoR. Dernière modification 22/07/2019

Les fichiers de l'enquête des temps et mouvements seront également documentés selon la spécification DDI.

Données financières

La collecte s'appuiera à la fois sur les différents documents de rapportage produits par la structure générant le site de dispensation (rapports financiers, rapports d'activités...) ainsi que une grille en cours de développement par l'équipe d'économistes.

Stockage et sauvegarde pendant le processus de recherche

3- Stockage et sauvegarde pendant le processus de recherche

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Modalités et lieu de stockage (type d'hébergement, support)
Procédure de sauvegarde
Utilisation de plateformes de sauvegarde et de partage institutionnelles

Sauvegarde et stockage des données au cours du projet

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Quels supports seront utilisés pour sauvegarder et stocker les données tout au long du projet ?	☐	☐	☐	☐
Dans quels formats les données seront-elles disponibles au cours du projet ?	☐	☐	☐	☐
Gestion et organisation des fichiers de données au cours du projet : - quelles conventions de nommage des fichiers ? - quelle structure des dossiers ? - selon quelle périodicité (gestion des versions) ?	☐	☐	☐	☐
Quelle est la volumétrie de données estimée ?	☐	☐	☐	☐
Où seront sauvegardées et stockées physiquement les données ? Sur quel type d'hébergement ?	☐	☐	☐	☐
Quelle sera la fréquence de sauvegarde (automatisée ?)	☐	☐	☐	☐
Où seront localisées géographiquement les données ?	☐	☐	☐	☐

Stockage et sauvegarde pendant le processus de recherche

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Modalités d'accès aux données pendant le projet
Droits d'accès (Pour les membres, les partenaires)

Quelle est la politique de sécurité des données de l'entité qui hébergera physiquement les données ?	☐	☐	☐	☐
Quelles mesures de protection des données seront mises en œuvre au cours du projet ?	☐	☐	☐	☐
Comment les données seront-elles récupérées en cas d'incident ?	☐	☐	☐	☐
Les données feront-elles l'objet d'échanges ou de partages avec des tiers acteurs ?	☐	☐	☐	☐
Des droits d'accès seront-ils nécessaires pour accéder aux données pendant les recherches, notamment en cas de partenariat ?	☐	☐	☐	☐

Stockage et sauvegarde pendant le processus de recherche : exemples

Projet du laboratoire IAB, données de séquençage
Biologie

Nous définissons ici trois espaces de stockage :

- un espace de stockage **sécurisé**, **privé** (*summer,* données froides et tièdes)
- un espace de stockage destiné au **calcul intensif** (*bettik,* données chaudes et tièdes)
- un espace d'**archivage** pérenne, public (l'entrepôt de données NCBI/GEO, données froides)

Au cours du processus de recherche les données sont stockées sur *summer* et *bettik*.

Les **données brutes** (*fastq,* niveau 1) que nous produisons sont stockées sur notre espace de stockage sécurisé hébergé par la plateforme *summer* de l'unité Grenoble Alpes Recherche - Infrastructure de Calcul Intensif et de Données (GRICAD UMS 3758). Cette solution de stockage et de sauvegarde est réparties dans trois centres de données de l'UGA situés sur le campus de Saint Martin d'Hères. L'infrastructure en place est basée sur les technologies du constructeur NetApp. Nous disposons de manière privative d'une volumétrie de 60 To (évolutif) associée à des fonctions évoluées assurant la sécurité, l'intégrité des données et la continuité de service.
Notre espace de stockage sécurisé est directement accessible depuis les serveurs du mésocentre de calcul CIMENT/GRICAD.

Les **données traitées** (niveau 2 et 3) sont stockées sur l'espace destiné au calcul intensif (*bettik)* du mésocentre CIMENT/GRICAD. *Bettik* est un espace de stockage distribué haute performance partagé (*scratch*) accessible depuis les clusters du mésocentre. Il offre une volumétrie conséquente (>1Po) partagée par l'ensemble des noeuds de calcul avec des temps d'accès minimaux et des débits élevés. *Bettik* utilise le système de fichier distribué *BeeGFS*.
Les données traitées sont produites par l'exécution de nos pipelines sur les infrastructures de calcul du mésocentre.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Comme précisé plus haut, les données brutes que nous générons (*fastq*) sont répliquées sur **3 sites distincts** ce qui permet d'assurer la continuité du service même en cas de sinistre important.

Notre espace de stockage est directement monté sur les serveurs de calcul du mésocentre CIMENT (GRICAD UMS 3758). L'accès aux données est contrôlé par la **gestion des droits** sous linux. Par défaut, uniquement les membres du groupe linux *epimedcore* (3 personnes, Sophie Rousseau, Ekaterina Flin et Florent Chuffart, EpiMed/IAB) peuvent accéder aux données. Si toutefois nous devions étendre cet accès à d'autres personnes (stagiaires, collaborateurs) nous pourrions le faire grâce à gestion avancée des **ACL** qui autorisent ou non l'accès aux fichiers à l'échelle de l'utilisateur linux.

Les données que nous traitons ne sont pas sensibles dans la mesure ou le ne comporte pas de caractère personnel. En conséquence, il n'y a pas de mise en œuvre particulière d'une politique institutionnelle.

Stockage et sauvegarde pendant le processus de recherche : exemples

Projet LipInTB, Utilisation de nouveaux inhibiteurs sélectifs pour décrypter le métabolisme lipidique et de la virulence chez M. tb

Biologie

Pour la durée du projet ANR, les données générées seront hébergés sur les postes de travail chiffrés des membres du projet LipInTB, et pour les données de protéomique et de cristallographie sur des serveurs dédiés et sécurisés. Des sauvegardes sur un serveur propre à chaque unité seront effectuées à intervalles réguliers. Les données de diffraction aux rayons X seront également sauvegardées par les sites de rayonnement synchrotron.

Les données seront stockées sur les postes de travail chiffrés de chaque responsable scientifique du projet, ou sur des serveurs dédiés et sécurisés pour les données de protéomique et de cristallographie.

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

L'ensemble des données sera stockée via le logiciel ProjectMeta (hébergé au LEGI) utilisant le protocole openDAP.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

- Pendant le projet, les données ne seront accessibles que par les partenaires du projet.
- A la fin du projet, l'ensemble des données sera ouverte à la communauté scientifique.
- Le projet ne contient aucune donnée sensible.
- Une copie des données de chaque utilisateur sera disponible auprès de chaque laboratoire.

Projet Palagram, PArticle-LAden GRAvity currents Modelling

Mécanique des fluides

Stockage et sauvegarde pendant le processus de recherche : exemples

Projet Atlas, Décrire, analyser et comprendre les effets de l'introduction de l'autodépistage du VIH en Afrique de l'Ouest à travers l'exemple du programme ATLAS en Côte d'Ivoire, au Mali et au Sénégal

SHS

Données d'enquêtes qualitatives

Les enregistrements audio des entretiens seront conservés sur l'ordinateur personnelle du/de la chercheur·e chargée de la conduite de l'entretien et protégés par mot de passe.

Les entretiens retranscrits et pseudonymisés seront déposés sur la plateforme https://cloud.atlasti.com/ dédiée à l'analyse collaborative de données qualitatives et notamment à la codification thématique d'entretiens. Cette plateforme est un produit développé et commercialisé par la société *Scientific Software Development GmbH* spécialisée dans les logiciels d'analyses qualitatives scientifiques. La plateforme est sécurisée et les accès individuels et privatifs. Selon la politique de confidentialité du service (https://atlasti.com/privacy/), la plateforme "ATLAS.ti Cloud" respecte le règlement général sur la protection des données (RGPD) de l'Union Européenne. Les communications seront chiffrées (https), les accès individuels (compte utilisateur) avec gestion des droits.

Données d'enquêtes quantitatives téléphoniques

La plateforme de collecte des données sera hébergée sur un serveur informatique sécurisé localisé dans les locaux du programme PACCI. L'accès à la plateforme (enquêteurs, superviseurs, chercheurs) se fera uniquement au travers d'une connexion chiffrée (https) et d'un compte utilisateur individualisé (les droits de chaque acteur pouvant être définis séparément).

Les analyses statistiques seront réalisées à partir de fichiers analytiques anonymisés extraits de cette plateforme. Les analyses seront effectuées avec le logiciel R et hébergées sur l'infrastructure de la TGIR Huma-Num (https://www.huma-num.fr/) qui mets à disposition un serveur R Studio Pro (https://r-tools.huma-num.fr) et un gestionnaire de versions Git (https://gitlab.huma-num.fr) situés sur des serveurs sécurisés du Centre national de la recherche scientifique (CNRS) en France. Les communications sont chiffrées (https), les accès individuels et la gestion des droits par utilisateur.

Exigences légales et éthiques, codes de conduite

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Formalités de déclaration de données sensibles (par exemple auprès du DPO, etc.) ?
Mesures prises pour garantir la confidentialité des données personnelles (pseudonymisation ou anonymisation, chiffrement, procédure d'accès…) ?
Modalités d'information des personnes (formulaires de consentement, etc.)

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière

Titularité des droits de propriété intellectuelle (données, code, etc.)
Licences d'usage des données ?
Respect des règles des différentes tutelles.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

Exigences légales et éthiques, codes de conduite

Exigences légales et éthiques, codes de conduite

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Qui sera titulaire des droits de propriété intellectuelle sur les données (qui aura le droit d'en contrôler l'accès) et sur les autres informations créées lors du projet ? Selon quelle législation ?	☐	☐	☐	☐
Un accord de consortium a-t-il été rédigé ?	☐	☐	☐	☐
Du matériel protégé par des droits spécifiques sera-t-il utilisé au cours du projet (ex : données personnelles, bases de données…) ?	☐	☐	☐	☐
Qui s'occupera des formalités à accomplir (consentement éclairé…), de l'obtention des autorisations d'utilisation et de diffusion éventuelles… ?	☐	☐	☐	☐
Quelles mesures seront prises pour garantir la confidentialité des données personnelles (pseudonymisation ou anonymisation, chiffrement, procédure d'accès…) et selon quelles normes?	☐	☐	☐	☐
Comment sont pris en compte les questions éthiques et le respect des codes déontologiques ?	☐	☐	☐	☐

Exigences légales et éthiques, codes de conduite : exemples

Projet Palagram,
PArticle-LAden
GRAvity currents
Modelling
Mécanique des fluides

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

- Aucune donnée sensible sur ce point

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

- Chaque expérimentateur/numéricien sera propriétaire par défaut des données qu'il/elle génère.
- Une fois mise sur openDAP, ces données seront partagés par les membres du consortium.
- La valorisation des données partagées ne peut être faite qu'en accord avec le propriétaire des données.
- A la fin du projet, sous réserve d'accord avec le propriétaire, les données seront partagées avec l'ensemble de la communauté scientifique.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

- Aucune donnée sensible sur ce point

Exigences légales et éthiques, codes de conduite : exer

Projet Atlas, Décrire, analyser et comprendre les effets de l'introduction de l'autodépistage du VIH en Afrique de l'Ouest à travers l'exemple du programme ATLAS en Côte d'Ivoire, au Mali et au Sénégal

SHS

Le protocole de recherche a été évalué et autorisé par 4 comités d'éthique :

- OMS : Ethical Research Committee, date :
- Côte d'Ivoire : Comité National d'Ethique des Sciences de la vie et de la santé, date : 28 mai 2019
- Mali : Comité d'Éthique de la Faculté de Médecine et de Pharmacie de l'université de Bamako , date :
- Sénégal : Comité National d'Ethique pour la Recherche en Santé, date :

L'ensemble des formulaires de consentement et des notices d'informations sont disponibles en annexe du protocole, sur le site https://atlas.solthis.org/

Les formulaires de consentement seront conservés pendant 5 ans après la fin du projet (soit jusque fin 2026) dans une armoire sous clé dans les locaux de Solthis situés à Abidjan pour les formulaires de consentement collectés en Côte d'Ivoire, Bamako pour les formulaires collectés au Mali et Dakar pour les formulaires collectés au Sénégal.

Données d'enquêtes qualitatives

Avant tout entretien en face à face, un formulaire de consentement éclairé sera signé par la personne enquêtée. Un consentement spécifique sera demandé pour les enregistrements audio. Afin de permettre aux enquêté·e·s d'exercer leur droit d'accès, de rectification et d'opposition, le numéro unique de l'entretien sera reporté sur le formulaire de consentement remis à l'enquêté.

Pour les entretiens menés par téléphone, le consentement sera demandé verbalement à l'enquêté et son anonymat préservé. Un formulaire de recueil du consentement verbal sera rempli par l'enquêteur·rice. Les données de contacts de la personne interrogée seront détruites juste après la passation de l'entretien téléphonique. Le numéro de l'entretien sera précisé à la personne enquêtée et il lui sera expliquée qu'elle doit le conserver si elle veut pouvoir exercer ultérieurement son droit d'accès, de rectification et/ou d'opposition.

Les enregistrements audio des entretiens seront conservés le temps de leur transcription et de leur pseudonymisation. Chaque chercheur·e s'engage à ce que les enregistrements audios soient détruits au plus tard à la fin du projet (fin 2021).

Comme mentionné sur les notices d'information, les données pseudonymisées seront conservées pour une durée maximum de 5 ans après la fin du projet, soit jusque fin 2026 au plus tard. Seules des données anonymisées feront l'objet d'un archivage de longue durée (voir le chapitre "Sélection et préservation").

Exigences légales et éthiques, codes de conduite : exemples

Projet LipInTB, Utilisation de nouveaux inhibiteurs sélectifs pour décrypter le métabolisme lipidique et de la virulence chez M. tb

Biologie

Non applicable.

Chaque partenaire respectera les règles de sa tutelle scientifique respective, en matière de propriété intellectuelle et de brevet (INSERM, CNRS ET AMU).

Tous les partenaires seront associés aux publications en fonction de leurs contributions respectives. Afin de promouvoir l'accessibilité, des publications en libre accès seront privilégiées, dans la mesure du possible. A l'issu du processus de publications et en respectant les politiques éditoriales des éditeurs, les publications issues du projet seront déposées dans une archive ouverte, libre d'accès (HAL, …)

Non applicable.

Partage des données et conservation à long terme

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Mode de diffusion des données :
Libre ? Restreint ? Embargo ?
Quels jeux de données librement accessibles ? (données liées aux publications?)
Lieu de diffusion des données (utilisation d'archives ouvertes comme HAL ou OpenEdition ?)
Délai de diffusion

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Utilité des données ('data utility') ?
Contexte, public, durée d'utilisabilité, etc.
Quelles données seront conservées sur le long terme ?
Procédures de la conservation à long terme (par ex : entrepôt certifié) ?
Utilisation d'une plateforme d'archivage pérenne (CINES en France) ?
Conditions d'archivage (budget nécessaire)
Responsabilité des données

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Utilisation d'un entrepôt de données pour l'obtention d'un DOI (par ex : Dataverse, ou Nakala)?

Partage des données et conservation à long terme

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Quels jeux de données du projet seront partagés (totalité des jeux de données ou sélection) ?	☐	☐	☐	☐
Y aura-t-il besoin de migrer vers un autre format (plus ouvert) pour partager les fichiers de données ?	☐	☐	☐	☐
Y a-t-il une obligation de partage ?	☐	☐	☐	☐
Y a-t-il une interdiction ou une restriction de partage ?	☐	☐	☐	☐
Y aura-il un embargo ? - *Pour quelles raisons (publication, protection de la propriété intellectuelle, dépôt de brevet…) ?* - *Pour quelle durée ?*	☐	☐	☐	☐
A partir de quand les jeux de données seront-ils disponibles ?	☐	☐	☐	☐
Pendant combien de temps les jeux de données seront-ils disponibles ?	☐	☐	☐	☐
Avec qui les données seront-elles partagées ? - Accès ouvert ou restreint ? - Une utilisation exclusive des données est-elle revendiquée et pour quelle raison ?	☐	☐	☐	☐
Comment chaque jeu de données sera-t-il partagé (dépôt dans un entrepôt, indexation dans un catalogue, …) ?	☐	☐	☐	☐
Quelle licence sera appliquée à chaque jeu de données ?	☐	☐	☐	☐
Un identifiant unique et pérenne (DOI ou autre) sera-t-il attribué à chaque jeu de données ?	☐	☐	☐	☐

Partage des données et conservation à long terme

En cas de partenariat, quel organisme se chargera de la demande d'identifiant ?	☐	☐	☐	☐
Un logiciel ou un outil spécifique sera-t-il nécessaire pour lire ou utiliser les données ? - *Si oui, lequel ?* - *Dans quel entrepôt seront déposés les scripts ?*	☐	☐	☐	☐
Quelles sont les réutilisations potentielles de ces données ?	☐	☐	☐	☐

Partage des données et conservation à long terme

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
Un plan de conservation des données a-t-il été établi ? - Quels jeux de données seront à conserver sur le moyen ou le long terme ? - Quels jeux de données seront à détruire (raisons contractuelles, légales ou réglementaires, intérêt scientifique, coûts…) ?	☐	☐	☐	☐
Une documentation complémentaire est-elle nécessaire pour permettre la réutilisation des données conservées à long terme (métadonnées supplémentaires, public autorisé…) ?	☐	☐	☐	☐
Quelle est la volumétrie estimée des jeux de données à conserver ?	☐	☐	☐	☐
Quelle sera la durée de conservation à long terme des jeux de données ?	☐	☐	☐	☐
Quelle plateforme d'archivage pérenne sera choisie ou quelles procédures seront mises en place pour la conservation à long terme ?	☐	☐	☐	☐
Quelles seront les garanties de financement pour couvrir les coûts associés à la conservation à long terme des jeux de données ?	☐	☐	☐	☐
Qui sera responsable de la conservation à long terme (nom, courriel, téléphone…) ?	☐	☐	☐	☐

Partage des données et conservation à long terme : exemples

Projet LipInTB, Utilisation de nouveaux inhibiteurs sélectifs pour décrypter le métabolisme lipidique et de la virulence chez M. tb

Biologie

Les résultats obtenus seront publiés dans des revues à comité de lecture appropriées. A l'issu du processus de publications et en respectant les politiques éditoriales des éditeurs, les publications issues du projet seront déposées dans une archive ouverte, libre d'accès (HAL, …)

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 (www.proteomexchange.org).

L'ensemble des données générées seront uniques et devront donc être conservées sans mention de délai.

Les données cristallographiques et protéomiques seront conservées sur des entrepôts de données accessibles à tous (PRIDE-EBI pour la protéomique et RSCB Protein Data Bank pour la cristallographie).

Dans la mesure du possible des formats standards et ouverts seront privilégiés à des fins le partage et de réutilisation.

Non applicable.

Partage des données et conservation à long terme : exemples

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

- openDAP permet le stockage des données de façon durable (10 ans).
- Une partie des données sera également déposée sur une plateforme type zenodo en fin de projet.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

- l'ensemble des données sera conservée sur openDAP pour une durée de 10 ans.
- Une partie des données sera également déposée sur une plateforme type zenodo.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

- Des exemples de lecture des fichiers métadonnées sous différents formats seront mis à disposition.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

- Les données zenodo sont identifiées par un DOI.

Projet Palagram,
PArticle-LAden
GRAvity currents
Modelling
Mécanique des fluides

Partage des données et conservation à long terme : exemples

Projet Hitrac, High-resolution traction force microscopy
Physique

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Les données ne seront partagées qu'après publication ou dépôt de brevet, leur diffusion restant limitée aux membres de l'équipe avant cela. Sauf demande explicite de l'éditeur lors de la dissémination des résultats sous forme de publication, les données seront partagées par traitement direct des demandes de données.

L'archivage sur bandes magnétiques permet une fiable conservation de long terme (plusieurs années ou dizaines d'années), tout en limitant leur impact environnemental.

La réutilisation des données à des fins commerciales ou de publication ne pourra se faire qu'avec l'accord écrit de la coordinatrice, et en citant le(s) publication(s) initiale(s).

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Afin d'éviter tout biais dans la sélection des données, toutes les données exploitables seront conservées. En l'absence de lieu de stockage dédié à l'Université Grenoble Alpes, les bandes seront stockées en doubles exemplaires dans deux laboratoires différents de l'université.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Les demandes attendues sont faibles, il est donc prévu de traiter les demandes en direct en partageant les données demandées via des site de partage de données volumineuses (e.g. FileSender/Renater)

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Non pertinent pour ce projet.

Partage des données et conservation à long terme : exemples

Les données personnelles (données nominatives ou données permettant l'identification des personnes par recoupement d'information) seront conservées au maximum pendant 5 ans après la fin du projet, soit jusque fin 2026 au plus. De même, les formulaires de consentement seront conservés pendant la même période afin de permettre aux personnes enquêtées d'exercer leur droit d'accès, de rectification et/ou d'opposition (voir section 3.1).

Seules des données anonymisées générées à partir des données brutes feront l'objet d'un archivage.

Données d'enquêtes qualitatives

Une procédure d'anonymisation des retranscriptions pseudonymisées des entretiens sera mise en place. Les rapports d'entretiens seront repris un à un et tous les passages contenant des informations susceptibles d'identifier la personne interrogée ou celles qu'elle mentionne seront supprimés (avec indication de la suppression). En cas de doute, la suppression sera favorisée. Les rapports d'entretiens anonymisés seront validés collégialement par les équipes de recherche qualitatives.

Données d'enquêtes quantitatives téléphoniques

L'anonymisation des données sera effectué au moment de la génération de fichiers de données analytiques à partir de la plateforme de collecte des données.

Données de l'enquête des temps et mouvements

Des fichiers analytiques anonymisés seront produits en fin de projet. Dans la mesure où l'enquête a été menée auprès d'une population spécifique (agent·e·s dispensateurs) de taille réduite, des outils d'aide à l'anonymisation des données tels que *ARX Data Anonymization Tool* (https://arx.deidentifier.org/) seront utilisés pour guider la sélection et la recodification des variables.

Projet Atlas, Décrire, analyser et comprendre les effets de l'introduction de l'autodépistage du VIH en Afrique de l'Ouest à travers l'exemple du programme ATLAS en Côte d'Ivoire, au Mali et au Sénégal

SHS

Les données déposées sur Zenodo disposeront chacune d'un DOI (Digital Object Identifier) permettant leur identification pérenne au cours du temps. Ces DOI seront notamment utilisés dans les publications scientifiques générées à partir de ces données.

Pendant la durée du projet, les fichiers de données seront déposés en accès à la demande, le temps que les équipes de recherche publient les résultats de leurs analyses.

Une fois le projet terminé et les articles correspondants publiés, les fichiers seront en accès libre et distribués sous licence *Creative Commons - Attribution - Partage dans les Mêmes Conditions 4.0 International* (https://creativecommons.org/licenses/by-sa/4.0/).

Responsabilités et ressources en matière de gestion des données

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

Responsable de la gestion des données au sein du projet

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Le budget (frais de stockage, coût de traitement, coût matériel, d'archivage)
Les ressources humaines (métier)

Responsabilités et ressources en matière de gestion des données

	Réponse complète et détaillée	Réponse renseignée mais incomplète	Non renseignée	Non concerné
La ou les personne(s) responsable(s) de la mise en œuvre du PGD sont-elles bien identifiée(s) ? Qui est responsable - De l'examen du PGD ? - Des mises à jour ? - Du respect des délais / financeur ?	☐	☐	☐	☐
Comment s'effectue la coordination des responsabilités de gestion des données dans le cas de partenariats ?	☐	☐	☐	☐
Les rôles respectifs des partenaires sont-ils bien précisés ? *Nom, affiliation, rôle et coordonnées des personnes responsables des différentes modalités de gestion des données :* - *saisie des données,* - *production des métadonnées,* - *qualité des données,* - *sauvegarde et stockage,* - *partage,* - *archivage,* - *…*	☐	☐	☐	☐
La répartition des ressources est-elle détaillée ? - budget alloué : frais de stockage pendant le projet, coût de matériel, frais de partage dans un entrepôt, coût d'archivage pérenne - ressources humaines affectées - temps estimé pour chaque étape du projet : préparation des données et métadonnées pour le dépôt dans un entrepôt et pour l'archivage pérenne…	☐	☐	☐	☐

Responsabilités et ressources en matière de gestion des données : exemples

Projet Palagram, PArticle-LAden GRAvity currents Modelling
Mécanique des fluides

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

- La gestion des données sera faite par tous les membres du consortium. Un responsable qualité sera désigné dans chaque laboratoire.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

- Le serveur de stockage est hébergé par le LEGI. L'achat d'espace stockage sera faite de façon collaborative par les partenaires du projet.
- le service informatique du LEGI est en charge du maintien de la plateforme.
- Chaque membre est responsable, le représentant local s'assurera de la qualité des données.

Projet LipInTB, Utilisation de nouveaux inhibiteurs sélectifs pour décrypter le métabolisme lipidique et de la virulence chez M. tb
Biologie

Chaque responsable scientifique sera en charge de l'archivage des données le concernant.

Chaque responsable scientifique sera en charge de la gestion des données le concernant.

Responsabilités et ressources en matière de gestion des données : exemples

Projet Hitrac, High-resolution traction force microscopy
Physique

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

La coordinatrice du projet est responsable de la rédaction, mise à jour, mise en oeuvre du DMP et de son respect par les personnes impliquées dans le projet. L'archivage se fera avec le support de la cellule informatique du laboratoire.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Le coût de stockage sur bandes magnétiques estimé pour l'ensemble du projet est faible (quelques centaines d'euros), et aucun autre coût n'est associé à l'archivage des données. Nous prévoyons un nombre de demandes faible voire inexistant pour le type de données que nous archiverons, le temps consacré à l'archivage et à la mise à disposition de données est donc de l'ordre d'1 jour/an. La classification et le stockage redondant des données est lui géré au jour le jour et intégré dans le temps nécessaire aux expériences et au traitement de données. Des disques durs résistants aux chocs sont d'ores et déjà disponibles pour le projet, de même que les espaces de stockage en ligne mise à disposition gratuitement par nos tutelles.

Chaque work package est coordonné par un responsable scientifique (voir protocole) qui est en charge de s'assurer de la bonne conduite des activités de son work package.

L'investigateur principal s'assurera de la bonne exécution globale du plan de gestion de données.

Un organigramme détaillé des rôle et responsabilités sera développé en début de projet.

Les ressources humaines, matérielles et logicielles requises pour la mise en œuvre du plan de gestion des données ont été prévues et inclues dans le budget du projet.

Projet Atlas, Décrire, analyser et comprendre les effets de l'introduction de l'autodépistage du VIH en Afrique de l'Ouest à travers l'exemple du programme ATLAS en Côte d'Ivoire, au Mali et au Sénégal
SHS

Conclusions de l’atelier

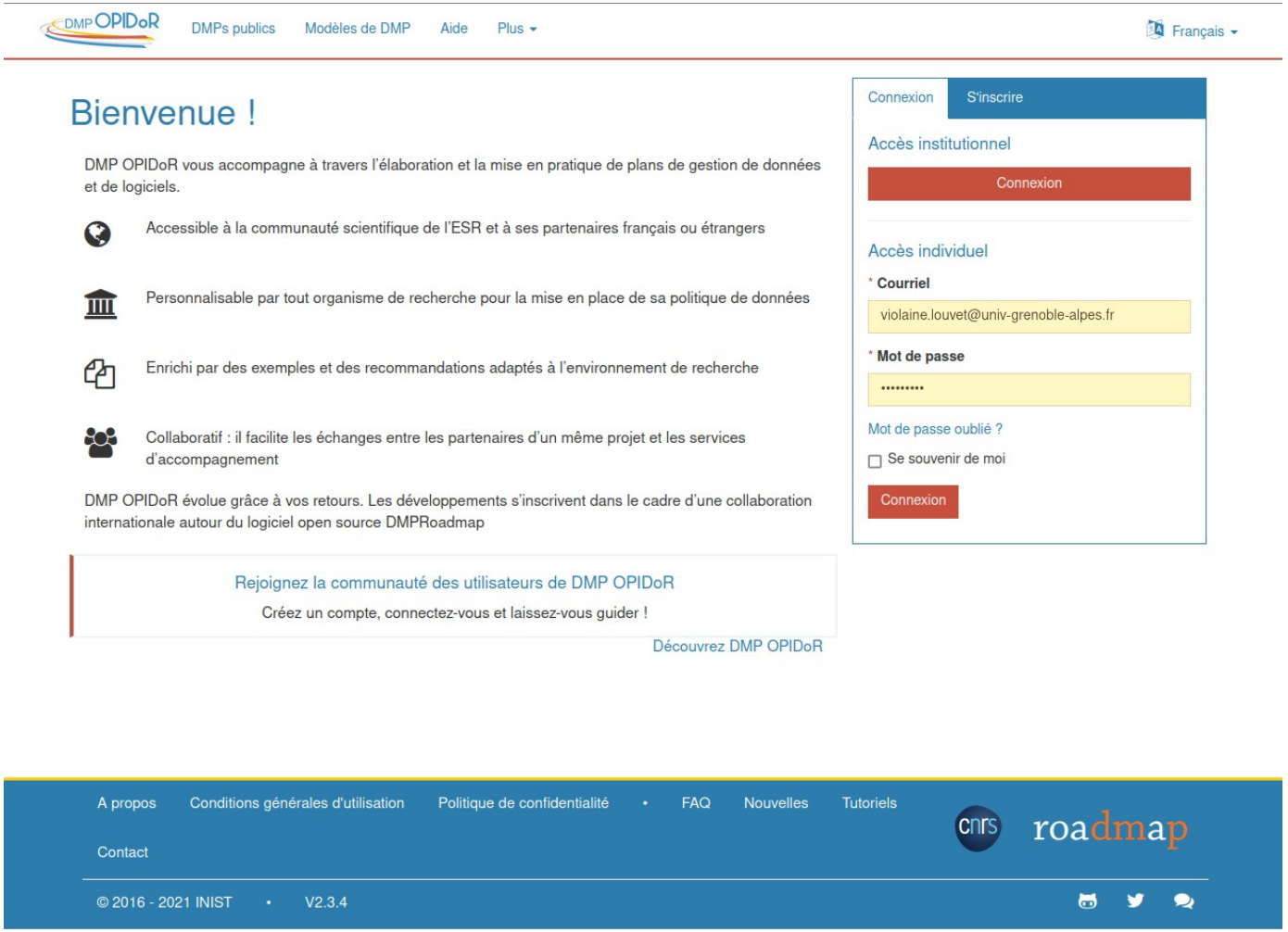
- On reforme un seul groupe présentiel / distanciel
- Les rapporteurs font une rapide synthèse des échanges de chaque groupe

Ce qu’il faut retenir

- Chaque projet scientifique est particulier : chaque PGD l’est aussi !
- Toutes les sections ne sont pas forcément pertinentes pour tous les projets
- L’objectif n’est pas juste la diffusion à l’issue du projet mais bien la gestion des données tout au long du projet
- Toutes les données ne sont pas forcément à ouvrir et à diffuser !
- L’idée est de permettre aux scientifiques de s’approprier les bonnes pratiques et de les mettre en oeuvre autant que possible

DMP Opidor

- Un outil d'aide à la création de PGD
- Data Management Plan pour une Optimisation du Partage et de l'Interopérabilité des Données de la Recherche :

  https://dmp.opidor.fr

- Développé par l'INIST (CNRS)