PGD CyPS (UMS37-PASS)

Organization	Institut Français de Bioinformatique
Created by	Cytométrie Pitié-Salpêtrière
Based on	PGD structure développé par l'IFB, 2.3.13
Project Phase	Avant soumission du PGD
Created at	23 Oct 2023

I. Introduction

Les objectifs de ce plan de gestion des données (PGD):

- Accompagner les plateformes et leurs utilisateurs dans le cycle de vie de leurs données ;
- PGD structure et PGD projet : nous proposons ici les champs à renseigner pour le PGD structure, accompagnés de recommandations ; ce PGD structure a pour but de venir renseigner (automatiquement, voir le prochain point) certains éléments des futurs PGDs projet des utilisateurs de la structure ;
- Machine actionable data management plan (maDMP) : ce PGD est proposé dans un format qui se veut compatible avec le projet maDMP de l'INIST et de l'IFB, qui permettra de réutiliser et partager facilement les briques de description (par exemple pour la mise en place de l'espace de stockage d'un projet, pour nourrir le PGD d'un projet...), afin d'éviter la double saisie et s'assurer que les éléments pertinents sont connus des acteurs concernés.

Dans ce contexte, la structure est définie d'une part comme l'infrastructure qui est génératrice de données à la demande des projets de recherche, et d'autre part, comme l'entité qui est la garante de la pérénnité des données une fois le projet arrivé à son terme.

Le PGD structure a pour vocation de servir ensuite de modèle à tout PGD de projet porté par l'infrastructure.

Le document qui suit décrit les éléments du cycle de vie de la donnée numérique, c'est-à-dire, les processus de création, d'identification, de documentation, de partage et d'archivage. Par souci de simplification pour cette première version du PGD structure, certaines caractéristiques constitutives sont imposées et certains types de données en sont exclus à ce stade. Néanmoins ce document, évolutif et destiné à être modulaire, s'enrichira en fonction des exigences et recommandations de diverses entités tutélaires, et des besoins émanant des utilisateurs actuels et potentiels.

L'objectif premier du PGD structure est d'accompagner les données de tous renseignements permettant ainsi de les rendre ainsi visibles, accessibles, et réutilisables quel que soit le demandeur. Cet objectif s'inscrit dans le paradigme FAIR (https://www.go-fair.org/fair-principles/), soit « Facile à découvrir, Accessible, Interopérable, Réutilisable », qui est l'un des piliers de la science ouverte.

Un autre objectif est l'automatisation de la collecte et de la diffusion des informations pertinentes aux données, afin d'optimiser ce processus fastidieux, de le rendre plus robuste en limitant l'intervention humaine, et, enfin, de garantir une visibilité maximale par l'exposition du cadre qui gouverne la réutilisation des données.

Ce projet « PGD structure » est une extension du projet « PGD structure bioimagerie », porté par et réalisé conjointement avec les Infrastructure Nationale en Biologie Santé (INBS) « France-BioImaging » (FBI), « Institut Français de Bioinformatique » (IFB), et « Centre National De Ressources Biologiques Marines » (EMBRC-France).

II. Informations générales

1. Informations sur la structure

Comment se nomme la structure ?

✓ Plateforme de Cytométrie Pitié-Salpêtrière

Si pertinent, quel est son acronyme ?

✓ CyPS

Si pertinent, listez ici l'URL (ou les URLs) de la structure

✓ http://www.cytometrie.pitie-salpetriere.upmc.fr/

Est-ce l'URL principale ?

✓ b. Oui

Si la structure est répertoriée dans ROR (Research Organization Registry), indiquez son entrée

✓ Sorbonne University (ROR (organisations) : https://ror.org/02en5vm52)

Quel est son type ?

✓ Plateforme d'acquisition

La structure comprend-t-elle plusieurs sites ?

✓ a. Non

La structure (ou service) fait-elle partie d'une infrastructure ou une structuration locale/régionale/nationale/internationale ?

✓ a. Non

Quelles en sont les tutelles ?

✓ Sorbonne University et Inserm

La structure (ou service) a-t-elle un ou plusieurs identifiants ?

✓ a. Non

La structure (ou service) est-elle labellisée ou certifiée ?

✓ b. Oui

Label(s) ou certification(s)

✓ GIS IBiSA

2. Personnes et/ou entités assumant des rôles et responsabilités pertinents à la gestion des données

a. S'agit-il d'une personne ou d'une organisation ?

✓ b. Organisation

Nom

✓ Plateforme de Cytométrie Pitié-Salpêtrière

Acronyme

✓ CyPS

Courriel

✓ cyps@sorbonne-universite.fr

L'organisation est-elle référencée dans Research Organization Registry (ROR) ?

✓ a. Non

A-t-elle d'autres référencements que ROR ?

✓ b. Oui

Type de l'identifiant

✓ b. HAL

Identifiant

✓ PASS-CYPS 1005723

Quel est le rôle ou responsabilité de l'organisation ?

✓ Plateforme d'acquisition de données

b. S'agit-il d'une personne ou d'une organisation ?

✓ a. Personne

Nom

✓ VINIT

Prénom

✓ Angélique

Courriel

✓ angelique.vinit@sorbonne-universite.fr et cyps@sorbonne-universite.fr

Quel est le rôle de la personne ?

✓ Responsable du plan de gestion de données
✓ Responsable de la production ou de la collecte des données
✓ Responsable de la gestion des échantillons
✓ Responsable de la gestion des instruments
✓ Responsable du traitement et de l'analyse des données
✓ Responsable du stockage des données
✓ Responsable du dépôt et de la diffusion des données

c. S'agit-il d'une personne ou d'une organisation ?

✓ a. Personne

Nom

✓ HOAREAU

Prénom

✓ Bénédicte

Courriel

✓ benedicte.hoareau@sorbonne-universite.fr

Quel est le rôle de la personne ?

✓ Responsable de la production ou de la collecte des données
✓ Responsable de la gestion des échantillons
✓ Responsable de la gestion des instruments
✓ Responsable du traitement et de l'analyse des données
✓ Responsable du stockage des données
✓ Responsable du dépôt et de la diffusion des données

d. S'agit-il d'une personne ou d'une organisation ?

✓ a. Personne

Nom

✓ CORNEAU

Prénom

✓ Aurélien

Courriel

✓ aurelien.corneau@sorbonne-universite.fr

Quel est le rôle de la personne ?

✓ Responsable de la production ou de la collecte des données
✓ Responsable de la gestion des échantillons
✓ Responsable de la gestion des instruments
✓ Responsable du traitement et de l'analyse des données
✓ Responsable du stockage des données
✓ Responsable du dépôt et de la diffusion des données

e. S'agit-il d'une personne ou d'une organisation ?

✓ a. Personne

Nom

✓ BLANC

Prénom

✓ Catherine

Courriel

✓ catherine.blanc@sorbonne-universite.fr

Quel est le rôle de la personne ?

✓ Responsable opérationnel de la plateforme

3. Information sur le PGD

Le PGD a-t-il ou aura-t-il un identifiant ?

✓ b. Oui, plus tard

Quelles sont la/les personne(s) impliquée(s) dans la rédaction du PGD ?

✓ Angélique VINIT Bénédicte HOAREAU

Une licence est-elle ou sera-t-elle attribuée au PGD ?

✓ a. Non

III. Données de la recherche

1. Les produits de la recherche

a. Nom

✓ **Données de cytométrie conventionnelle, spectrale et de masse**

Description succinte

✓ Fichiers fcs, lmd et pdf

Le type du(des) produit(s) de la recherche

✓ Jeux de données cytométrie

✓ Jeu de données autres

Le produit de la recherche a-t-il ou aura-t-il un identifiant ?

✓ a. Non, cela n'est pas pertinent

b. Nom

✓ **Données de cytométrie de masse en image**

Description succinte

✓ Fichiers mcd, txt, fcs, pdf, tiff et png

Le type du(des) produit(s) de la recherche

✓ Image

✓ Jeux de données cytométrie

✓ Texte

Le produit de la recherche a-t-il ou aura-t-il un identifiant ?

✓ a. Non, cela n'est pas pertinent

2. Description des produits de la recherche

Listez ici les données produites par la structure

✓ Tous

a. Listez ici le(les) produit(s) de la recherche concerné(s) par la question

✓ **Données de cytométrie conventionnelle, spectrale et de masse**

Pour information et si pertinent, quel est le niveau de traitement des données ?

✓ Données brutes

✓ Données primaires

✓ Données analysées

S'agit-il d'images ?

✓ a. Non

S'agit-il de jeux de données cytométrie ?

✓ b. Oui

Quelle est la modalité d'acquisition ?

✓ Cytométrie en flux conventionnelle

✓ Cytométrie en flux spectrale

✓ Cytométrie de masse (CyTOF)

✓ Tri cellulaire

Quels logiciels, technologies ou processus sont utilisés pour générer ou collecter les données ?

Pour les données brutes

✓ Les données brutes sont collectées et générées sur les appareils suivants en utilisant les logiciels ci-précisés: - BD FACSAria avec le logiciel BD FACSDiva, - Aurora de Cytek avec le logiciel SpectroFlo - CyTOF XT de Standard BioTools avec le logiciel CyTOF software

Pour les données primaires

✓ Les données primaires sont obtenus par le logiciel CyTOF software sur la base des données brutes.

Pour les données analysées

✓ Les données analysées sont obtenues par analyse des données brutes ou primaires générées par les logiciels: BD FACSDiva, Spectroflo, FlowJo, Maxpar Pathsetter, Omiq ou Cytobank

Quelles mesures de contrôle de qualité sont prises ?

✓ Contrôle qualité sur machines

Quelles sont les actions menées pour permettre la re-création des données à l'identique ou à l'équivalent ?

✓ Utilisation d'un protocole standardisé de génération des données
✓ Tenue d'une traçabilité par l'opérateur (e.g. cahier de laboratoire, notebook, etc...)
✓ Utilisation de standards dans la communauté (outils, technologies, formats, etc...)
✓ Utilisation d'un environnement logiciel empaqueté (application et dépendances)
✓ Rappel d'un réglage préalablement enregistré sur le logiciel constructeur de l'instrument

S'agit-il de jeux de données cytométrie ?

✓ b. Oui

Quels sont les formats de fichiers utilisés ?

✓ .fcs

Pour les formats autres que les formats standards : précisions supplémentaires

Quel est le nom du format ?

✓ Fichier lmd

Quel est le suffixe du format ?

✓ .lmd

Pourquoi conserver un format non standard ?

✓ Format constructeur de données brutes

b. Listez ici le(les) produit(s) de la recherche concerné(s) par la question

✓ **Données de cytométrie de masse en image**

Pour information et si pertinent, quel est le niveau de traitement des données ?

✓ Données brutes
✓ Données analysées

S'agit-il d'images ?

✓ b. Oui

Quelle est la modalité d'acquisition ?

✓ Cytométrie de masse en image (Hyperion)

Quels logiciels, technologies ou processus sont utilisés pour générer ou collecter les données ?

Pour les données brutes

✓ CyTOF Software

Pour les données analysées

✓ MCD Viewer, Visiopharm, Halo, Steinbock ou R

Quelles mesures de contrôle de qualité sont prises ?

✓ Contrôle qualité sur machine (lame de calibration)

S'agit-il de jeux de données cytométrie ?

✓ b. Oui

Quelle est la modalité d'acquisition ?

✓ Cytométrie de masse en image

Quelles mesures de contrôle de qualité sont prises ?

✓ Contrôle qualité sur machines

Quelles sont les actions menées pour permettre la re-création des données à l'identique ou à l'équivalent ?

✓ Utilisation d'un protocole standardisé de génération des données
✓ Tenue d'une traçabilité par l'opérateur (e.g. cahier de laboratoire, notebook, etc...)
✓ Utilisation de standards dans la communauté (outils, technologies, formats, etc...)
✓ Utilisation d'un environnement logiciel empaqueté (application et dépendances)

Quels sont les formats de fichiers utilisés ?

✓ OME-TIFF ou autre format standard supporté par Bio-Formats (DICOM par exemple), ou tout autre logiciel libre
✓ Format constructeur

S'agit-il de jeux de données cytométrie ?

✓ b. Oui

Quels sont les formats de fichiers utilisés ?

✓ .txt

Pour les formats autres que les formats standards : précisions supplémentaires

Quel est le nom du format ?

✓ Format mcd

Quel est le suffixe du format ?

✓ .mcd

Pourquoi conserver un format non standard ?

✓ Format constructeur de données brutes

3. Documentation et qualité des données

Listez ici le(les) produit(s) de la recherche concerné(s) par la question

✓ **Tous**

De la liste ci-dessous, comment caractérisez-vous les annotations mises en oeuvre ?

✓ Les données sont décrites avec un vocabulaire contrôlé

Quelles sont les méthodes d'organisation des données ?

✓ Organisation par équipement (appareil) puis par utilisateur et par date d'acquisition

Quelles sont les méthodes de nommage des fichiers ?

✓ Date(AnnéeMoisJour)_Nom(ChoisiParUtilisateur)

Les composants individuels du jeu de données sont-ils liés les uns aux autres ?

✓ a. Non

Le jeu de données est-il lié à d'autres produits de la recherche ?

✓ b. Oui

Donnez une description succinte

✓ Les jeux de données analysées (fcs, pdf, png, etc) sont liés aux données brutes (fcs, mcd et txt)

Des versions différentes du jeu de données sont-elles créées ?

✓ a. Non

Quel type de métadonnées sont-elles mises en œuvre ?

✓ Métadonnées décrivant l'acquisition de la donnée brute (SOP)

✓ Métadonnées décrivant la donnée brute

Comment les métadonnées sont-elles créées, collectées et tracées ?

✓ Les métadonnées sont créées lors de la création des données sur les instruments. Sont retenues à minima : la date d'acquisition de l'échantillon ou la date de son analyse avec le nom de l'échantillon choisi par l'utilisateur. Les données sont ensuite positionnées dans dossier de l'utilisateur associé.

Qui est chargé de documenter les métadonnées et les informations de contexte et de vérifier si elles sont correctes et complètes ?

✓ Angélique Vinit ou Bénédicte Hoareau en fonction de qui fait l'acquisition ou l'analyse d'un fichier

Allez-vous attribuer des identifiants persistents (PIDs) ?

✓ a. Non

Pour quelle raison ?

✓ Les données sont récupérées par les utilisateurs de la plateforme pour production scientifique. Ces publications sont écrites par les porteurs de projet et ce sont donc les scientifiques qui prendront la responsabilité d'attribuer des PIDs si nécessaire.

4. Gestion des données chaudes et tièdes

a. Listez ici le(les) produit(s) de la recherche concerné(s) cette question

✓ **Données de cytométrie conventionnelle, spectrale et de masse**

Quelle est la taille estimée ?

✓ 25 TB

Utilisez-vous d'autres unités que celle de la taille ?

✓ b. Oui

S'agit-il de jeux d'images ?

✓ a. Non

S'agit-il de jeux de données cytométrie ?

✓ b. Oui

Autre unité

✓ Nombre de tubes acquis. Environ 2300 tubes par an.

Avez-vous des commentaires à faire au sujet de la taille des données ?

✓ Par unité, les fichiers fcs peuvent faire de quelques MB à plusieurs GB. Cette variation importante dans la taille des fichiers dépend du nombre de marqueurs utilisés (entre 1 et 40) et du nombre d'évènements acquis (entre quelques milliers et plusieurs millions). Ces paramètres sont défini par les utilisateurs de la plateforme pour chacun de leurs échantillons.

b. Listez ici le(les) produit(s) de la recherche concerné(s) cette question

✓ **Données de cytométrie de masse en image**

Quelle est la taille estimée ?

✓ 2 TB

Utilisez-vous d'autres unités que celle de la taille ?

✓ b. Oui

S'agit-il de jeux d'images ?

✓ b. Oui

Autre unité

✓ Nombre d'images. Environ 400 images par an.

S'agit-il de jeux de données cytométrie ?

✓ a. Oui

Avez-vous des commentaires à faire au sujet de la taille des données ?

✓ Par unité, les fichiers mcd peuvent faire de quelques MB à quelques GB. Cette variation dans la taille des fichiers dépend du nombre de marqueurs utilisés (entre 15 et 35 en moyenne) et de la taille de la zone acquise (entre une centaine de µm2 et plusieurs mm2). Ces paramètres sont définis par les utilisateurs de la plateforme pour chacun de leurs échantillons.

c. Listez ici le(les) produit(s) de la recherche concerné(s) par la question

✓ **Tous**

Oú les données sont-elles stockées ?

✓ Machine d'acquisition

✓ Serveur de stockage

Avez-vous des commentaires à faire à ce sujet ?

✓ Une fois créées les données brutes et primaires sont transférées sous 1 à 3 jour de la machine d'acquisition vers un serveur de stockage.

De quel support de stockage parle-t-on ?

✓ Machine d'acquisition

✓ Serveur de stockage

Quelles sont les stratégies de sauvegarde employées ?

✓ Sur la machine d'acquisition les données ne sont conservées que 1 jour à 1 semaine. Ensuite elles sont transférées sur 2 serveurs différents tous 2 hors sites, soit 2 copies.

Quelle est la fréquence des sauvegardes ?

✓ Tous les jours de la semaine

✓ Chaque mois

Quelle est la durée de conservation des sauvegardes ?

✓ Les données sont transmises dans la semaine suivant l'acquisition à l'utilisateur. Elles sont ensuite conservées au moins 6 mois sur nos serveurs de sauvegarde.

Quelles sont la/les personne(s)/entité(s) responsable(s) des sauvegardes ?

✓ Angélique Vinit ou Bénédicte Hoareau selon la personne qui a fait l'acquisition

Avez-vous d'autres commentaires à faire au sujet de la sauvegarde ?

✓ Une sauvegarde quotidienne est programmée pour les données de cytométrie spectrale et de masse (en flux et en image). Une sauvegarde mensuelle est effectuée pour les données de cytométrie conventionnelle.

Quelles sont les mesures ou dispositions mises en place pour garantir la sécurité des accès aux données ?

✓ Les données sont stockées sur des serveurs dont l'accès est protégé par un mot de passe. Les données sont transférées (par lien, par mail ou par filesender) uniquement à l'utilisateur et au responsable du projet si ce n'est pas la même personne.

De quel support de stockage parle-t-on ?

✓ Serveur de stockage

Les données seront-elles partagées durant le temps de traitement des données ?

✓ b. Oui

Quelles sont les modalités de partage ?

✓ Limité en externe, avec approbation individuelle

Avez-vous des commentaires à faire au sujet des coûts ?

✓ Pour le serveur installé actuellement il y a eu 12 000 euros d'investissement et environ 2 000 euros de maintenance annuelle. Un transfert sur un serveur institutionnel est envisagé pour un coût annuel de 1 000 euros en moyenne.

Quel est le devenir de ces données ?

✓ a. Elles vont être transférées au donneur d'ordre

Sinon décrivez la modalité de ce transfert

✓ Elles sont transférées grâce à un lien sur un serveur par utilisateur ou par transfert des données par Filesender.

5. Gestion des données froides

Listez ici le(les) produit(s) de la recherche concerné(s) par la question

✓ **Tous**

Quels sont les critères et/ou les règles de sélection des données à entreposer ou archiver ?

✓ Données de projets de recherche publiques

✓ Les données sont utilisées dans une publication

Avez-vous des commentaires à faire au sujet de la date d'entreposage ou d'archivage ?

✓ Nous n'archivons pas les données. Au-delà de 6 mois, nous pouvons être amené à prévenir les utilisateurs d'une date de suppression. Si le volume de stockage nous le permet, nous les gardons au moins 1 an sur le serveur.

Oú les données seront-elles stockées ? Indiquez l'URL

✓ http://134.157.199.131:5000/

L'entrepôt est-il certifié ? Si oui, indiquez la certification

✔ a. Non

Pour quelle raison l'entrepôt n'est-il pas certifié ?

✔ Le serveur est présent dans la pièce prévue à cet effet par l'université.

Oú seront stockées les métadonnées ?

✔ a. Les métadonnées sont stockées au même endroit que les données

Quelle est la durée de stockage envisagée ?

✔ Plusieurs années

Les données seront-elles partagées ?

✔ b. Oui

Quelles sont les modalités de partage ?

✔ Limité en externe, avec approbation individuelle

Si pertinent comment l'identité de la personne accédant aux données sera-t-elle vérifiée ?

✔ Les données sont envoyés par mail au porteur de projet concerné

Y aura-t-il un embargo ?

✔ a. Non, les données sont disponibles inconditionnellement

À partir de quand les données seront-elles accessibles ?

✔ a. Dès leur stockage en entrepôt

Quelles sont les restrictions gouvernant le partage des données ?

✔ a. Il n'y a aucune restriction

Avez-vous des commentaires à faire au sujet des coûts ?

✔ Actuellement les coûts de stockage correspondent aux frais de maintenance de serveur soit entre 1 000 et 2 000 euros par an.

6. Questions autour du cadre légal et éthique

Listez ici le(les) produit(s) de la recherche concerné(s) par la question

✔ **Tous**

Quel est le cadre juridique applicable ?

✔ a. La loi Française

Ces données contiennent-elles des données personnelles ?

✔ a. Non

Ces données contiennent-elles des données sensibles ?

✔ a. Non

Les données sont-elles protégées par des droits de propriété intellectuelle ou industrielle ?

✔ a. Non

La loi sur le droit d'auteur s'applique-t-elle à ce jeu de données ?

✔ a. Non

Quels sont les produits de la recherche financés au moins pour moitié par des fonds publics ?

✔ Tous