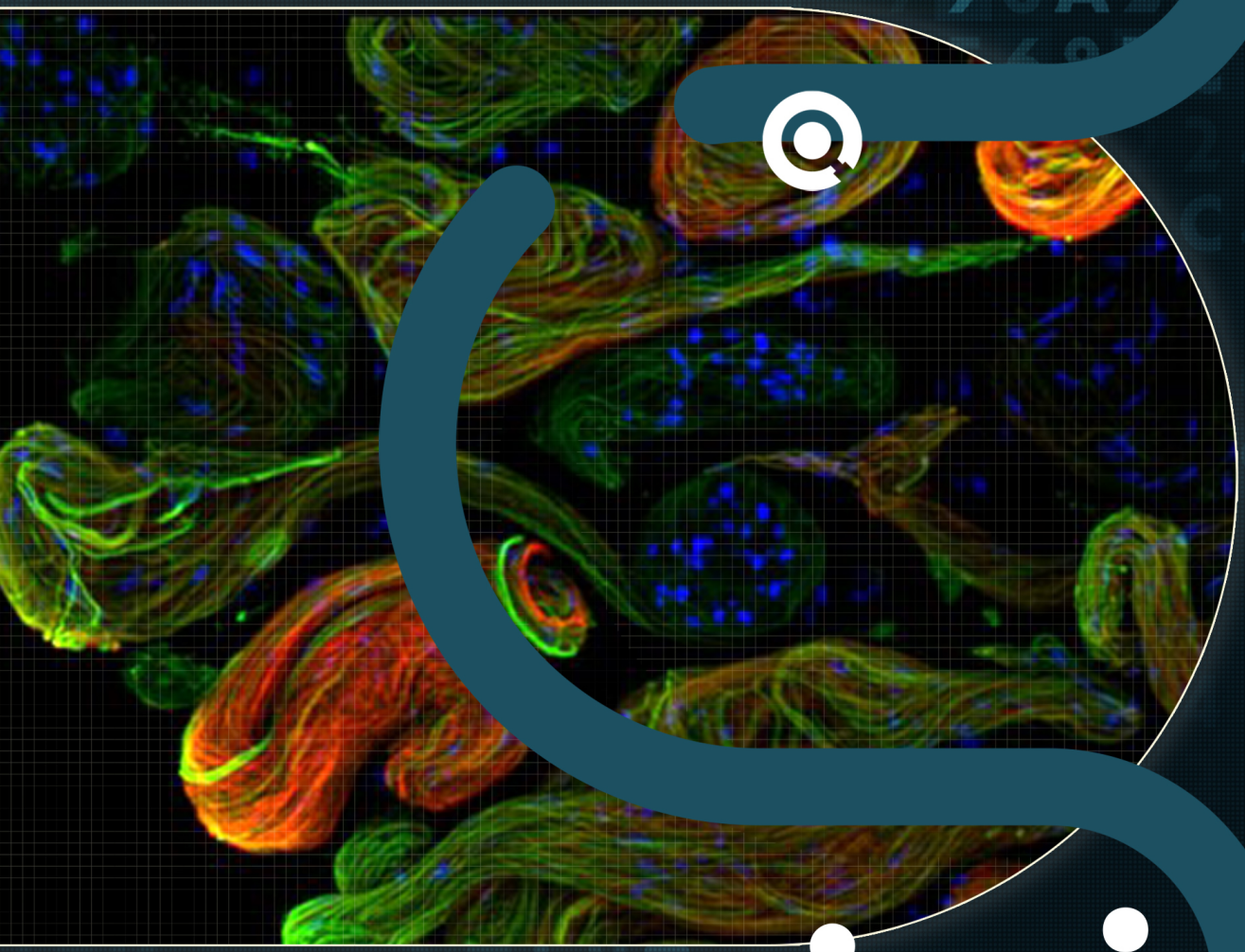
cnrs

MFM
Microscopie de
Fluorescence
Multidimensionnelle

Guide
des bonnes pratiques



Pour la gestion des données de la recherche en BioImagerie

Cas d'usage en microscopie photonique

GeDe

Alexis Lebon[1], Julio Mateos Langerak[2], Faisal Bekkouche[3], Guillaume Gay[4], Mathieu Vigneau[5], Damien Schapman[1], Thomas Guilbert[6]

1 : PRIMACEN, Université Rouen Normandie, INSERM US51, CNRS UAR2026, Rouen 76000 France
2 : Montpellier Ressources Imagerie, BioCampus, Université de Montpellier, CNRS, INSERM
3 : Laboratoire de Biologie du Développement de Villefranche-sur-Mer, UMR7009 - Institut de la mer de Villefranche, EMBRC France, CNRS, Sorbonne Université
4 : France BioImaging, BioCampus Montpellier, Université de Montpellier
5 : RESTORE lab UMR 1301-Inserm 5070-CNRS EFS UPS ENVT, Toulouse
6 : IMAG'IC, Université Paris Cité, CNRS, Inserm U1016, Institut Cochin, Paris 75014 France

Abstract :

Ce guide a été conçu pour répondre aux défis récents de la gestion des données de recherche dans le contexte de la science ouverte. Il s'appuie sur les réflexions interdisciplinaires menées au sein de réseaux et de groupes de travail de la MITI (GeDeM, RTmfm et DOREMITI) et d'instituts du CNRS, mettant en avant des pratiques FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables). En prenant pour exemple la gestion des données en microscopie photonique, ce document propose un cas d'usage concret tout en exposant les notions fondamentales de la gestion FAIR. Le guide suit les étapes du cycle de vie des données, enrichi d'une phase initiale dédiée à la planification et à la préparation des projets. Cette approche met en lumière l'importance de chaque étape, depuis l'acquisition jusqu'à la publication, afin d'assurer la pérennité, la diffusion et la réutilisation des données au-delà de leur contexte initial. Non exhaustif, ce guide s'inscrit dans les efforts nationaux pour la science ouverte, offrant un accompagnement aux ingénieurs et aux chercheurs dans la gestion des données.

V2.3_2025-01-29

Table des matières

Introduction

La gestion rigoureuse et cohérente des données de la recherche constitue aujourd'hui un enjeu de taille pour la production de nouvelles connaissances scientifiques. Guidés par le "Plan National pour la Science Ouverte" [01], les différents organismes de recherche s'emparent de ces questions primordiales pour participer à la réflexion et à la mise à disposition des outils, méthodes et infrastructures répondant aux besoins des communautés scientifiques en matière de gestion et de partage des données scientifiques.

"La science ouverte est la diffusion sans entrave des résultats, des méthodes et des produits de la recherche scientifique. Elle s'appuie sur l'opportunité que représente la mutation numérique pour développer l'accès ouvert aux publications et — autant que possible — aux données, aux codes sources et aux méthodes de la recherche." [94]

Le Groupe de Travail inter-réseaux DOREMITI (Données de la Recherche de la MITI) de la Mission pour les Initiatives Transverses Interdisciplinaires (MITI) du CNRS a édité en janvier 2023 un Guide des bonnes pratiques sur la gestion des données de la recherche [02] Ce guide, très complet, permet au Groupe de Travail GeDeM (Gestion des Données et Métadonnées) [03] du RTmfm (Réseau Technologique de microscopie de fluorescence multidimensionnelle) [04] d'établir un cas d'usage de la gestion des données d'imagerie produites par les plateformes de microscopie photonique, ou dans les laboratoires.

Nous avons choisi de reprendre la structure du guide du GT DOREMITI sur le cycle de vie des données. Ce document a pour vocation d'aider toute personne souhaitant s'informer sur la gestion des données et de fournir des clés pour tendre vers la FAIRisation des données (cf. 1.4) issues de nos différentes plateformes. Il contient des recommandations du GT GeDeM. Ce document est également une aide pour la compréhension du vocabulaire utile à la rédaction d'un Plan de Gestion de Données de type entité (anciennement structure) dédié aux plateformes de microscopie photonique. Les produits de la recherche qui sont décrits ici concernent les données de type images issues de microscope photonique, codes sources informatiques, fichiers textes et tableurs.
Ce guide n'a pas vocation à être exhaustif car toutes nos plateformes sont différentes, et à une généralité, il existe toujours des exceptions. Ce domaine évoluant très rapidement, il conviendra de s'informer régulièrement.

De nombreuses ressources anglophones existent déjà sur ce sujet. Parfois complexes, une excellente approche pour débuter est de se référer au site RDMkit [05].

RECOMMANDATIONS DU GT GEDEM

"Les cas particuliers ne le sont jamais assez pour déroger aux règles.
Mais, à la pureté, privilégie l'aspect pratique."

extrait de *The Zen of Python* [06], Tim Peters

1. Imaginer et préparer

Imaginer est la première étape du cycle de vie de nos données (figure 1). C'est une phase préparatoire qui correspond à la connaissance et à l'identification des problématiques générales, techniques et juridiques associées à la gestion des données dans un projet de recherche ou dans la pratique quotidienne de nos métiers. Une vision d'ensemble du paysage de la gestion des données de la recherche est fournie dans le Guide des bonnes pratiques sur la gestion des données de la recherche **[07]**. Il est nécessaire de se familiariser avec cet écosystème qui est au service du partage et de l'ouverture des données de recherche.

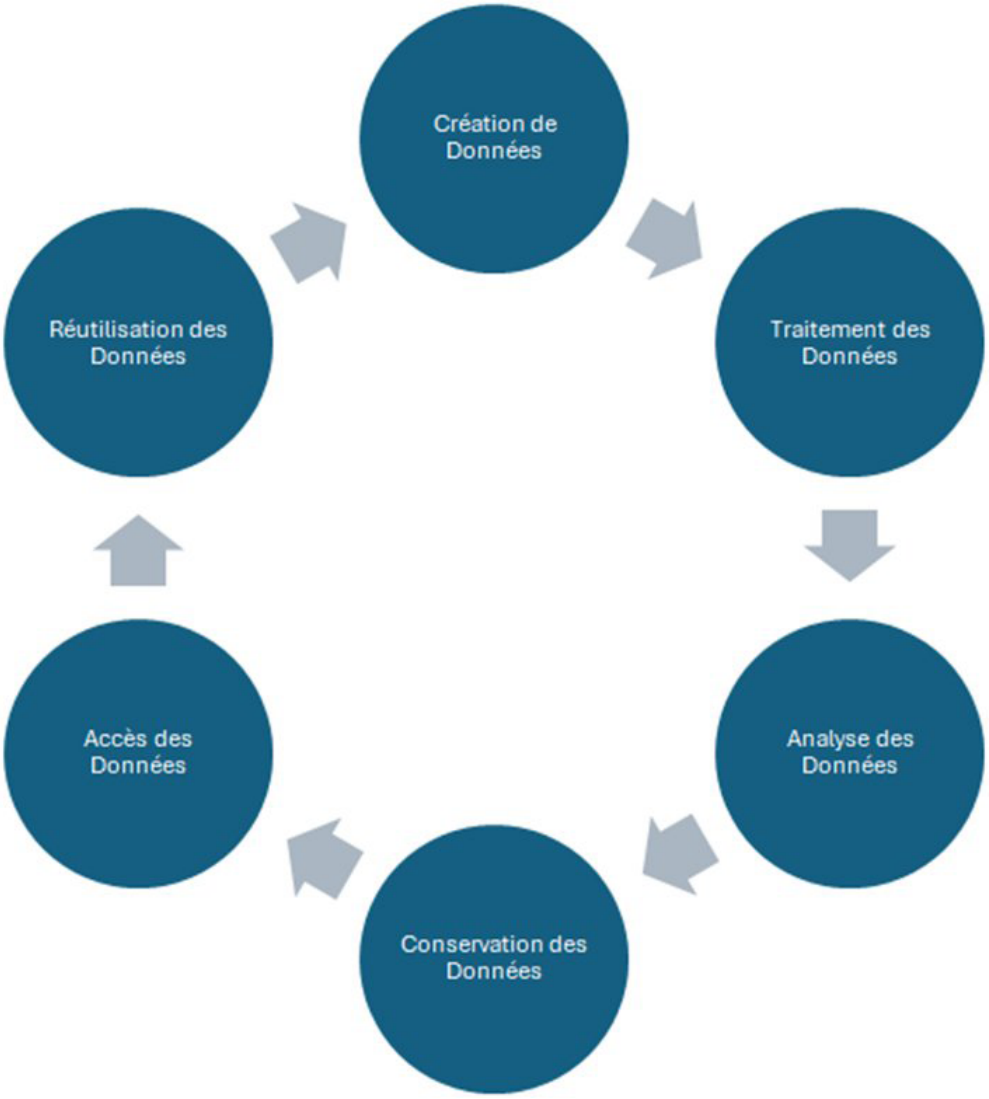


Figure 1 : Cycle de la vie des données

Nous commencerons donc par décrire ce que sont nos données, le contexte technologique national, quelles législations leurs sont associées, et comment les rendre le plus FAIR (Faciles à trouver, Accessibles, Interopérables, et Réutilisables, cf. **1.4**) possible dans le cadre du deuxième plan national pour la science ouverte [08].

Kemmer *et al.* [09] ont publié un article en juin 2023 traitant des meilleures manières de construire un écosystème de données d'images FAIR pour la communauté des microscopistes.

1.1 / Les données en plateforme de microscopie photonique

Dans le cas précis des plateformes de microscopie photonique et de la bioimagerie, les données sont majoritairement des images multidimensionnelles. Dans la rédaction d'un plan de gestion de données entité (cf. 2.2), les données issues d'une plateforme sont décrites comme étant les produits de la recherche et concernent donc les images, les métadonnées qui leur sont associées et tous documents permettant de retrouver les résultats et d'assurer la reproductibilité (les images dérivées, les vidéos de time-lapses ou de rendus 3D, les codes sources informatiques développés à façon, les fichiers d'annotations, les fichiers tableurs contenant des résultats...). La science ouverte [10], ou open science, est un mouvement dont l'objectif est de rendre universellement accessibles ces informations.

A noter que dans le cadre de ce guide, le cas des données sensibles à caractère personnel ainsi que les données de santé ne sera pas particulièrement traité. Pour plus d'information, il est possible de se reporter vers le Health Data Hub ou sur le cloud dédié de l'Inserm certifiant les Hébergeurs de Données de Santé.

1.2 / Comprendre et respecter la législation en vigueur

Gérer les données de la recherche suppose de clarifier en amont les modalités de partage et de mise à disposition des données de la recherche et le cadre juridique applicable aux projets de recherche. Ainsi, la loi Valter (2015) [11] et la Loi pour la république numérique (2016) [12] élargissent toutes deux le champ d'application de la Loi CADA [13], et ont pour objectif de favoriser la réutilisation de l'information publique. Ces deux lois sont à l'origine du principe d'ouverture ou d'open data par défaut : "Aussi ouvert que possible et pas plus fermé que nécessaire". Ainsi, il n'y a pas de droits d'auteur sur les données mais les chercheurs sont fortement incités à partager leurs données. Cependant, le partage des données ne porte que sur des données achevées (voir le code des relations entre le public et l'administration, article L. 311-2 : l'application du principe d'ouverture des données publiques est liée à la notion de « document achevé »). Pour la recherche, ce serait par exemple à la fin du projet ou à la publication d'un article.

Les données et les codes issus de la recherche sont considérés comme des documents administratifs : cela implique un droit d'accès sur demande, une obligation de diffusion gratuite et une libre réutilisation. Pour permettre leur partage ou leur diffusion, les divers documents, codes sources et logiciels produits dans nos institutions doivent donc être protégés par une licence autorisée par la loi [14].

1.3 / Hygiène numérique

Le développement du numérique et d'Internet a révolutionné nos manières de vivre et de travailler. La protection des données personnelles et professionnelles est primordiale. L'hygiène numérique est faite de petits gestes simples, trop peu souvent appliqués voire ignorés, ou non connus de la majorité des utilisateurs d'un outil informatique, qui permettent l'autodéfense numérique. Une liste de mesures permet de mettre en place une première barrière de protection et de s'informer des différents risques [95].

La démarche d'ouverture des données de recherche dans laquelle nos établissements sont engagés est une démarche positive à de nombreux égards (préservation pérenne, reproductibilité, etc.). Elle doit néanmoins, et de manière urgente, être considérée aussi du point de vue de son impact environnemental.

La gestion des données participe à cet impact avec une soixantaine de zettaoctets de données créées en 2020 et des projections à 170 zettaoctets pour 2025.

Une réflexion sur les outils, les infrastructures et les formats à utiliser s'impose, de même qu'une gestion FAIR rigoureuse avec *une sélection stricte des données utiles, nécessaires, validées et suffisamment bien qualifiées (avec des métadonnées de qualité) pour éviter de sauvegarder et de conserver des données inutilisables.*

1.4 / Connaître et comprendre les principes FAIR

Il s'agit d'un ensemble de principes directeurs [15] visant à rendre les données de la recherche: 1- **F**aciles à trouver, 2- **A**ccessibles, 3- **I**nteropérables et 4- **R**éutilisables (**FAIR**) par les êtres humains et les machines. Ces principes sont une formalisation technique de ce qui est considéré comme l'état de l'art de la bonne gestion des données et des métadonnées.
Il s'agit en particulier de mettre l'accent sur le renforcement de la capacité des machines à rechercher et utiliser automatiquement les données afin de favoriser leur réutilisation par la communauté (privé, public, grand public).

Les principes FAIR [16] ont pour objectif de guider le partage et la publication des données. Toutefois, s'il y a une volonté forte en faveur du partage et de la réutilisation des données (les principes sont adoptés par de plus en plus d'organismes de financement de communautés scientifiques et sont également préconisés dans le plan national pour la science ouverte et dans la feuille de route du CNRS), il faut bien garder à l'esprit qu'appliquer les principes FAIR n'implique pas l'ouverture systématique des données. Le principe de base "**aussi ouvert que possible, aussi fermé que nécessaire**" reste en vigueur y compris lorsque les principes FAIR sont appliqués.

1.5 / Prévoir la traçabilité des données

Dans un environnement où l'information arrive en masse, pouvoir assurer la traçabilité des données est essentiel. Les données numériques représentent un enjeu majeur pour la recherche, il est donc important d'intégrer une traçabilité de leur production au sein des structures de recherche pour disposer de données fiables et réutilisables. Le réseau Qualité en Recherche - particulièrement investi sur ce sujet - a élaboré en 2018 un guide de référence : Traçabilité des activités de recherche et gestion des connaissances [17], à destination des agents des unités de recherche. Ce guide a pour objectif de fournir des recommandations et bonnes pratiques concernant l'utilisation des cahiers de laboratoire électroniques, fiches projets, etc...

1.6 / Envisager la curation des données

La curation des données est l'activité permettant de vérifier la conformité du jeu de données. Elle vise à assurer une bonne compréhension des données publiées et à favoriser leur réutilisation, en accord avec les principes FAIR (éviter les doublons, vérifier l'intégrité des métadonnées, etc...). La curation des données est essentielle dans la pratique de gestion des données, car elle contribue à leur pérennité, leur qualité et leur ré-exploitation. Elle s'avère toutefois difficile à définir, car sa pratique se situe très souvent à la croisée de différentes disciplines. Elle s'applique tout au long du cycle de vie de la donnée et intègre des tâches de nature parfois différente comme la sélection, la vérification, la normalisation ou encore l'enrichissement nécessaires à la publication des données.

"Les activités de curation de données permettent de faciliter la découverte et la récupération de données, de maintenir la qualité des données, de leur ajouter de la valeur et d'en fournir pour de futures ré-utilisations. Ce nouveau champ inclut la représentation, l'archivage, l'authentification, la gestion, la préservation, la récupération, et l'utilisation."
Digital Humanities Data Curation

La qualité d'une donnée peut entrer en compte dans le choix de conserver ou non cette donnée. Dans le cadre de la microscopie photonique, il est particulièrement difficile d'associer une valeur Qualité à une image car elle est intrinsèquement liée à la question biologique posée, ainsi qu'à la technique de microscopie photonique utilisée. Cependant, il est possible de définir une image de qualité comme étant une image acquise dans une configuration où l'échantillon reçoit le moins de photons possibles, mais suffisamment pour répondre à la question biologique posée. Ce critère de qualité, correctement défini, constituerait une première étape dans l'accompagnement des utilisateurs de microscope pour les aider à améliorer le paramétrage de leurs acquisitions, et ainsi éviter de générer des données non exploitables. Le Working Group 10 du consortium QUAREP-LiMi [18] travaille à l'établissement d'un tel outil.

1.7 / Prévoir le stockage, le partage et l'archivage des données

Se préoccuper du stockage et de l'archivage des données est un élément clé d'une bonne gestion des projets de recherche. Dans un premier temps, il est essentiel de définir comment les données seront stockées durant la réalisation du projet. Ensuite, dans une logique de préservation et de partage, l'archivage doit être pensé dès la création des données, et ce, en amont du projet. L'objectif est de décrire, documenter, contextualiser les données pour pouvoir ensuite assurer leur diffusion et leur préservation à long terme. Il concerne tous types de données (bases de données, questionnaire d'enquête, données brutes, photos, etc.). Au-delà du stockage, il s'agit de faire en sorte qu'une donnée soit ré-exploitable (intègre, lisible, intelligible) dans 10, 20 ou 50 ans par une nouvelle communauté de chercheurs (cf. 6.2).

Un élément essentiel dans ce processus est le recours aux entrepôts de données dédiés à la recherche (cf. 7.6). Ces entrepôts permettent non seulement un stockage sécurisé, mais aussi une indexation, un partage organisé et une accessibilité durable des données. Ils jouent un rôle crucial dans l'intégration des bonnes pratiques FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable).

Les entrepôts de données ne sont pas adaptés à leur conservation pérenne. Ce rôle revient aux

systèmes d'archivage électroniques (SAE, pour plus d'éléments introductifs voir NF Z 42-013), conformes à des normes strictes qui garantissent l'intégrité, l'authenticité et la traçabilité des données sur le long terme. Ces systèmes complètent les entrepôts en prenant en charge l'archivage, indispensable pour préserver les données et les rendre accessibles aux générations futures.

Des outils existent pour aider à la sélection des données archivables, notamment le référentiel de gestion des archives de la recherche [19]. Ce référentiel est organisé par thématiques et indique pour chaque type de document sa durée de conservation, son sort final (tri, conservation, destruction) et les aspects légaux à connaître.

1.8 / S'informer et se former en gestion des données de la recherche

La formation continue des personnels est fondamentale pour suivre l'évolution des métiers et des technologies.

Au CNRS, la formation continue est pilotée par le Service Formation et Itinéraire Professionnel (SFIP) [20]. Ce service met en œuvre des actions adaptées aux orientations et à la stratégie de l'établissement au travers de deux principaux dispositifs de formation : les Actions Nationales de Formation (ANF), fortement orientées sur les technologies et l'ingénierie ; les "Écoles Thématiques", d'un contenu davantage scientifique et plutôt en relation avec les chercheurs. Le SFIP soutient également des actions régionales de formation.

Dans ces dispositifs institutionnels de formation, le RTmfm est fréquemment au cœur des propositions de thématiques, de la programmation, du montage et de l'organisation. Chaque année, de nombreuses formations sont réalisées par les réseaux, et les supports de formations présentés sont habituellement capitalisés sous une forme ou une autre (résumé, pdf, vidéo) sur les sites des réseaux.

N'hésitez pas à contacter le groupe de travail GeDeM du RTmfm (gedem-rtmfm@groupes.renater.fr) ou encore le groupe de travail inter-réseaux DOREMITI (donnees-inter-reseaux@services.cnrs.fr) pour toute information concernant les besoins en formation sur la gestion des données.

Exemples de webinaires et formations passés :
Petits Webinaires du RTmfm 2023 : Les données de la recherche et leur protection. Cas d'usage en microscopie photonique (GT GeDeM, RTmfm)
ANF 2023 "OMERO FAIRly" (GT GeDeM, RTmfm, FBI)
FAIR data IBISA 2023 : "Principes FAIR pour la gestion des données d'une plateforme IBISA"
MiFoBio 2023: 3 ateliers et une table ronde (GDR ImaBio, GT GeDeM, RTmfm, FBI)
Journée cycle de vie des données en biologie (Univ. Côte d'Azur, RTmfm, FBI)

RECOMMANDATIONS DU GT GEDEM

Du Plan de Gestion des Données à la FAIRisation des données, il est important de se former et de former les utilisatrices et utilisateurs des plateformes, et de se familiariser avec le vocabulaire de la gestion des données en BioImagerie.

2. Concevoir et planifier un projet de recherche (PGD)

Une bonne gestion des données de la recherche commence par une bonne planification du projet de recherche. Il s'agit ici de définir les tâches à accomplir pour réaliser le projet de recherche, d'élaborer un planning, de rechercher d'éventuels partenaires et financements, d'élaborer les spécifications nécessaires, de définir les données et les métadonnées qui seront utiles, de penser au futur plan de diffusion, et bien d'autres actions de préparation et de planification. Rebecca A. Senft et son équipe ont publié en 2023 un guide complet [21] pour accompagner ces différentes étapes.

2.1 / Description du projet

Pour ces travaux de conception et de planification, du point de vue des plateformes de microscopie photonique, une grande part incombe aux chercheurs / utilisateurs des machines. Cela étant, les plateformes sont souvent un appui pour la gestion de projets en proposant par exemple un Plan de Gestion de Données entité (PGD), document qui va pouvoir aider les chercheurs à rédiger leur PGD de type projet. Ces discussions entre plateformes et chercheurs sont primordiales et permettent d'optimiser les volumes d'acquisition.

Un exemple notable de définition conjointe de projet de recherche / plateforme a été mis en place sur la plateforme PRIMACEN de Rouen.

Au sein de la plateforme d'imagerie PRIMACEN, un processus complet a été mis en place via une interface en ligne [22], afin d'accompagner les chercheurs/utilisateurs dans leurs conceptions et leurs démarches d'imagerie :

- Première étape : les chercheurs/utilisateurs sont invités à remplir un document en ligne nommé "demande d'accès" afin de décrire leur projet de recherche : titre du projet, description succincte en une dizaine de lignes, coordonnées d'utilisateurs, statut au sein de leur structure de recherche, les délais de faisabilité du projet, leurs affiliations, les équipements souhaités et/ou proposés, la nature de leurs échantillons, la nature de la demande (utilisation, collaboration, prestation), ainsi que leurs besoins en analyses de données.
- Deuxième étape : les personnels référents de la plateforme répondent sous 15 jours maximum à ladite "demande d'accès", soit en engageant une discussion via un système de messagerie spécifique pour optimiser les discussions, soit en se réunissant pour identifier précisément les besoins et définir l'accompagnement nécessaire pour répondre à la demande des chercheurs/utilisateurs.
- Troisième étape : les personnels référents de la plateforme rédigent un document nommé "cahier des charges" définissant les accords d'accessibilité des chercheurs/utilisateurs aux équipements de la plateforme. Selon chaque modalité d'accès, le cahier des charges est plus ou moins complexe. Ce document fait office de contrat cadre entre la plateforme d'imagerie et les chercheurs/utilisateurs.
- Quatrième étape : le nouvel utilisateur fait remplir et signer le document par toutes les parties considérées de son laboratoire de recherche, et le transmet par voie électronique en le téléchargeant sur l'interface prévue à cet effet.
- Cinquième étape : les personnels de la plateforme accusent réception du document via la même interface.
- Sixième étape : comme mentionné dans le cahier des charges, et sauf cas contraire, le nouvel utilisateur prend contact avec le(s) responsable(s) des formations des équipements concernés pour définir une date de formation.
- Septième étape : après réalisation et validation de la formation par le personnel de la plateforme, le nouvel utilisateur se voit notifier ses droits d'accès aux équipements, lui permettant ainsi de les réserver et d'accéder aux ressources informatiques (stockage, plateforme de dépôt).

Selon certaines modalités d'accès comme la collaboration ou la prestation, la complexité du cahier des charges modifie différentes étapes décrites ci-dessus.

En fonction du type d'expérience menée, l'utilisateur peut avoir différents besoins qui doivent être définis :
- type de microscope à utiliser,
- conseils en préparation d'échantillons,
- formation pour devenir utilisateur autonome,
- comprendre le mode de gestion des données de la plateforme,
- annotation des expériences, des données (nommage, métadonnées, tags...),
- espace de stockage à prévoir (plus ou moins conséquent, lié au type de microscope),
- traitement d'images,
- collecte des données.

À ce stade, il est aussi nécessaire de prévoir le mode de collecte et de stockage des données afin d'organiser, en amont, la traçabilité qui permettra de garantir la réutilisation des données :
- il existe une base de données accessible via login / mot de passe,
- un serveur de partage existe sur le réseau interne,
- transfert sécurisé institutionnel de fichiers volumineux via Internet : https://filesender.renater.fr/

Concrètement, toutes ces étapes doivent être discutées lors du premier rendez-vous entre la plateforme et le porteur du projet.

Une fiche contenant les bonnes pratiques du microscopiste pourra être remise à l'utilisateur autonome, fiche qui sera détaillée lors de sa formation :
- comment bien nommer ses fichiers [23],
- toujours conserver une copie de ses données brutes,
- comment collecter / stocker ses données,
- combien de temps les données utilisateurs sont conservées sur l'ordinateur d'acquisition.

2.2 / Élaborer le Plan de Gestion de Données de sa plateforme

À l'échelle d'une plateforme, la rédaction d'un plan de gestion de données (PGD) entité (anciennement dit de structure) permet de formaliser les points précédents en un seul document. Les plateformes numériques DMP OPIDoR de l'INIST [24] et DSW de l'IFB [25] fournissent un service permettant de rédiger de façon collaborative un PGD. DSW de l'IFB a été choisi pour développer une trame de PGD entité orientée multi-omique avec de nombreux contributeurs représentatifs des communautés de la Biomagerie, Protéomique, Cytométrie en flux, Génomique, Métabolomique... Cette trame permet d'éditer des PGDs modulaires pour un seul type de technologie/ou plateforme ou plusieurs. La finalité d'un tel document sera d'être disponible pour tout chercheur souhaitant utiliser une plateforme technologique, et d'être ainsi intégré à leur PGD de type projet. En effet, une description complète des produits de la recherche à l'échelle d'une plateforme présente plusieurs avantages :
- mieux connaître ses processus d'acquisition pour définir le mode opératoire permettant d'obtenir un échantillon observable en microscopie photonique,
- définir les stockages nécessaires à la collecte temporaire (le temps du projet) et pérenne (archivage) des données. Cela implique, en amont, de travailler en mode projet (gestion de projet) avec une équipe informatique,

- fournir aux utilisateurs le mode de fonctionnement de la gestion des données de la plateforme, depuis leur création jusqu'à leur récupération,
- formaliser l'utilisation des codes sources le cas échéant,
- harmoniser le fonctionnement des plateformes à l'échelle d'une unité, d'une région...
- mettre en place une gestion et une conduite de projets pour faire interagir les différents acteurs intervenant dans la chaîne de collecte : ingénieurs, opérateurs, chercheurs, analystes, ...

Enfin, un mot concernant la gestion des risques associés aux données de la recherche dans nos plateformes, qui consiste à :

- garantir la disponibilité de l'outil de travail pour l'ensemble des personnels de la structure,
- garantir la confidentialité des informations,
- garantir l'intégrité des informations (curation) et donc de la recherche,
- assurer la protection des données à caractère personnel et / ou données sensibles collectées, produites ou gérées par la structure (données scientifiques et techniques, données de gestion administrative, données individuelles),
- assurer la protection juridique (risques administratifs, risques pénaux, perte d'image de marque).

Une analyse de risques telle qu'évoquée dans le Guide des bonnes pratiques pour les Administrateurs Systèmes et Réseaux [26] apparaît comme une réponse aux besoins de protection des données de nos unités de recherche. Pour plus d'informations dans ce domaine, cf. 1.8.

Le PGD n'est pas un document statique mais un document dynamique. La nature de la recherche fait qu'au démarrage d'un projet, nous avons une vision très limitée des données (type, volumétrie...) à acquérir et, encore moins, de leurs exploitations. C'est la raison pour laquelle il faut revisiter et versionner régulièrement le PGD afin de le faire évoluer avec le projet.

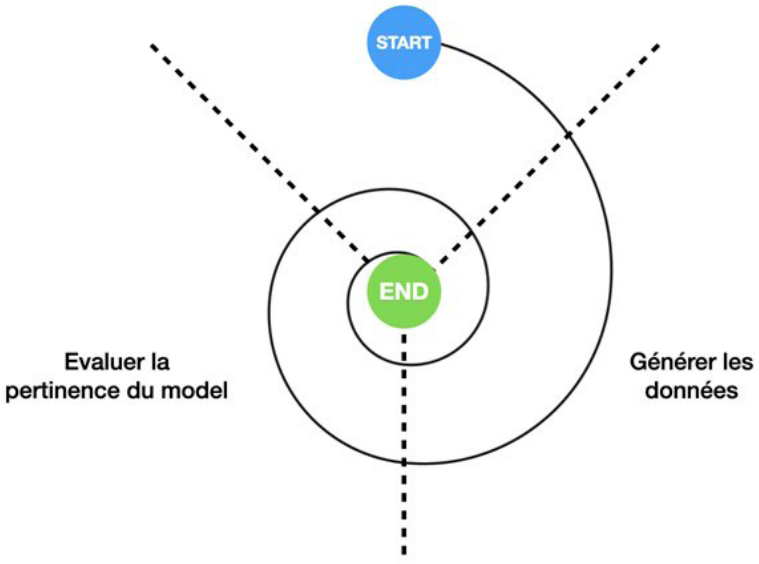


Figure 2. Le PGD est un document évolutif

Voici deux exemples de PGD rédigés via l'interface dédiée DMP_OPIDoR pour les plateformes IMAG'IC de l'Institut Cochin [96] et MICA de l'Université Côte d'Azur [97].

RECOMMANDATIONS DU GT GEDEM

Aide à la rédaction d'un Plan de Gestion de Données entité : ne pas hésiter à contacter les personnes compétentes qui aident à la rédaction de ce type de documents (DMP OPIDoR, DSW).

Générer le moins de données possible

Adapter la technique à la question biologique : compter des noyaux ou faire de la colocalisation de protéines uniques ne requiert pas la même configuration d'un système de microscopie.

Maîtriser l'échantillonnage spatio-temporel : compter des noyaux peut se faire avec un objectif 10x en binning de 4 (moins de données générées), tandis qu'une colocalisation de protéines uniques requerra une adaptation stricte de l'échantillonnage spatial avec le critère de Nyquist-Shannon. Pour la dimension temporelle, l'imagerie de signaux calciques pourra se faire en streaming (20Hz, ou 50ms de temps d'exposition) tandis qu'une expérience de tracking cellulaire ne nécessitera pas plus d'une image toutes les 10 minutes.

Se familiariser avec les lois statistiques : préparer son expérience en tenant compte des lois de puissance permettant d'adapter le nombre d'échantillons à imager pour atteindre une significativité statistique nécessaire et suffisante.

3. Collecter et stocker

Cette phase du cycle de vie de la donnée concerne les aspects d'acquisition et de collecte des données, ainsi que la constitution des jeux de données ("dataset" en anglais) avec leurs métadonnées descriptives. Dans l'objectif de rendre FAIR ces données issues de nos microscopes (Faciles à trouver, Accessibles, Interopérables, et Réutilisables), il est nécessaire de les décrire avec un maximum de métadonnées. Cet effort nécessaire prend tout son sens quand il s'agit de quantifier ou de comparer des données, de les réutiliser et de rendre l'expérience la plus reproductible possible.

Cette phase "Collecter" nécessite de :
- générer des images via le logiciel pilotant l'appareil,
- fournir les métadonnées utiles à la description des données brutes (type de microscope, paramètres d'acquisition, projet, propriétaires,...) via un fichier alimenté manuellement ou via un logiciel dédié (cf. 3.1).
- mettre en place des chaînes de collecte et de transferts des données : du microscope jusqu'aux espaces disques de stockage pérenne, en passant par le stockage tampon (et éventuellement jusqu'aux stations d'analyse où les traitements pourront être réalisés, avec la documentation adaptée),
- utiliser des protocoles, si possible normalisés ou standardisés, pour permettre d'appliquer les mêmes paramètres d'acquisition sur l'ensemble des données du projets, et rendre ainsi ces données comparables,
- disposer de cahiers de laboratoire (si possible électronique, cf 3.2.3) pour consigner le protocole d'obtention des échantillons, les relevés et métadonnées observés, les fichiers analysés, les résultats, ...

3.1 / Utiliser des normes et des standards d'interopérabilité

Lors de la collecte, l'interopérabilité est un point important afin que la donnée soit exploitable à travers différents logiciels. Cela se traduit par l'utilisation de formats standards d'images permettant d'assurer le bon fonctionnement et les échanges entre différents systèmes/logiciels informatiques. Appliquée à la bioimagerie, l'interopérabilité permet de rendre les données accessibles, réutilisables et comparables. Pour y parvenir, il faut utiliser des protocoles d'accès et des formats de données "ouverts", normés ou standardisés : au niveau des formats de fichiers d'une part, et au niveau des outils informatiques d'autre part, outils qui serviront à échanger, diffuser, lire et analyser les données. Il est important de considérer le cycle de vie des données et des métadonnées dans leur globalité (entretenir, mettre à jour, stocker, gérer la suppression des données, publier) afin de les valoriser et de ne pas surcharger les espaces de stockage.

Deux composants essentiels de la gestion de données sont à prendre en compte :
- **Les fichiers contenant la donnée** (organisation et nommage, format et critères d'interopérabilité, pérennité). Ces fichiers présentent déjà des métadonnées fournies par l'appareil (meta-dataset),
- **Les métadonnées et la documentation associée** (définitions, identifiants pérennes pour les données et syntaxes d'échange).

Pour cela, il est nécessaire d'utiliser des référentiels de métadonnées. Ils peuvent être standards ou normés; ce sont des documents importants qui se chargent de définir les informations nécessaires pour décrire les données. De ce fait, ils sont utilisés pour donner toutes les informations nécessaires à la compréhension et à l'utilisation des données, et ainsi faciliter leur exploitation et leur réutilisation

3.1.1 Stratégie d'organisation des données

Les données générées sur nos plateformes sont idéalement enregistrées directement sur les ordinateurs pilotant les microscopes. Il faut ainsi prévoir de mettre en place une arborescence locale - hiérarchisée par projets ou par équipes de recherche - permettant à l'utilisateur de sauvegarder ses données d'une session à l'autre. Il est déconseillé de sauvegarder les données d'une expérience de microscopie directement sur un disque réseau pour des raisons d'encombrement du réseau et de coupures intempestives qui pourraient corrompre les données.

Chaque utilisateur ne devrait avoir qu'un accès limité au dossier du microscope contenant ses images (ou celles de son équipe) générées à l'acquisition, afin d'éviter que les données ne soient compromises par d'autres utilisateurs. Pour cela, il convient de mettre en place sur l'ordinateur d'acquisition une authentification individuelle et de définir, pour chaque utilisateur, des droits de lecture et d'écriture uniquement sur son dossier personnel (ou le cas échéant celui de son équipe). La configuration des multi-sessions peut se faire via les comptes locaux ou par un contrôleur de domaine si l'ordinateur est raccordé à un domaine informatique (par exemple en utilisant l'Active Directory de Microsoft).

Attention à la volumétrie des données. Les postes d'acquisition présentent généralement un espace de stockage limité, sans pour autant être munis d'un système permettant de répartir et/ou dupliquer les données localement (RAID) [27]. Ainsi, une stratégie de suppression des données au bout d'un certain temps (2 mois à l'Institut Cochin par exemple) est à envisager. Pour les microscopes générant beaucoup de données (time-lapses, light-sheet, SMLM...), il convient de maîtriser l'espace disponible avant chaque acquisition et de ne jamais enregistrer les données sur le disque dur qui contient les fichiers systèmes de l'OS (disque C:/). Pour limiter l'encombrement des disques, un quota d'espace disponible peut être appliqué pour chaque utilisateur ou chaque équipe.

Pour fiabiliser localement le stockage des données, l'utilisation d'un serveur NAS (Network Attached Storage) offrira une solution de transfert et de sauvegarde efficace. Il facilitera également l'accès aux fichiers depuis plusieurs appareils. Les données ainsi migrées sur le serveur NAS, i devront ensuite être transférées vers des espaces de stockage dédiés, sécurisés, fiables et intègres (cf. 3.3 & 6). Un certain nombre de logiciels font office de plateforme d'accès et de gestion des données. Ils permettent de gérer les données et leurs métadonnées associées, de fournir des interfaces de recherche, de géolocaliser les données, et parfois de visualiser des données avec des graphiques. Cette organisation des données facilite grandement leur analyse ultérieure.

- Le standard ISA (Investigation Study Assay) [28] permet de structurer ses données en trois niveaux (le projet de recherche, l'étude, l'expérience), et d'y associer des ressources telles que l'échantillon ou les publications. La suite logicielle open source ISA tools [29] est disponible pour permettre de répondre à ce standard.
- MethodsJ2 [30] est un outil logiciel basé sur ImageJ/Fiji qui vise à améliorer la reproductibilité en microscopie, en capturant des métadonnées d'images provenant de plusieurs sources, en les consolidant et en générant automatiquement le texte des méthodes en vue d'une publication.
- En outre, certaines disciplines adoptent une organisation hiérarchique fixe de leurs jeux de données, comme BIDS [31] pour l'imagerie du cerveau (y compris la microscopie photonique).

Il est important de noter ici que les tutelles souhaitent mutualiser les moyens de stockage par l'utilisation en prestation d'un des 29 mésocentres répartis sur le territoire national [92].

Si une plateforme de microscopie photonique n'a pas de moyens autres que des disques réseaux de type NAS pour transférer et archiver les données, l'idéal est d'envisager l'utilisation de serveurs de type base de données avec un accès personnalisé permettant l'import, l'export, la consultation et la sauvegarde des données produites. À l'heure actuelle, des plateformes de gestion de données dédiées à la microscopie photonique ont déjà fait leur preuve :

- OMERO [32] : logiciel client-serveur pour la visualisation, la gestion et l'analyse d'images de microscopie. Ce dépôt centralisé et sécurisé permet d'organiser les images en projet et de les partager simplement via un navigateur internet ou une application. L'infrastructure nationale France-BioImaging (FBI.Data) dispose d'un portail [33] sur ce logiciel pour en faciliter le déploiement.
- openCID [34] : développée à l'Institut Cochin, cette solution web permet le stockage, la consultation des images et de leurs métadonnées associées, la consultation d'images pyramidales, le traitement d'image en ligne... Cette base de données permet également de prendre en compte d'autres types de données.
- TPS Data Management : Adapté à une gestion intégrée des données de type phénotypage, avec une maîtrise de la donnée depuis la production jusqu'au stockage (Vidéo explicative [35]).

3.1.2 Stratégie de collecte et de structuration des Métadonnées

Parmi les standards généralistes les plus utilisés, le Dublin Core Centre [36], issu d'un consensus international et multidisciplinaire, définit un ensemble d'items de métadonnées obligatoires pour décrire les données :

1. **Titre :** Nom de l'image/du dataset
2. **Créateur :** Nom de la personne, de l'organisation ou du service ayant effectué l'acquisition (identifiants ORCID ou idHal est un plus)Sujet : Contexte biologique
3. **Description :** Présentation du contenu de l'image, des modalités d'acquisition
4. **Éditeur :** Nom de la personne, de l'organisation ou du service responsable de la mise à disposition ou de la diffusion des données
5. **Contributeur :** Nom de la personne, de l'organisation ou du service responsable ayant participé à la génération du dataset
6. **Date :** Date de création ou de mise à disposition de la donnée
7. **Type :** Type de données (images 2D, piles d'images, time-lapses, ...)
8. **Format :** Image ou vidéo (.tif, .avi...)
10. **Identifiant :** Référence univoque à la ressource dans un contexte donné (D.O.I., URI, ISBN)
11. **Source :** Origine de l'échantillon biologique, lieu de stockage
12. **Langue :** Langue du contenu intellectuel de la ressource
13. **Relation :** Référence à une ressource apparentée éventuelle
14. **Couverture :** Domaine dans lequel le dataset peut être exploité
15. **Gestion de droits :** Informations sur les droits associés à la ressource (IPR, copyright, etc.)

D'un point de vue plus spécifique, la microscopie photonique est un domaine scientifique où une image privée de ses métadonnées est inutilisable. Un jeu minimal de métadonnées recommandées décrivant les propriétés de configuration et d'acquisition d'une image est détaillée dans la publication

REMBI [37] et particulièrement dans ce tableau [38].

Pour les articles scientifiques, le consortium QUAREP-LiMi a publié en 2023 une checklist [39] qui offre aux auteurs, aux lecteurs et aux éditeurs des recommandations clés pour le formatage et l'annotation des images, la sélection des couleurs, la disponibilité des données et les flux de travail pour l'analyse des images.

Par ailleurs, le WG7 Metadata de QUAREP-LiMi (cf. 3.2.1) travaille à une spécification plus complète (NBO-Q, cf. figure 3) [40].

Pour les systèmes commerciaux et homemade, la solution Micro-Meta App [41] permet de décrire les systèmes via une interface intéractive riche [42] et génère un fichier JSON (format léger et écrit en clair d'échange de données) correspondant à une configuration hardware pour joindre à une publication. En outre, certaines disciplines adoptent une organisation hiérarchique fixe de leurs jeux de données, comme BIDS [31] pour l'imagerie du cerveau (y compris la microscopie photonique).

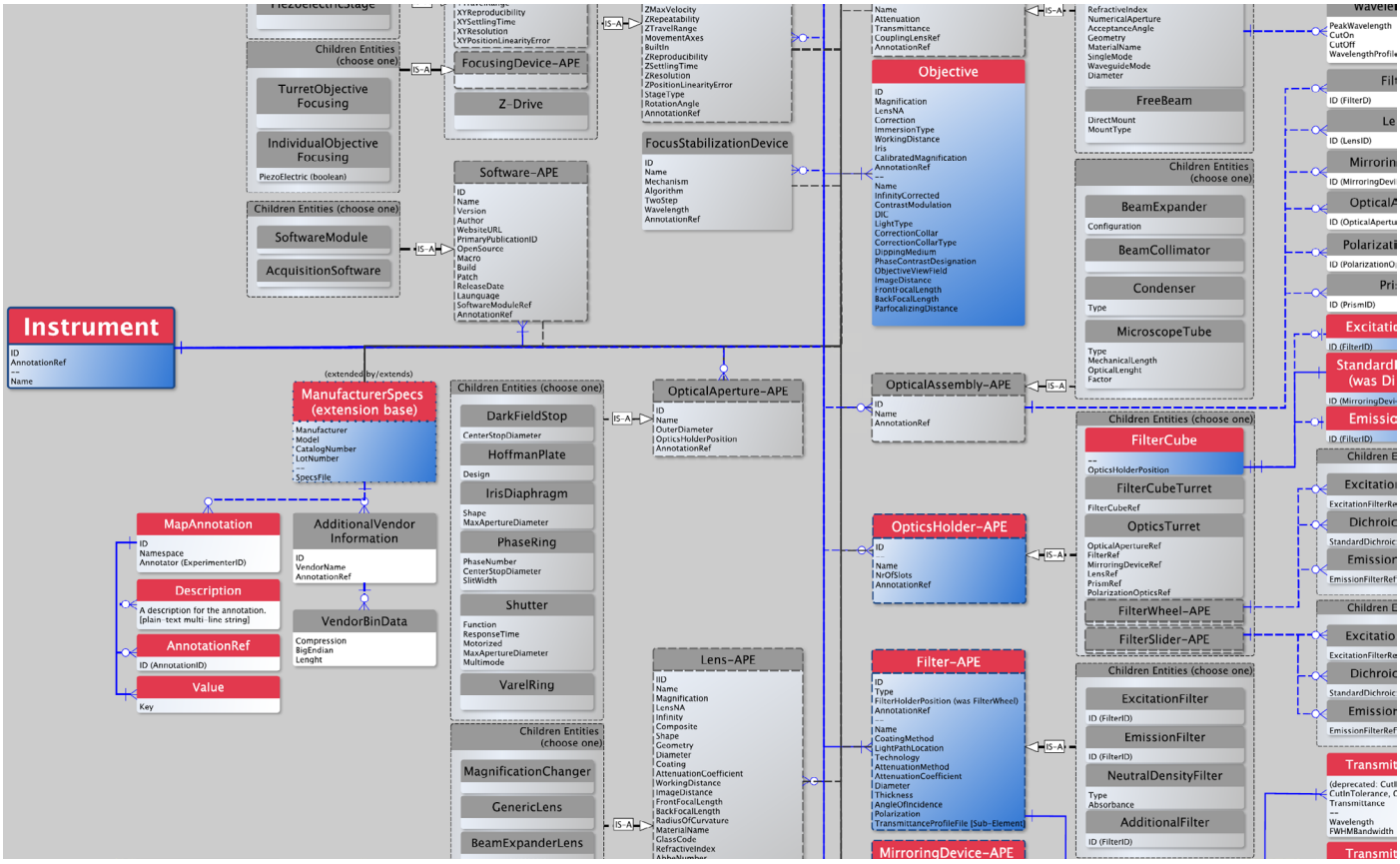


Figure 3 : zoom sur une partie de la description des métadonnées en microscopie photonique par NBO-Q .

La collecte des métadonnées peut se faire de différentes façons, plus ou moins efficaces :

- Manuellement, directement par l'opérateur, en remplissant un tableur ou un document texte, un cahier électronique, ou en alimentant une application nomade dédiée ou une application web. Ces applications web (homemade ou existantes) permettent en général d'extraire les métadonnées des formats propriétaires afin de les rendre directement lisibles par les contributeurs, ce qui facilitera le travail de l'opérateur et favorisera l'adoption de tels systèmes. Pour accompagner les utilisateurs, ces solutions de collecte doivent être simples, efficaces et conviviales. C'est une des pistes explorées par FBI.Data et l'Institut Français de BioInformatique.
- Une contrainte de saisie de métadonnées / tags à l'importation des fichiers images dans une base de données peut être envisagée pour forcer les utilisateurs à renseigner ces informations.
- Une nomenclature imposée sur le nommage des fichiers (cf. 3.1.3). Cette stratégie à l'avantage de standardiser les métadonnées qui pourront facilement être extraites. Le groupe de travail GeDeM du RTmfm travaille à l'heure actuelle sur un guide de bonnes pratiques de nommage des fichiers [23].
- Pour les utilisateurs possédant une base de données de sauvegarde et d'archivage de type OMERO, l'outil MDEmic [43] permet d'éditer les métadonnées pour permettre d'ajouter des descriptions (conditions expérimentales, échantillons, etc...). Cela permet ainsi d'exporter ces métadonnées en fichier texte ou table dans le but d'améliorer la reproductibilité lors, par exemple, de publications scientifiques.

3.1.3 Stratégie de nommage des dossiers et des fichiers

Pour retrouver facilement les données, il convient d'adopter un nommage des fichiers. De bonnes pratiques existent pour classer de manière efficace les données et gérer les versions :

- Éviter les noms trop longs
- pas de caractères spéciaux (accent, pourcentage, apostrophe...)
- Éviter les espaces et privilégier les tirets bas (underscore).
-
- Utiliser des dates au format ISO-8601 : AAAA-MM-JJ
- Intégrer l'acronyme du projet
- Intégrer le nom du créateur
- Type de données (Raw, traitées, quel traitement...)
- Conditions
- Numéro de version

Exemple : 2023-01-15_TNT_MicroAlgue_ALebon_Controle_RAW_v01.csv

Le GT GeDeM a ses propres préconisations [44].

L'INRA récapitule les préconisations principales [45] en termes de nommage. Attention toutefois au nombre de caractères (idéalement inférieur à 128 pour éviter les problèmes avec le système de fichier).

3.2 / Les systèmes d'acquisition : maîtriser l'acquisition et la collecte des données

3.2.1 / Reproductibilité de la collecte des données

Il est important que le processus de collecte des données soit clairement défini et validé. Par exemple, il conviendra de s'assurer que les systèmes d'acquisition sont bien étalonnés. Par nature, la recherche n'est pas un processus répétitif, elle est pleine d'aléas et d'incertitudes contrairement à un processus industriel. Or, pour comparer des analyses, il est nécessaire d'avoir une reproductibilité dans le processus d'acquisition. La confiance dans la qualité d'une recherche va consister à établir et vérifier que les différentes étapes d'une étude peuvent être répétées [46] en obtenant le même résultat par différents chercheurs, à des moments différents. Il est donc essentiel de s'assurer que l'ensemble des activités soit tracé et maîtrisé ; cela est nécessaire pour toute la chaîne fonctionnelle d'une analyse (des pipettes, balances, jusqu'aux équipements d'analyse) et s'applique également aux microscopes photoniques.

À l'échelle nationale, le groupe de travail Mesures de Métrologie en Microscopie (GT-3M_RTmfm [47]) développe des outils, des organigrammes, des protocoles et des bonnes pratiques de mesures et de suivi de la qualité des technologies de microscopie photonique à fluorescence [91]. À l'international, le consortium QUAREP-LiMi [48], constitué d'académiques, d'industriels et d'éditeurs, travaille sur la normalisation de ces procédures.

3.2.2 / Transferts des données

Récupérer les données relève souvent de la mise en place de chaînes de collecte composées de plusieurs étapes, plusieurs transferts de fichiers, voire plusieurs transformations de données (conversion en vue d'une analyse directement sur les stations dédiées par exemple). Pour accompagner les utilisateurs, il est possible d'écrire des programmes (en langage Python, BASH...) pour automatiser au maximum les transferts des données au travers de cette chaîne. Il peut alors être utile d'avoir des systèmes de contrôle, de surveillance ou de monitoring, qui permettent de s'assurer que les données arrivent à bon port, au bon format, et à l'endroit où elles sont attendues. L'élaboration de "dashboard" ou "tableau de contrôle" peut être envisagée pour ce type de surveillance (Apache Airflow, Framework "Dash", ...).

3.2.3 / Cahiers de laboratoire électroniques

Le processus de collecte des données de la recherche doit être décrit et enregistré en vue d'une réutilisation potentielle. Le cahier de laboratoire est un outil non obligatoire, mais fortement recommandé pour toute structure générant des données donnant lieu à des connaissances diffusables et valorisables. Il constitue un véritable outil scientifique, et ce dès le commencement d'un projet. Les cahiers de laboratoire répondent également aux obligations légales et contractuelles, en apportant la preuve de l'invention et en légitimant ses inventeurs. Nous pouvons citer elabFTW [49], logiciel open source proposé par le CNRS, ou le progiciel LabGuru [50] proposé par l'Inserm.

3.3 / Environnements de stockage - Sauvegarder les données

Dès la phase de collecte, il convient de se préoccuper des aspects de stockage et de sauvegarde qui seront plus largement abordés dans la phase 6 du cycle de vie des données. En effet, dès le début d'un projet, il est nécessaire d'estimer le volume de stockage nécessaire à la collecte de données, et de mettre en place les moyens adaptés pour la sauvegarde des données récoltées.

En pratique, pour avoir des performances optimales, les données générées sont, dans un premier temps, enregistrées localement sur le poste d'acquisition (cf. 3.1.1 & 4.2). Ces postes n'étant pas à l'abri d'une compromission, il faut voir cet espace comme un stockage tampon. Il convient ensuite de transférer les données (manuellement ou automatiquement de façon synchrone ou asynchrone) sur des serveurs distants, sécurisés avec redondance et sauvegarde. Attention toutefois à ce que le débit soit adapté à la taille des données à transférer (des travaux d'upgrade des infrastructures réseaux sont envisageables en collaborant avec les services informatiques).

La duplication des données par stockage redondant sur des supports différents de ceux de l'équipement utilisé (poste de travail fixe, mobile, serveur, ...) est un des principes de base d'une bonne conservation. Il est recommandé d'opter pour un stockage centralisé, en suivant la règle communément admise du 3-2-1 : 3 copies sur 2 supports différents, dont 1 copie stockée dans un lieu géographiquement distant. Pour garantir la mise en œuvre de cette approche, il est essentiel de collaborer en amont avec une équipe informatique afin de s'assurer que les dispositifs de stockage soient adaptés, facilement accessibles et suffisamment dimensionnés pour répondre aux besoins du projet.

Divers outils de sauvegarde des données sont couramment utilisés dans les environnements informatiques, tels que **BackupPC** [51], **Cobian Backup** [52], **Bacula** [53], et **rdiff-backup** [54]. Parmi eux, **Rsnapshot**, basé sur **rsync**, permet également de réaliser des sauvegardes incrémentielles efficaces sur les systèmes Linux/UNIX.

Un nouveau paradigme dans la sauvegarde consiste à introduire et utiliser des fonctionnalités de déduplication. Cette technologie consiste à réduire les volumes sauvegardés et les durées de sauvegarde en découpant les gros fichiers en fragments (blocs), et en ne sauvegardant qu'une seule fois les fragments identiques. Un volume maîtrisé facilitera les réplications et limitera l'impact environnemental.

Dès lors qu'il y a collecte de données personnelles (données permettant l'identification directe ou indirecte d'une personne), il est important de respecter des principes essentiels sur la durée de conservation des données, le droit à l'information et l'obligation de sécuriser les données (RGPD - Règlement Général sur la Protection des Données). Pour ce qui concerne les données de santé, la certification Hébergeur de Données de Santé (HDS) est obligatoire. Il ne faut pas hésiter à se rapprocher du correspondant du Délégué à la protection des données (DPD) de votre établissement. Nous pouvons néanmoins préciser que la collecte des données sensibles peut devenir compatible avec le RGPD s'il existe un consentement explicite entre les parties prenantes pour une ou plusieurs finalités spécifiques, ou s'il est justifié que la collecte est d'intérêt public, selon certaines conditions.

La durée de conservation des données est abordée dans la partie 6.

RECOMMANDATIONS DU GT GEDEM

Mettre en place des stratégies d'**arborescence**
(de type Projet / Étude / Expérience) et de **nommage des fichiers.**

Après avoir collecté les données et leurs métadonnées sur l'ordinateur pilote du microscope, il convient de les mettre à disposition de l'utilisateur.

Utilisateur local :

- Mésocentre national (+++)
- Compte utilisateur sur serveur BDD backupé (+++)
- Sur un NAS backupé avec logiciel de gestion des données (++)
- Sur un NAS avec arborescence maîtrisée (+)

Utilisateur extérieur :

- Mésocentre national (+++)
- Compte utilisateur sur serveur BDD backupé (+++)
- Transfert par Renater sécurisé (+)
- Disque externe (---)

La maîtrise de la collecte des **métadonnées** prend du temps. Assurez-vous d'abord de leur intégrité en vérifiant, sur chaque microscope, l'accord entre la configuration du microscope, les paramètres d'acquisition et les métadonnées résultantes.

4. Traiter

Cette phase du cycle de vie des données correspond au prétraitement des données brutes issues des acquisitions et de la collecte. Il s'agit souvent de regrouper, de choisir et de qualifier les données pertinentes parmi celles qui ont été collectées, puis de les mettre en forme dans des formats standards interopérables, et de les préparer en vue de leur analyse ultérieure.

Cette partie est donc structurée en deux sections décrivant cette préparation des données :
- Préparer les fichiers de données, en vue de leur analyse, en utilisant des formats interopérables.
- Vérifier et s'assurer de la qualité des données.

4.1 / Préparer les fichiers de données en vue de leur traitement

Les données brutes sont issues des capteurs des microscopes. Ces appareils génèrent des images accompagnées de leurs paramètres d'acquisition. Ces informations se présentent souvent sous la forme de fichiers dans des formats propriétaires peu exploitables et peu interopérables en l'état (.lif pour Leica Microsystems, .czi pour Carl Zeiss, .nd2 pour Nikon, .vsi pour Evident...). Un connecteur est nécessaire pour ouvrir de tels fichiers de manière autonome, sans passer par le logiciel constructeur. Le plus répandu est la bibliothèque Java Bio-Formats [55] permettant de lire ces formats de fichiers d'images. La communauté peut cependant faire face à des formats de fichiers un peu plus retors, comme le .mrxs (mode de compression JPEGXR pour les contrastes fluo), qui doit être converti en .zarr, avant une nouvelle conversion en .ome.tiff [56]. Ainsi les données sont conservées, mais nous avons constaté certaines modifications au niveau des métadonnées. En vue de l'analyse, il est recommandé de s'assurer que la donnée brute validée et sauvegardée sur le support de stockage soit accessible en lecture seule pour éviter qu'elle ne soit ni modifiée ni supprimée. Seul le porteur du projet pourra disposer des droits en lecture/écriture.

Dans une optique de gestion FAIR, il est important de se préoccuper du format des données afin de les rendre "ouvertes" et interopérables. La notion de format "ouvert" est importante pour que les données puissent être partagées, interopérables et préservées sur le long terme. Si l'objectif est le traitement massif des données, il faut choisir des formats capables de supporter des entrées / sorties intensives sur des infrastructures de calcul. À l'heure actuelle, le format TIFF [57] est privilégié pour les données brutes. Mais ce format, datant des années '80, ne répond plus aux besoins actuels.

En effet, les jeux de données devenant de plus en plus volumineux, ils ne rentrent plus en mémoire vive (RAM), ce qui rend leur ouverture compliquée. L'utilisation de nouveaux formats devient alors indispensable.

Le format HDF5 (Hierarchical Data Format, version 5) [58] est un format de fichier de type conteneur, c'est-à-dire assimilable à une arborescence de dossiers / fichiers contenus dans un même fichier. C'est un format très utilisé pour traiter ou simuler des données grâce au calcul intensif, car il offre des possibilités de compression et d'écriture/lecture très efficaces.

Le nouveau format de fichier OME-zarr (spécification NGFF Next Generation File Formats) [59], utilisé en imagerie photonique, permet de combiner les avantages du format HDF5 :
- un accès aux données par morceaux (chunks) de petite taille, très efficace pour l'accès distant,
- une structure pyramidale (la même image avec différentes définitions),
- la lecture et la visualisation des images à travers le réseau,
- la manipulation et l'analyse des données sont alors simplifiées.

L'utilisation de solutions de compression sans perte pour sauver de l'espace de stockage est également possible. Le format de compression Jetraw en est un excellent exemple.

4.2 / Mettre en place un contrôle qualité des données

La qualité des données est une notion qui se retrouve sur toutes les étapes du cycle de vie de la donnée. Elle recouvre cependant des concepts différents (qualité des données, des métadonnées, de l'archivage, ...) et peut toujours être vue sous deux angles :

- Qu'est-ce qu'une donnée de qualité ?
- Quelle organisation faut-il mettre en place pour arriver à obtenir des données de qualité et reproductibles ?

En microscopie photonique, les images de qualité sont celles qui peuvent répondre de manière fiable aux questions scientifiques, en tenant compte des métadonnées, du traitement et de l'analyse des images ainsi que du contrôle de l'état de santé du microscope dans le temps. Par conséquent, en fonction de la technique de microscopie et de la question scientifique, la qualité de l'image ne peut pas être déterminée de manière unique. Ce point est également abordé dans le paragraphe 1.7, avec le travail du WG10 de QUAREP-LiMi. Ce WG travaille à établir un score systématique de la qualité des données, score qui peut être calculé automatiquement à l'import des images dans une base de données. Par exemple, un calcul simple et automatisé du rapport signal à bruit selon la définition de SVI **[60]** permet d'enrichir les métadonnées. De la même façon, les données images saturées peuvent être rapidement repérées dans le but de sélectionner les contenus les plus pertinents (curation, cf 1.7).

RECOMMANDATIONS DU GT GEDEM

Avant même de commencer une acquisition en imagerie, il est souhaitable d'avoir en tête une idée de la figure qui illustrera les potentiels résultats de l'expérience. Savoir comment seront analysées les données avant de les générer permet de choisir précisément les configurations système et logicielle (format, résolution, échantillonnage spatial...).
Ceci implique une bonne connaissance des logiciels d'acquisition et d'analyse ou alors l'implication d'un ingénieur de plateforme dès la conception du projet.

5. Analyser

Derrière le terme “analyser”, on entend l’extraction de l’information contenue dans l’image, via l’utilisation de logiciels de visualisation, de logiciels d’analyse ou de programmes informatiques écrits sur-mesure. L’analyse des données permet ainsi de produire des modèles statistiques pertinents pour la problématique scientifique.

Pour analyser des datasets de grande taille ou en grand nombre avec des programmes développés à façon, il est préférable d’utiliser des plateformes adaptées à la puissance de calcul nécessaire :
- Calculateur sur un des 29 Mésocentres nationaux publiques (HPC - High Performance Computing, Cloud, HTC - High Troughput Computing) ;
- Station d’analyse locale performante (processeurs multicœurs avec fréquence élevée, disques rapides SSD, carte(s) graphique(s) de type “gaming”, mémoire vive conséquente) ;
- Kit de développements GPU (boitier permettant de regrouper d’importantes quantités de mémoires et plusieurs cartes graphiques dont la vitesse de calcul des processeurs est très performante) ;
- Calculateurs déportés privés.

5.1 / Conception d’outils d’analyse

Malgré la richesse des logiciels de traitement et d’analyse, l’analyste peut se retrouver sans solution pour exploiter ses images, du fait de données complexes et/ou trop volumineuses. Il est alors possible de se tourner vers la programmation informatique. Plusieurs langages de programmation (Python, BASH, Matlab, Macro ImageJ...) permettent d’écrire un logiciel sur mesure, dédié à un type d’analyse complexe. Il va s’agir d’écrire un algorithme qui va découper tout le processus d’analyse et faire appel aux outils informatiques (variables, boucles répétitives, tests conditionnels, fonctions...) en les combinant avec les différentes fonctionnalités d’analyse vu en 5.1 (segmentation, machine learning...) ou en combinant différents logiciels existants.

Attention à la qualité logicielle ! Il est assez aisé d’écrire soi-même un programme simple de quelques lignes de code, sans l’aide d’un informaticien. Mais lorsqu’il s’agit de concevoir des programmes plus complexes, il est nécessaire d’avoir des pratiques de codage collaboratives et industrialisées. Ces bonnes pratiques doivent assurer la qualité du produit développé, son évolution et/ou sa maintenance. Il faut penser à bien commenter le code, le documenter, réaliser des tests unitaires, le déposer dans des forges logicielles pour suivre les versions et la diffusion. Pensez à l’utilisation de plateformes / forges comme GitHub [65] pour la gestion des codes sources.

5.2 / Utilisation de logiciels en local

Il existe un grand nombre de logiciels de traitement et d’analyse d’images. Il ne faut pas confondre les deux disciplines. Un traitement générera une nouvelle image, souvent avec des valeurs de pixels modifiées qui impacteront entre autres les données quantitatives. Ces logiciels peuvent être en open source (ImageJ / Fiji [61], ICY [62] ou Napari [63] par exemple) ou propriétaires (Aivia Leica Microsystems) [64], Arivis (Carl Zeiss) [65], Amira (Thermo Fisher) [66] et Imaris (Oxford Instruments) par exemple [67]). L’analyse consiste entre autres à extraire des valeurs quantitatives, ou à détecter/segmenter des objets ou des ROI (Region Of Interest).

Pour assurer la reproductibilité des analyses, il est important de décrire correctement le workflow de traitement : quels logiciels? quelles versions? quels paramètres sont utilisés en entrée et en sortie ?

- En entrée : il est nécessaire de s'assurer que les métadonnées soient correctement lues par le logiciel et également bien adaptées à l'analyse en tant que telle (échelles, dynamiques, etc.).
- En sortie : les fichiers résultats (.csv, données tabulaires), des données images ou .zip type «label» (appelées communément données secondaires).
    - Maintenir la traçabilité des données secondaires avec les données brutes,
    - S'assurer de l'intégrité des métadonnées (unités des valeurs dans les tableaux, clarté des noms de colonnes / variables etc.)
    - S'assurer de la propreté de leur mise en forme, typiquement pour les données tabulaires (notion de *tidy data* [93]).

Il est également possible d'utiliser des logiciels d'analyse installés sur des serveurs afin de centraliser les licences et les ressources.

5.3 / Mettre à disposition ses outils d'analyses

Le choix de la licence, le lieu de stockage, l'utilisation d'une forge logicielle (GitLab, GitHub...) permettent de mettre à disposition de la communauté scientifique ses outils d'analyse. L'archive Software Heritage [68] permet de conserver et préserver le code source de tels logiciels. A noter qu'il est désormais possible d'associer un dépôt Software Heritage à un article scientifique déposé sur HAL, et donc de faciliter sa réutilisation.

RECOMMANDATIONS DU GT GEDEM

Être critique sur les réglages par défaut et les résultats proposés par les logiciels d'analyse. La mise en forme des résultats ainsi que l'intégrité des métadonnées des données secondaires générées sont primordiales.

Pour limiter la subjectivité liée à l'analyse et permettre une comparaison des résultats, les données doivent être analysées de la même façon et par le même analyste sur l'entièreté du projet. L'automatisation du process est un plus à ne pas négliger pour éviter un biais de sélection de type *cherry picking*.

6. Préserver et archiver

Préserver, sécuriser l'information, sauvegarder et archiver les données sont des phases essentielles de la gestion rigoureuse des données. Il n'est cependant pas toujours aisé de faire la distinction entre ces notions et celles relatives à l'utilisation d'une sémantique et d'une procédure adaptées. De plus, il apparaît souvent compliqué de préserver une donnée pour un usage futur dont on ignore le plus souvent les tenants et les aboutissants.

6.1 / Principaux statuts de la donnée

Donnée Stockée

Il s'agit de la première étape, qui consiste à déposer les données sur un support numérique, généralement sur le poste d'acquisition, afin de les rendre accessibles. L'expérimentateur peut aussi utiliser son ordinateur de bureau ou personnel, ou encore un disque partagé, pour disposer d'une copie locale des données s'il ne peut pas les conserver sur la machine principale et souhaite pouvoir y accéder localement ou hors connexion (cf. 3). Cependant, cette pratique est fortement déconseillée. En photonique, les données sont généralement transférées vers les stations d'analyse. Avoir les données en local permet d'éviter les délais d'accès, qui varient selon la qualité du réseau, et d'améliorer ainsi les performances d'analyse.

Certaines structures sont même équipées de bases de données dédiées, comme OMERO ou CID, qui représentent une autre stratégie de gestion des données. La tendance actuelle est de traiter les images avant de les transférer dans ces bases de données. Cependant, certaines structures ont choisi de transférer systématiquement les données d'acquisition vers la base de données, puis de procéder au traitement et au nettoyage des images une fois qu'elles y sont stockées.

Si le système d'information le permet, il est également possible de récupérer les images via le réseau en utilisant un transfert sécurisé (comme SFTP ou Samba), sur un serveur sécurisé. Il est recommandé d'opter pour des serveurs sécurisés et situés dans un environnement également sécurisé. Cela inclut un système d'information avec un service d'annuaire (LDAP ou Active Directory) qui gère les comptes utilisateurs, les mots de passe et les accès sécurisés, tout en répondant aux exigences de protection des données, telles que celles définies par le Règlement Général sur la Protection des Données (RGPD).

Il est conseillé de configurer des sessions Utilisateurs sur le poste d'acquisition et d'analyse, idéalement gérées via un service d'annuaire (par exemple, Active Directory). Cela permet à chaque opérateur de stocker ses données localement de manière sécurisée, tout en garantissant un accès privé et protégé. En l'absence d'un tel annuaire, la création de comptes individuels pour chaque utilisateur devient plus complexe et moins efficace. De plus, sans accès personnalisé, il devient difficile de garantir la confidentialité et l'intégrité des données, ce qui expose à des risques de vol, d'altération ou de suppression. Enfin, pour certains logiciels, l'accès dédié permet de conserver les réglages spécifiques à chaque utilisateur, en lui offrant une session de travail individualisée.

Donnée Sauvegardée

La sauvegarde consiste à dupliquer les données sur un support externe, généralement un serveur distant, afin de pouvoir les récupérer en cas de perte ou de dégradation. Elle est principalement utilisée pour les données dites «vivantes» ou «chaudes».

Il est fortement recommandé d'utiliser une sauvegarde incrémentielle, qui ne sauvegarde que les données modifiées depuis la dernière opération. Cela permet d'optimiser l'espace de stockage et d'améliorer l'efficacité du processus. En ce qui concerne la rétention des données, il est conseillé de conserver les sauvegardes pendant plusieurs jours, voire plusieurs semaines, en fonction de la capacité de stockage disponible. Conserver différentes versions des données sur une période donnée permet de revenir à une version antérieure si nécessaire.

Bien que les périphériques USB soient souvent utilisés pour cette tâche, ils présentent plusieurs inconvénients, notamment des risques d'infection virale, de vol, de chutes ou de pertes. Il est donc préférable d'éviter leur utilisation.

Donnée Archivée

L'archivage consiste à conserver des données dans un lieu spécifiquement dédié pour une très longue période. Contrairement aux données «vivantes» ou «chaudes», les données archivées ne sont pas destinées à être récupérées régulièrement. L'objectif de l'archivage est de garantir la conservation à long terme de ces données tout en facilitant leur réutilisation. Toutefois, une fois archivées, ces données ne sont plus modifiables. En revanche, leur format peut être adapté au fil du temps pour éviter l'obsolescence logicielle et garantir leur accessibilité à long terme.
Bien que les périphériques USB soient souvent utilisés pour cette tâche, ils présentent plusieurs inconvénients, notamment des risques d'infection virale, de vol, de chutes ou de pertes. Il est donc préférable d'éviter leur utilisation.

Les données archivées sont souvent qualifiées de «mortes», «froides» ou même «glaciales», car elles ne sont plus activement utilisées. Cette phase démarre typiquement à l'issue d'un projet, quand on sait que les données n'évolueront plus. Une première période dite d'archivage intermédiaire consiste à conserver les données dans leur intégralité pendant leur durée d'utilité administrative (de 5 à 25 ans), période à l'issue de laquelle les archivistes éliminent les données n'ayant ni valeur légale ni intérêt stratégique, historique ou scientifique, après accord des Archives de France pour les données publiques. Cette décision est parfois complexe, car elle soulève la question suivante : doit-on conserver ou supprimer ces données ? Il n'existe pas de réponse simple à cette question, car elle dépend de nombreux facteurs, notamment de la valeur potentielle des données pour la communauté scientifique à long terme.

Il est important de noter que toutes les données sauvegardées ne sont pas nécessairement destinées à être archivées. L'archivage représente un coût élevé, et il doit y avoir un réel intérêt scientifique, historique ou patrimonial à conserver ces données. Avec l'augmentation exponentielle du volume des données scientifiques, cette gestion devient de plus en plus complexe et coûteuse.

Une fois qu'une donnée est archivée, il est essentiel de supprimer les versions locales qui peuvent être dupliquées de manière excessive et s'accumuler dans des sous-dossiers pendant des années, sur les PC d'acquisition par exemple. Des processus automatisés, tels que des «robots», peuvent être utilisés pour supprimer ces données qui ont dépassé leur durée de vie utile. Ces robots peuvent être configurés pour supprimer automatiquement les fichiers après une période définie, une fois que la donnée devient «froide»— 2 mois à l'Institut Cochin par exemple. Cependant, avant de procéder à la suppression, il est crucial de s'assurer que la sauvegarde a bien été effectuée.

En France, le principal opérateur habilité à archiver les données et documents numériques produits par la communauté de l'Enseignement Supérieur et de la Recherche (ESR) est le CINES (Centre Informatique National de l'Enseignement Supérieur).

Certaines institutions disposent également de systèmes de stockage sur bandes magnétiques, un autre moyen de conserver des volumes importants de données à un coût relativement faible.

Donnée Préservée/Pérennisée

L'objectif est de protéger les données contre tout type de dégradation (cf. 3.3 notion de sauvegarde) :

- Dommages matériels : Les supports de stockage (disques durs, bandes magnétiques) se détériorent avec le temps.
- Dommages logiques : Les corruptions de fichiers ou erreurs humaines peuvent altérer les données.
- Destruction accidentelle : Il est important de disposer de copies redondantes des données pour pallier les pertes dues à des suppressions accidentelles ou des sinistres.

Idéalement la conservation se fait dans un lieu géographique différent. Cela protège contre les catastrophes locales (incendies, inondations, cyberattaques).

- Sites secondaires : Répartir les copies entre différents bâtiments.
- Cloud sécurisé : Utilisation de fournisseurs de services cloud conformes aux réglementations
- locales (par exemple, RGPD en Europe).
- Centres de données distants : Les données peuvent être stockées dans un centre ou une infrastructure située dans une autre région ou un autre pays.

En informatique, la règle «3-2-1» est une pratique bien connue pour la gestion des sauvegardes. 3 copies des données (Une copie principale et deux sauvegardes redondantes), 2 supports différents (ex. disque dur, cloud, bandes) et 1 copie hors site : Conserver une sauvegarde dans un emplacement distinct (cloud ou autre bâtiment) pour se protéger des sinistres locaux.

6.2 / Maintenir les données dans le temps

Les données déposées (datasets, tableaux de résultats, fichiers analysées, métadonnées) sur des plateformes de gestion de données en microscopie photonique sont généralement sauvegardées et archivées par les départements informatiques. Ils assurent également la sauvegarde de la database et des données qu'elle contient. Le projet FBI.data [69] de l'infrastructure France-BioImaging recommande de mettre en place une couche d'abstraction entre les données et le logiciel des plateformes afin de pouvoir changer de logiciel de gestion, sans perdre l'intelligibilité des données. Pour aller dans ce sens, il peut être conseillé aux départements informatiques d'utiliser le système iRODS (Integrated Rule-Oriented Data System) [70]. iRODS est un outil permettant un accès transparent aux données réparties sur différents sites et différents supports hétérogènes (systèmes de fichiers sur disque, bases de données, systèmes de bandes, etc.).

Toutefois, une alternative plus simple et flexible pourrait être OneData, qui est en train de se déployer dans le monde de la recherche et tend à se généraliser (Cette solution est utilisée par de nombreux projets européens dans le cadre de l'EOSC). Il s'agit d'un système de stockage distribué pour la

gestion de grands volumes de données. Contrairement à iRODS, qui peut s'avérer complexe à mettre en place et à gérer en raison de sa configuration avancée, OneData se distingue par son interface intuitive et sa capacité à gérer efficacement de grands volumes de données, sans nécessiter d'expertise technique approfondie. En outre, OneData est conçu pour s'intégrer facilement aux systèmes existants tout en garantissant une gestion unifiée des données, ce qui en fait une option plus accessible et pratique pour les équipes informatiques des plateformes de microscopie photonique.

L'autre point est de faire en sorte que les données restent au maximum lisibles après plusieurs années. Ainsi, il est conseillé d'exporter les paramètres depuis votre logiciel d'analyse vers un format de fichier lisible standard, ceci afin de garantir la reproductibilité de votre protocole. Pour cela, il faut privilégier les formats ouverts suivants :

- CSV (plutôt que des documents Excel/Calc) ;
- TXT (plutôt que des fichiers Word/Writer) ;
- XML/JSON (pour les données structurées).

Il est cependant possible de spécifier un format de fichier dans le répertoire JHOVE s'il n'existe pas d'alternative. Déclarer un format permet de le signaler comme devant être maintenu dans le temps.

RECOMMANDATIONS DU GT GEDEM

Se rapprocher des équipes informatiques des Départements des Systèmes d'Informations pour avoir leur expertise sur le stockage et ne pas chercher à réinventer la roue.
Se rapprocher du service archives de son institution ou de sa / ses tutelles hébergeantes.

7. Publier et diffuser

Cette dernière étape du cycle de vie des données représente la finalité de toute une politique de gestion de données FAIR, puisqu'elle vise, dans un contexte de Science ouverte, à publier et à diffuser les données de manière à ce qu'elles soient correctement faciles à trouver, accessibles et surtout ... "réutilisables", selon des formats ouverts et des processus interopérables. L'accompagnement des réseaux métiers et technologiques s'exerce sur diverses actions comme par exemple :

- la documentation des données via des métadonnées descriptives provenant de vocabulaires contrôlés (thesaurus disciplinaires) et de leurs formats d'exploitation pour en assurer la réutilisabilité.
- l'établissement de catalogues de données (idéalement moissonnables) nécessaires pour trouver et identifier les données;
- le processus de dépôt des données dans des entrepôts respectant les critères "TRUST" [71] ou des plateformes techniques, pour en permettre l'accès centralisé;
- l'aide au choix d'entrepôts de données effectué par DORANum ou bien encore par le CoSO ;
- l'utilisation d'outils logiciels et de protocoles interopérables permettant d'échanger ouvertement les données;
- la description et l'identification des données avec des "datapapers", et des identifiants pérennes (D.O.I.);
- la représentation des données sous forme de graphes;
- le monitoring des flux de données au moyen de tableaux de bords;
- etc.

Ainsi, les réseaux travaillent sur l'harmonisation et la génération [72] des informations (métadonnées, données, modes opératoires, échantillons, publications, visualisation et interfaces graphiques) nécessaires à la mise en œuvre des supports de diffusion et de valorisation pertinents en rapport avec l'objectif du projet initial.

Cette étape de publication et de diffusion est en outre accompagnée désormais d'une action nécessaire d'identification des données via des identifiants pérennes (de type DOI [73] par exemple) lors du dépôt dans des entrepôts de données.

7.1 / Finaliser le Plan de Gestion de Données projet

La fin d'un projet est marquée par la finalisation de la rédaction du plan de gestion de données qui a été initié en début de projet par l'équipe porteuse. Il est nécessaire pour eux de s'assurer que les premières informations saisies sont encore valides, et d'effectuer une mise à jour en ajoutant les dernières informations disponibles.d'alternative. Déclarer un format permet de le signaler comme devant être maintenu dans le temps.

7.2 / Publier les Métadonnées

Les catalogues de métadonnées représentent un moyen cohérent et rigoureux pour décrire et publier des jeux de données. Ils permettent de faciliter la recherche et l'identification des données (F de FAIR). Pour être interopérables, ces catalogues s'appuient en général sur des normes pour représenter les métadonnées.

7.3 / Diffuser avec des protocoles interopérables

Outre les formats de fichiers qui doivent répondre à des standards ouverts pour être partagés et réutilisables, il est également nécessaire de se préoccuper de diffuser les données par des protocoles d'échanges standards interopérables entre machines.

7.4 / Utilisation de thésaurus / ontologie

Les thésaurus et les lexiques sont utilisés pour représenter le contenu de documents, rendant ainsi plus fiable la recherche bibliographique. Une ontologie cherche à décrire de façon formelle un domaine de connaissance, en identifiant les types d'objets de ce domaine, leurs propriétés et leurs relations. Dans les deux cas, le but est d'utiliser un vocabulaire commun pour simplifier les échanges et faciliter l'interopérabilité.

De la préparation de l'échantillon jusqu'au traitement de l'image, il est primordial d'utiliser un vocabulaire commun dans le but de pouvoir moissonner autant que possible des données déjà existantes, mais aussi de rendre disponible ses propres données.
Le laboratoire de Catarina Strambio de Castilla a réalisé un glossaire de mots standardisés [74] très complet avec des spécifications des métadonnées [75].

7.5 / Utilisation d'identifiants pérennes

Afin d'être cités et réutilisés de manière univoque, les données et documents numériques se doivent de disposer d'un identifiant pérenne et unique.
Dans le domaine des données, les DOI (Digital Object Identification) sont des identifiants pérennes favorisant le référencement et la citation des jeux de données. Ils permettent de citer un jeu de données homogène de manière univoque et durable dans le temps, et de le lier aux publications ou à tout autre produit de recherche. Ils garantissent un lien stable vers la ressource en ligne et font correspondre en permanence l'identité de la ressource à sa localisation sur le web. Ces identifiants pérennes sont fournis automatiquement lorsqu'on dépose sur un entrepôt de confiance.

Dans les autres cas,l es DOI sont obtenus auprès du consortium international "DataCite" [76]. l'INIST du CNRS [77] est membre fondateur de DataCite, et agence d'attribution des identifiants DOI en France pour l'Enseignement Supérieur et Recherche (ESR). L'allocation de DOI sur des données implique des devoirs de la part du déposant, et en particulier de maintenir un lien permanent vers les données identifiées pendant une certaine durée, au travers d'une page de description (appelée aussi "landing page"). Cette page permet de fournir les métadonnées principales pour décrire les données et y accéder. Pour créer une "landing page" - page d'accueil pour décrire un jeu de données -, il faut s'assurer que certaines métadonnées obligatoires sont correctement mentionnées et renseignées pour permettre une recherche. Le site Datacite rappelle quelles sont les métadonnées obligatoires [78]. La page de Doranum [79] donne plus d'informations sur les identifiants pérennes.

Attention, la pérennité demandée est purement une question de service et n'est en aucun cas inhérente à un objet, ni conférée par une syntaxe de nommage particulier. Maintenir la pérennité du lien vers la localisation de la ressource est de la responsabilité du déposant ou du créateur de l'identifiant.

7.6 / Les entrepôts de données

Il est essentiel de différencier le stockage local des données (par exemple, sur l'ordinateur d'acquisition ou sur un réseau local) de leur archivage à long terme. Un entrepôt de données est une plateforme de stockage numérique centralisée dédiée au partage et à l'ouverture des données de la recherche. Il est à noter que la durée de stockage des données n'est jamais très simple à trouver. Est-ce 10 ans ? Il est nécessaire de bien se renseigner.

Il existe plusieurs catégories d'entrepôts de données :
- Généralistes comme Zenodo ou Dryad, adaptés à diverses disciplines scientifiques.
- Thématiques, spécialisés dans un domaine précis, tels que BioImage Archive pour la bioimagerie.
- Institutionnels, gérés par des organisations ou des institutions, comme le Dataverse Cirad.

Dans le domaine de la bioimagerie, 3 entrepôts spécialisés se distinguent :
- IDR (Image Data Resource) : Utilise la technologie OMERO, offrant la possibilité de visualiser des images multidimensionnelles via un dataviewer. A noter que IDR sélectionne les données, et ne prends que celles qui ont une haute valeur de réusabilité et peuvent être considérées comme des données «références».
- BioImage Archive : Conçu pour héberger des images biologiques générales. BioImageArchive accepte toutes les images, sous condition qu'elles soient liées à un travail publié ou en cours de publication, à l'exception des données dites sensibles (données de santé au sens de la CNIL ou données avec personnes identifiables).
- EMPIAR (Electron Microscopy Public Image Archive) : Spécialisé dans les images de microscopie électronique.

Les entrepôts de données permettent la publication de jeux de données (datasets) qui incluent généralement des images brutes et dérivées, ainsi que des fichiers associés au projet de recherche, comme des tableurs ou des documents texte. Sur les entrepôts génériques, ces données sont souvent accessibles sous forme d'archives compressées (.zip). Cependant, des entrepôts spécialisés, tels que BioImage Archive ou IDR (Image Data Resource), permettent également d'accéder aux images directement dans leur format d'importation natif. Ces plateformes offrent par ailleurs la possibilité de visualiser et de télécharger les données, enrichissant ainsi leur (re)utilisation et leur valorisation dans le cadre de recherches scientifiques.

Un serveur de données local peut également être utilisé comme un entrepôt de données, avec des solutions telles que OMERO ou CID. Dans ce contexte, il est possible de générer une URL publique pointant vers le dataset concerné, identifié par un DOI unique, permettant ainsi le partage des données. Cependant, ce type d'infrastructure ne constitue pas un entrepôt de confiance. La question de la pérennité des données, de leurs métadonnées et de leur accès, est essentielle dans ce cas et il ne faudra pas chercher à réinventer la roue (cf 7). En effet, la plateforme qui met les données à disposition est entièrement responsable de garantir leur accès à long terme. Cela implique aussi de

s'assurer de la continuité du service, de gérer les évolutions technologiques (cf. figure 4).

Pour répondre aux besoins des chercheurs ne disposant pas de solution de dépôt de confiance pour leurs données (ni entrepôt thématique, ni entrepôt institutionnel), un entrepôt multidisciplinaire nommé «Recherche Data Gouv» [80] a été développé pour les données de la recherche et est accessible depuis 2022, à l'instar de HAL [81] pour les publications.

Cet entrepôt repose sur le logiciel Dataverse et propose une infrastructure adaptée au dépôt, au partage et à la préservation des données. Par ailleurs, un module catalogue est en cours de développement. Ce module permettra de signaler et moissonner les métadonnées de jeux de données externes (déposés sur un autre entrepôt), grâce à l'utilisation du protocole d'échange standard OAI-PMH, et ainsi d'éviter le double dépôt.

Lors de la création d'un espace institutionnel, celui-ci dispose de 5 To de stockage, avec une taille maximale de fichier limitée à 50 Go. C'est pourquoi les données issues de la microscopie ne sont pas adaptées à cet entrepôt. Il est préférable de privilégier des entrepôts thématiques, tels que BioImage Archive ou IDR, ce dernier étant particulièrement dédié aux données de valeur ajoutée.

Il existe plus de 3000 entrepôts où publier des données, recensées par des plateformes spécialisées comme Re3Data.org ou DataCite.

Re3Data [82] est un répertoire d'entrepôts de données de recherche maintenu par DataCite, qui fournit des informations détaillées sur les caractéristiques de chaque entrepôt, facilitant ainsi le choix en fonction des besoins spécifiques. De plus, des outils comme Cat OPIDoR [83] (catalogue de services dédiés aux données de la recherche, hébergé par l'INIST) sont également disponibles pour orienter et accompagner les chercheurs dans leurs démarches liées à la gestion et au partage des données scientifiques.

Pour le partage des données de code informatique, qu'il soit collaboratif ou non, nous considérons que chaque développeur de nos plateformes d'imagerie utilise un gestionnaire de version pour son code. Lorsqu'un code source est déposé sur GitHub ou une autre forge logicielle, il est conçu pour être partagé avec la communauté de développeurs, favorisant ainsi un échange collaboratif.

Il est recommandé, dans le cadre d'une science ouverte et pour permettre une reproductibilité des recherches publiées, de donner accès à tous les scripts et codes développés pour le projet de recherche donnant lieu à publication. Une solution robuste est de déposer ses codes sur l'infrastructure Software Heritage (SWH), et éventuellement HAL, afin de pérenniser son archivage, son accès et son référencement.

Repository	Zenodo	Figshare / Figshare+	Dryad	Bioimage Archive	Image Data Repository	Broad Bioimage Benchmark Collection	The Cell Painting Gallery
URL	https://zenodo.org/	https://figshare.com/	https://datadryad.org/	https://www.ebi.ac.uk/bioimage-archive/	https://idr.openmicroscopy.org/	https://broad.io/BBBC	https://broad.io/cellpaintinggallery
Qualifications	"All the digital artefacts"	Research data outputs	Non-human identifiable data of any kind that authors are willing to make CC0	Non-medical, non-Electron Microscopy images of any kind	"Reference image datasets" - complete, can be associated with other resources, likely to be re-analyzed	Image sets with descriptions and ground truth	Microscopic image sets suitable for image-based profiling
Also non-image data?	Yes	Yes	Yes	Not directly (BioStudies)	Not directly	No	Sometimes
Size limit?	50 GB per collection (soft)	20 GB Figshare, 5TB Figshare+ (soft)	300 GB (soft)	No	1 TB (soft)	No	No
Cost to depositor?	No ("donations encouraged")	Free to $20 GB, $395 to 100 GB, $585/250 GB beyond	$120, + $50/every 10 GB over 50 GB (some funders provide sponsorship)	No	No	No	No
Strictness of metadata requirements	None	Low	Low	Medium	High	Medium	High

Figure 4 : Mode de fonctionnement des entrepôts de données les plus populaires.
Source : https://zenodo.org/records/7628604

Un identifiant SWHID peut alors être utilisé pour citer les codes au sein des publications et communications et y donner accès de façon pérenne. SWH moissonne automatiquement les principales forges publiques régulièrement. Un dépôt ponctuel est cependant toujours possible (voir tutoriel). Comme pour le cas des données, un soin particulier apporté aux métadonnées permettra d'optimiser la réutilisabilité, visibilité et citabilité, ainsi que de rendre crédits aux auteurs. La question de la licence logiciel est également à prendre en compte. Si votre unité de recherche utilise une forge, rapprochez vous de votre supports informatique sur ce point. Pour des développements packagés sous forme de logiciel, et particulièrement pour les projets de recherche dans le cadre d'un accord de consortium intégrant des partenaires industriels, il peut être intéressant de contacter le service valorisation de votre institution afin de déterminer le type de licence recommandée. Enfin, le code peut-être parallèlement signalé et/ou déposé dans HAL, un partenariat SWH-HAL permettant d'optimiser la visibilité et la pérennité des travaux ainsi produits.

Plus d'informations sont disponibles sur le site d'ouvrir la science pour les productions logicielles [85].

RECOMMANDATIONS DU GT GEDEM

Données images : partagez vos datasets sur BioImage Archive (gratuit, pas de limite de taille d'archive, requiert un niveau d'exigence en métadonnées raisonnable) **[86]**.
Données codes sources informatiques : déposez vos codes sources sur GitHub ou Gitlab (idéalement le GitLab de l'IN2P3 **[87]** ouvert à toutes institutions avec login EduGain) avec une description de leur fonctionnement dans le fichier ReadMe.md. Nous vous conseillons également d'archiver les codes sources et logiciels dans le répertoire Software Heritage **[88]**.
Données de workflows d'analyses utilisant plusieurs outils informatiques : ces workflows peuvent être décrits, archivés et partagés sur protocol.io **[89]**. Le partage de workflow déployables (utilisables directement) est également possible sur workflowhub **[90]**.

Conclusion

La rédaction de ce guide a été motivée d'une part, par les évolutions récentes liées aux problématiques de gestion des données de la recherche dans le cadre d'une science ouverte, et d'autre part, par le regroupement et la réflexion interdisciplinaire des membres de réseaux de la MITI (GeDeM pour le RTmfm et le GT inter-réseaux DOREMITI) et d'Instituts CNRS. Nous tenons ici à remercier encore une fois le travail réalisé par le GT DOREMITI sur lequel la rédaction de ce document est très largement inspiré. Nous avons souhaité rédiger un cas d'usage précis et concret sur la gestion des données en microscopie photonique tout en rappelant aussi souvent que nécessaire les notions fondatrices de la gestion FAIR.

Les réseaux métiers et technologiques sont particulièrement actifs et investis dans la veille technologique et la diffusion de savoirs nécessaires pour une bonne gestion FAIR des données. Grâce à leurs actions, ils constituent le relai nécessaire pour diffuser les bonnes pratiques utiles pour le travail dans les laboratoires. Du fait des différentes approches, outils, concepts et vocabulaires entre nos différents métiers, nous avons retenu la solution fédératrice de relier nos actions aux étapes communément admises du "cycle de vie des données" auquel nous avons également ajouté une étape "Imaginer et Préparer" et "concevoir planifier", en les distinguant, pour bien prendre en compte les phases préparatoires de planification d'un projet. L'intérêt de cette représentation est de bien montrer toutes les étapes nécessaires pour aboutir à une publication des données, et les rendre réutilisables à l'échelle d'une équipe de recherche ou d'une plateforme technologique. Ainsi, les données de la recherche issues de nos microscopes pourront être mises à profit dans d'autres projets.

Le cycle de vie doit assurer aux données les meilleures conditions pour leur utilisation, leur archivage pérenne, et leur réutilisation pour d'autres besoins et d'autres projets que ceux pour lesquels elles ont été initialement constituées.

Outre des différences méthodologiques liées aux disciplines qu'on retrouvera dans les étapes 3 (collecter), 4 (traiter), 5 (analyser), les réseaux se retrouvent sur des concepts communs dès lorsqu'il faut préparer un projet (étapes 1,2), préserver les données (étape 6) ou publier et diffuser les données (étape 7).

Ce document est le fruit d'un travail collaboratif qui a consisté à collecter, sélectionner et mettre à disposition des ressources vers les actions phares des plateformes, enrichis d'informations et de conseils.

Ce guide n'a pas la prétention d'être exhaustif, mais il illustre les thèmes de fort intérêt de ces dernières années menés par les réseaux métiers et technologiques, et qui s'inscrivent dans la politique nationale liée à la science ouverte. Il sera complété au fil du temps par d'autres entrées et actions d'intérêt organisées par nos réseaux.

Les pratiques et conseils cités dans ce guide ne se substituent pas aux recommandations présentées par les agences de financement, les établissements, ou les instituts, ..., mais sont là pour éclairer, accompagner les personnels de la recherche en charge de la gestion des données.

Il est désormais clair qu'il faut considérer la gestion de données comme une tâche à part entière dans les projets de recherche. Il faut désormais anticiper comment les données seront acquises/collectées, stockées, diffusées et bien entendu respecter les règlements de manière à "ouvrir les données autant que possible, les fermer autant que nécessaire".

Le lecteur aura par ailleurs connaissance des infrastructures existantes. Il pourra se positionner sur les pratiques utilisées en fonction de sa discipline, faire appel et se rapprocher des réseaux métiers et technologiques pour l'aider à acquérir des bonnes pratiques de gestion des données.

Nous espérons, à travers ce guide, apporter notre pierre à l'édifice pour une meilleure prise en compte du travail consacré aux données de la recherche et pour que ces données de la recherche soient accessibles, bien documentées, réutilisables et donc réutilisées dans le cadre de la science ouverte.

Remerciements

Ce document est le fruit de l'implication de nombreuses personnes. Les auteurs remercient le GT DOREMITI qui nous a fourni une base de travail remarquable et en particulier, Alain Marois, Laure Bézard, Antoine Blanchard, et Cécile Arènes et Pierre Brochard pour leur relecture constructive.

Merci à Perrine Paul-Gilloteaux pour ses remarques toujours pertinentes.

Merci enfin au réseau RTmfm de la MITI du CNRS de permettre de mettre en place de si nombreuses et si fructueuses interactions entre ses membres.
Un merci tout particulier à Anne-Antonella Serra.

Glossaire

Analyse d'images : reconnaissance des éléments et des informations contenus dans une image.

Archivage : L'archivage est avant tout un ensemble de procédures. Cela peut concerner plusieurs types de fichiers tels que les e-mails, les documents, les vidéos, les photos, les bases de données, etc. Le rôle principal de l'archivage est de sélectionner, décrire, conserver et communiquer sur une longue durée les données devenues inutiles au quotidien. Lorsque vous disposez d'un système d'archivage, vos fichiers sont archivés en temps réel au fur et à mesure de leur création ou de leur réception.

Cahier de laboratoire électronique : outil de journal de bord utilisé pour détailler au quotidien les projets de recherche, les expériences et les protocoles utilisés.

CNRS : Centre national de la recherche scientifique

Dataset : Jeu de données issu des postes d'acquisition.

DMP : cf "PGD / DMP".

DOI : Digital Object Identification

Données (de la recherche) : enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique.

DOREMITI : Données de la Recherche de la MITI

Entrepôt de données : système de stockage digital qui connecte et harmonise de grandes quantités de données provenant de nombreuses sources.

FAIR : ensemble de principes directeurs visant à rendre les données de la recherche, Faciles à trouver, Accessibles, Interopérables et Réutilisables (FAIR) par les êtres humains et les machines.

GeDeM : Gestion des Données et Métadonnées

IN2P3 : Institut National de Physique Nucléaire et de Physique des Particules (CNRS Nucléaire et Particules)

Inserm : Institut national de la santé et de la recherche médicale

Logiciel Libre : cf. Open Source.

Métadonnée : donnée qui fournit de l'information sur une autre donnée. Il s'agit de renseignements qui sont générés automatiquement ou ajoutés manuellement afin de les restituer dans leur contexte.

Métrologie : Science de la mesure. Elle définit les principes et les méthodes permettant de garantir et maintenir la confiance envers les mesures résultant des processus de mesure.

MITI : Mission pour les initiatives transverses et interdisciplinaires

NAS : Network Attached Storage. Serveur de stockage en réseau.

OME : Open Microscopy Environnement

OMERO : logiciel client-serveur pour la visualisation, la gestion et l'analyse d'images de microscopie.

Ontologie (informatique) : En informatique et science de l'information, une ontologie informatique est un modèle de données contenant des concepts et relations permettant de modéliser un ensemble de connaissances et de données dans un domaine précis.

Open source (logiciel, code) : logiciel ou code informatique publié sous une licence dans laquelle le détenteur du droit d'auteur accorde aux utilisateurs le droit d'utiliser, d'étudier, de modifier et de distribuer le logiciel et son code source à toute personne et à toutes fins.

PGD / DMP : Plan de Gestion des Données / Data Management Plan. Document décrivant la gestion des données produites dans le cadre d'un projet de recherche. Ce document est évolutif, en général trois fois : au début, au milieu et à la fin d'un projet. Le PGD peut être dit de projet ou d'entité (anciennement structure) s'il s'agit du travail d'une plateforme par exemple.

Produit de la recherche : Dans la rédaction d'un PGD / DMP, les produits de la recherche décrivent les différents types de données qui sont générées et donc leurs gestions respectives.

RTmfm : Réseau Technologique de microscopie de fluorescence multidimensionnelle

Stockage : Le stockage des données consiste à recueillir et conserver des informations numériques dans un espace (information, application, fichier, vidéo, etc.). L'espace de stockage n'a d'autre but que d'accueillir une quantité de fichiers en tout genre. C'est une méthode utilisée principalement pour la conservation de données sur le court terme. Le stockage peut se faire sur un support physique (disques durs, clés USB, serveurs physiques), sur un support en ligne (Cloud) et/ou sur un NAS (Network Attached Store).

Traitement d'images : procédure permettant d'améliorer la qualité des images afin de faciliter les analyses.

Bibliographie

1. https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/
2. https://doremiti.cnrs.fr/guide.html
3. https://rtmfm.cnrs.fr/gt/gt-gedem/
4. https://rtmfm.cnrs.fr/
5. https://rdmkit.elixir-europe.org/bioimaging_data
6. https://en.wikipedia.org/wiki/Zen_of_Python
7. https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html
8. https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte-pnso/
9. https://doi.org/10.1007/s00418-023-02203-7
10. https://www.inserm.fr/nos-recherches/science-ouverte/
11. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000031701525/
12. https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/
13. https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000339241/2020-12-15/
14. https://www.data.gouv.fr/fr/pages/legal/licences/
15. https://www.ouvrirlascience.fr/fair-principles/
16. https://doi.org/10.1038/sdata.2016.18
17. https://qualite-en-recherche.cnrs.fr/wp-content/uploads/2021/08/guide_tracabilite_activites_recherche_gestion_connaissances.pdf
18. https://quarep.org/
19. https://www.archivistes.org/Referentiel-de-gestion-des-archives-de-la-recherche
20. https://carrieres.cnrs.fr/vos-avantages/#formation
21. https://doi.org/10.1371/journal.pbio.3002167
22. https://primacen.fr/Access
23. https://rtmfm.cnrs.fr/wp-content/uploads/2023/07/slides.html#/title-slide
24. https://dmp.opidor.fr/
25. https://dsw.france-bioinformatique.fr/wizard/?originalUrl=%2Fwizard%2Fdashboard
26. http://gbp.resinfo.org/
27. https://fr.wikipedia.org/wiki/RAID_(informatique)
28. https://isa-specs.readthedocs.io/en/latest/
29. https://isa-tools.org/

30. https://doi.org/10.1038/s41592-021-01290-5

31. https://bids.neuroimaging.io/

32. https://www.openmicroscopy.org/omero/

33. https://auth.omero-fbi.fr

34. https://www.opencid.fr/

35. https://www.youtube.com/watch?v=-8kwkwAxqMQ

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

37. https://doi.org/10.1038/s41592-021-01166-8

38. https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01166-8/MediaObjects/41592_2021_1166_MOESM1_ESM.pdf

39. https://doi.org/10.1038/s41592-023-01987-9

40. https://github.com/WU-BIMAC/NBOMicroscopyMetadataSpecs

41. https://github.com/WU-BIMAC/MicroMetaApp.github.io

42. https://doi.org/10.1038/s41592-021-01315-z

43. https://doi.org/10.1038/s41592-021-01288-z

44. https://rtmfm.cnrs.fr/wp-content/uploads/2023/07/slides.html#/title-slide

45. https://science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/gerer-des-donnees-et-des-codes/nommage-et-organisation-des-fichiers

46. https://doi.org/10.1038/s41592-018-0035-x

47. https://rtmfm.cnrs.fr/gt/gt-3m/

48. https://quarep.org/

49. https://www.elabftw.net/

50. https://www.labguru.com/

51. https://backuppc.github.io/backuppc/

52. https://www.cobiansoft.com/

53. https://www.bacula.org/

54. https://rdiff-backup.net/

55. https://github.com/ome/bioformats

56. https://github.com/Imagimm-CIML/convert-mrxs-to-ometiff

57. https://www.adobe.com/fr/creativecloud/file-types/image/raster/tiff-file.html

58. https://www.hdfgroup.org/solutions/hdf5/

59. https://ngff.openmicroscopy.org/latest/

60. https://svi.nl/Signal-to-Noise-Ratio

61. https://imagej.net/ij/

62. https://icy.bioimageanalysis.org/

63. https://napari.org/stable/

64. https://www.leica-microsystems.com/fr/produits/logiciel-du-microscope/p/aivia/

65. https://www.zeiss.com/microscopy/fr/produits/logiciel/analyse-dimages-avancee.html

66. https://www.thermofisher.com/fr/fr/home/electron-microscopy/products/software-em-3d-vis/amira-software.html

67. https://imaris.oxinst.com/

68. https://github.com/

69. https://france-bioimaging.org/about/work-packages/fbi-data/

70. https://irods.org/

71. https://www.ouvrirlascience.fr/les-principes-trust-des-entrepots-de-donnees/

72. https://arxiv.org/abs/2401.13022

73. https://www.doi.org/

74. https://doi.org/10.1038/s41592-021-01327-9

75. https://github.com/WU-BIMAC/NBOMicroscopyMetadataSpecs

76. https://doi.datacite.org/

77. https://opidor.fr/identifier/

78. https://support.datacite.org/docs/schema-mandatory-properties-v43

79. https://doranum.fr/identifiants-perennes-pid/

80. https://recherche.data.gouv.fr/fr

81. https://hal.science/

82. https://www.re3data.org/

83. https://cat.opidor.fr/index.php/Cat_OPIDoR,_wiki_des_services_d%C3%A9di%C3%A9s_aux_donn%C3%A9es_de_la_recherche

84. https://zenodo.org/

85. https://www.ouvrirlascience.fr/science-ouverte-codes-et-logiciels/

86. https://www.ebi.ac.uk/bioimage-archive/

87. https://gitlab.in2p3.fr/users/sign_in

88. https://www.softwareheritage.org/?lang=fr

89. https://www.protocols.io

90. https://workflowhub.eu/

91. https://doi.org/10.1083/jcb.202107093

92. https://calcul.math.cnrs.fr/pages/mesocentres_en_france.html

93. https://tidyr.tidyverse.org/articles/tidy-data.html

94. https://www.ouvrirlascience.fr/initiez-vous-a-la-science-ouverte/

95. https://cyber.gouv.fr/publications/guide-dhygiene-informatique

96. https://dmp.opidor.fr/plans/30160/export.pdf?export%5Bquestion_headings%5D=true

97. https://dmp.opidor.fr/plans/2389/export.pdf?export%5Bquestion_headings%5D=true

Notes

Crédits photos : © Faisal Bekkouche, Aleksandra Lawera & Julio Mateos-Langerak
Légende : «Gestion des données en microscopie photonique : de la planification à la diffusion, en assurant leur accessibilité, leur pérennité et leur réutilisation dans une démarche FAIR.»

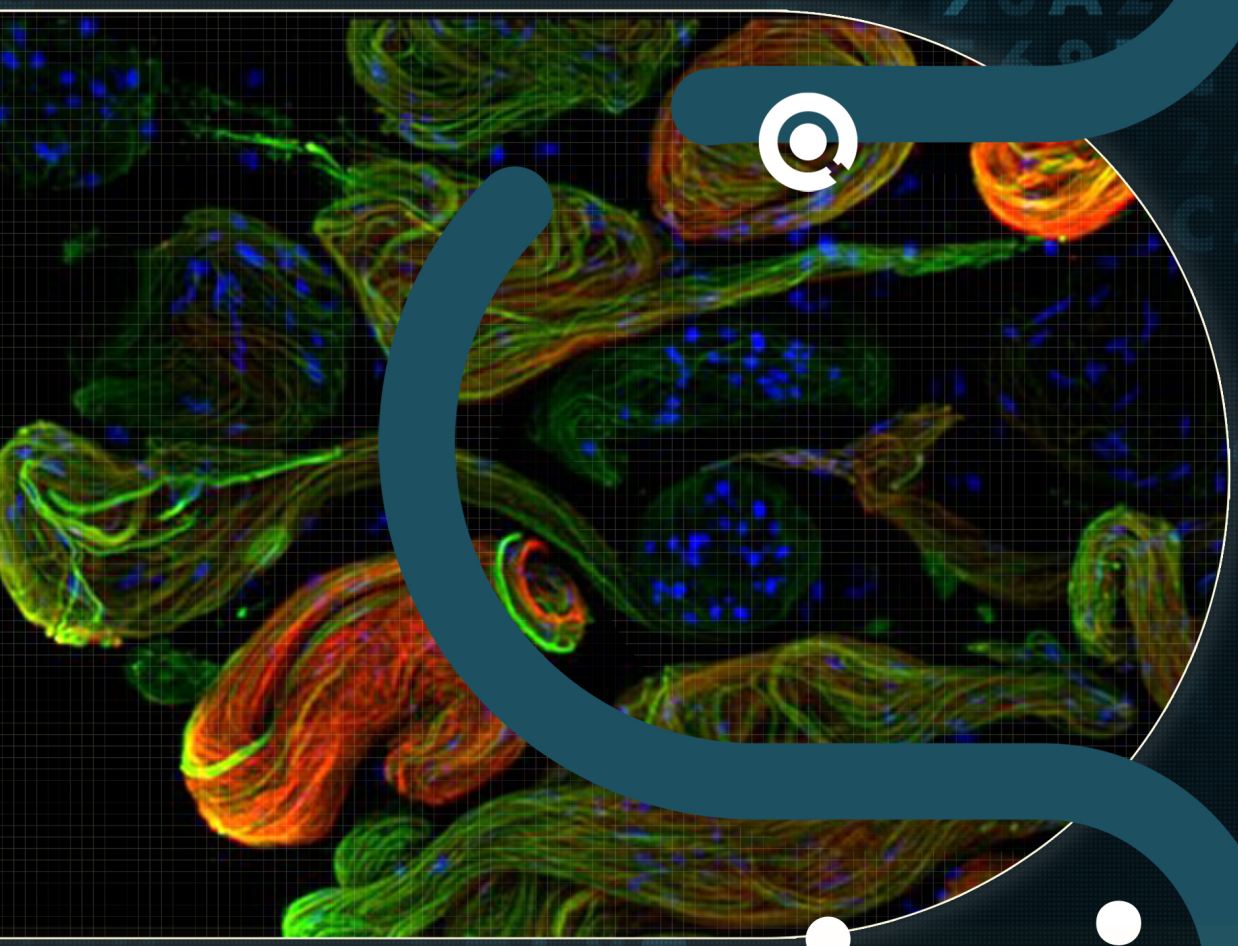
Mise en page et impression : CNRS/IFSeM/Secteur de l'imprimé/Valérie PIERRE

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cnrs

MFM
Microscopie de
Fluorescence
Multidimensionnelle

GeDe



cnrs