Guide de bonnes pratiques sur la gestion des données de la Recherche

Alain Rivet, Christine Hadrossek, Joanna Janik, Maurice Libes, Violaine Louvet, Marie-Claude Quidoz, Geneviève Romier

Guide de bonnes pratiques sur la gestion des données de la recherche

Alain Rivet
Centre de Recherches sur les macromolécules Végétales (CERMAV)
601 Rue de la chimie
CS40700
38041 Grenoble cedex 9

Christine Hadrossek
Direction des Données Ouvertes de la Recherche (DDOR)
3, rue Michel Ange
75794 Paris cedex 16

Joanna Janik
Direction des Données Ouvertes de la Recherche (DDOR)
3, rue Michel Ange
75794 Paris cedex 16

Maurice Libes
OSU Institut Pythéas
Campus de Luminy
13009 Marseille

Violaine Louvet
Grenoble Alpes Recherche Infrastructure de Calculs Intensifs et de Données (GRICAD)
Bâtiment IMAG
700, Avenue Centrale
38400 Saint Martin d'Hères

Marie-Claude Quidoz
Centre d'écologie fonctionnelle et évolutive (CEFE)
1919 Route de Mende
34090 Montpellier

Geneviève Romier
Centre de Calcul de l'IN2P3
21 avenue Pierre de Coubertin
CS70202
69627 Villeurbanne cedex

Résumé

La gestion rigoureuse et cohérente des données de la recherche constitue aujourd'hui un enjeu majeur pour la production de nouvelles connaissances scientifiques. Guidés par le « Plan National pour la Science Ouverte » qui prône la diffusion sans entrave des publications et des données de la recherche, les différents organismes de recherche et les instituts du CNRS 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.

Plusieurs réseaux métier du CNRS (Calcul, Devlog, QeR, rBDD, Renatis, Resinfo, Medici), le réseau SIST de l'Institut National des Sciences de l'Univers (INSU), accompagnés de la Direction des Données Ouvertes de la Recherche (DDOR) et de l'Institut de l'Information Scientifique et Technique (INIST) du CNRS ont publié leur « -Guide de bonnes pratiques sur la gestion des données de la recherche ». Ce guide a pour objectif la mise en commun de bonnes pratiques de gestion des données de la recherche telles qu'appliquées par les réseaux métier du CNRS dans la gestion et la valorisation des données scientifiques afin que ces données puissent être accessibles et réutilisables.

Mots-clefs

Données, FAIR[1], DMP[2]*, entrepôt, science ouverte*

1 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. Améliorer les pratiques de gestion des données de la science devient nécessaire pour garantir l'intégrité scientifique et la traçabilité de la recherche produite, mais aussi pour rendre accessibles, partager, permettre la réutilisation ou la reproductibilité des données. Guidés par le « Plan National pour la Science Ouverte », les différents organismes de recherche et les instituts du CNRS 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.

Gérer les données de la recherche est un processus complexe qui suppose un travail long et coûteux, des moyens techniques et humains parfois importants et qui comprend plusieurs étapes avant d'aboutir à la publication et l'archivage de données fiables, de qualité, respectueuses du droit des personnes et de la législation en vigueur.

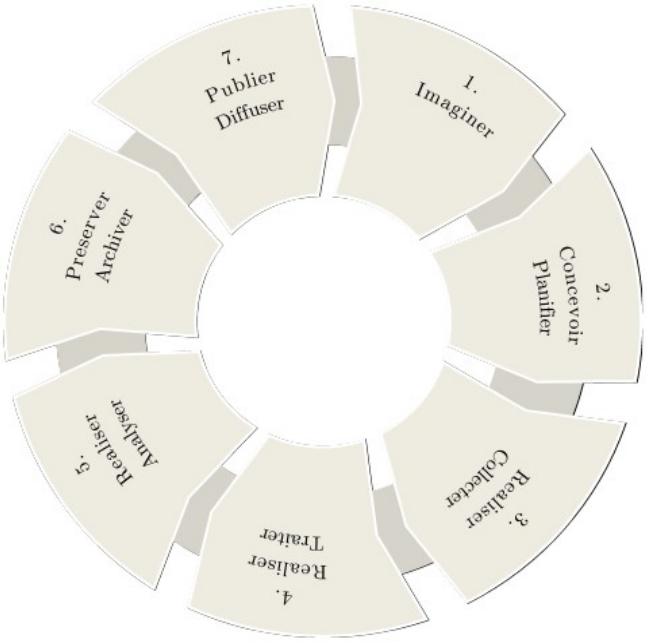

Figure 1 – Cycle de vie des données répondant aux actions des réseaux métier

1 *FAIR* pour *Findable, Accessible, Interoperable, Reusable*

2 *DMP* pour *Data Management Plan*

Pour formaliser les différentes étapes de gestion des données, nous nous sommes servis du cycle de vie des données élaboré au sein d'un groupe de travail inter-réseaux, intitulé « Atelier Données » (Figure 1). Il s'agit d'un cercle vertueux dans lequel les différentes actions des réseaux vont s'intégrer et dont les différentes étapes correspondent aux différentes phases d'un projet scientifique.

2 L'apport des réseaux métier du CNRS

Dans leurs différentes pratiques, les réseaux métier du CNRS, regroupés au sein de la Mission pour les Initiatives Transverses et Interdisciplinaires (MITI) ou soutenus par les instituts sont en première ligne pour participer à ce mouvement d'ouverture et de partage des données. Les personnels des organismes de recherche qui les constituent travaillent à la mise en place de bonnes pratiques de gestion et participent au processus de production des données scientifiques aux côtés des équipes de recherche.

C'est précisément ce travail de soutien que nous présentons dans ce document. À travers des nombreux séminaires, communications et formations présentés, ce guide fournit les meilleures pratiques du moment en matière de gestion des données.

Ce guide de bonnes pratiques sur la gestion des données de la recherche est la production du groupe de travail inter-réseaux « Atelier Données » de la MITI, composé de plusieurs réseaux métier : Calcul, Devlog, QeR, rBDD, Renatis, Resinfo, Medici, du réseau SIST (gestionnaires de données environnementales et labellisé par l'INSU), de l'INIST et de la Direction des données ouvertes de la recherche (DDOR-CNRS).

Il fournit donc un point de vue transversal intéressant et traduit les efforts et le soutien mis en place par les personnels d'appui à la recherche au sein des réseaux, dans la gestion et la valorisation des données scientifiques.

3 Objectifs du guide

L'objectif de ce guide est de fournir les bonnes pratiques en matière de gestion *FAIR* des données scientifiques, dans les différentes phases du cycle de vie de la donnée et de valoriser les apports des réseaux métier du CNRS en soutien à la Recherche.

Les réseaux métiers fournissent en effet le support d'ingénierie nécessaire pour développer les moyens et les compétences que demande une gestion *FAIR* des données, telle que décrite dans le Plan National de la Science Ouverte (PNSO). À ce titre ils :

– organisent de nombreuses actions de formation ou de sensibilisation ;

– développent des compétences et expertises en prenant appui sur des pratiques techniques et organisationnelles standardisées qui font leurs preuves sur le terrain ;

– diffusent des bonnes pratiques, recommandations, solutions techniques et organisationnelles au sein des communautés grâce à la veille technologique et juridique réalisée très régulièrement.

Dans ce Guide, nous avons donc inventorié et recensé tous les travaux réalisés au sein de nos réseaux métier qui rendent compte de la gestion des données de la recherche depuis une dizaine d'années. Nous les avons raccrochés aux différentes étapes du cycle de vie des données. Nous fournissons également au lecteur de nombreux liens vers des ressources complémentaires qui accompagnent nos travaux, et lui permettent d'approfondir le sujet.

Ce guide est donc un document un peu hybride, proche du vade-mecum, composé d'un inventaire d'actions de formation (conférences, séminaires, présentations), de liens, de recommandations

professionnelles, complétés de définitions utiles, pour approfondir le sujet de la gestion des données dans les réseaux métier.

Il s'adresse à toute personne désireuse de se former à la gestion des données de la recherche, et son objectif est d'aider le lecteur à analyser son besoin et trouver des solutions parmi l'éventail des communications qui sont présentées. Il constitue aussi une invitation à se rapprocher des réseaux métier.

Ce Guide réalisé par Pierre Navaro avec Jupyter Book est à la disposition de la communauté selon les termes de la licence *Creative Commons Attribution 4.0 International* (CC BY 4.0) sur le site de l'« Atelier-Données » (https://gt-atelier-donnees.miti.cnrs.fr).

4 Les étapes du cycle de vie

Le contenu du guide va être maintenant présenté à travers les différentes étapes du cycle de vie de la donnée. Nous nous limitons dans ce document à une brève présentation de ces différentes étapes qui sont largement détaillées dans le guide.

4.1 Imaginer et préparer

Imaginer est la première étape de notre cycle de vie. 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. C'est l'étape où l'on doit se projeter, s'informer, comprendre pour anticiper et envisager sereinement le déroulement d'un projet. C'est une étape initiale importante pour appréhender globalement la gestion des données, l'écosystème dans lequel elle s'inscrit avec ses contraintes et opportunités, les outils et infrastructures disponibles ou nécessaires, les politiques d'accompagnement et la multiplicité des acteurs qui interagissent, les réglementations en vigueur ou encore les compétences et expertises à acquérir.

L'apport des réseaux métier est ici important en termes de croisement des disciplines et des métiers pour apporter un éclairage global dans les évolutions des métiers et des compétences pour répondre au mieux aux besoins des communautés scientifiques.

4.2 Concevoir et planifier

Dans cette étape du cycle de vie de la donnée, il s'agit de définir les tâches à accomplir pour réaliser le projet de recherche, élaborer un planning, rechercher d'éventuels partenaires et financements, élaborer les spécifications nécessaires (i.e. définir précisément les éléments fonctionnels et techniques souhaités), définir les données et les métadonnées qui seront utiles, penser au futur plan de diffusion et bien d'autres actions de préparation et de planification.

Pour ces travaux de conception et de planification, les réseaux métier apportent un appui sur la gestion de projet, les méthodologies de conduite de projet permettant par exemple la définition des indicateurs utiles au projet, les outils pour assurer l'interopérabilité des systèmes mis en œuvre.

Ils fournissent des recommandations et des retours d'expérience pour la rédaction des plans de gestion de données, pour la définition du type de données à collecter, l'identification de nouveaux supports de publication, etc.

À ce stade, il est aussi nécessaire de prévoir le mode de collecte et de stockage afin d'organiser la traçabilité en amont, traçabilité qui permettra de garantir la réutilisation des données.

4.3 Collecter

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*) avec leurs métadonnées descriptives. Il s'agit donc, dans cette phase, de travailler sur les processus d'acquisition des données qui peuvent être obtenues au moyen de divers médias selon le domaine étudié : capteurs environnementaux, instruments, sondages, modèles numériques, etc. Une fois les données acquises, il est nécessaire et indispensable dans l'objectif de les rendre *FAIR*, de les décrire avec leurs métadonnées associées.

La description de ces jeux de données nécessite d'utiliser, autant que faire se peut, des référentiels de vocabulaires contrôlés (thésaurus), si possible standardisés et les plus appropriés au domaine étudié. Il est conseillé de gérer les jeux de données dans un environnement technique qui permette d'assurer la sauvegarde, l'archivage, l'accessibilité et l'interopérabilité des données. Cette gestion se fait via des infrastructures techniques, des bases ou des supports qui doivent être fiables et bien documentés, et ce dans le respect des règles de traitement spécifiques des données personnelles.

Cette phase « Collecter » nécessite :

– de disposer des données et de fournir les métadonnées nécessaires pour apporter toutes les informations utiles à la description des données brutes elles-mêmes (libellés des paramètres, unités de mesure, localisation, propriétaires, etc.), ainsi que sur les dispositifs d'acquisition (capteurs de mesures, modèles numériques, etc.) ;

– de mettre en place des chaînes de collecte, du capteur jusqu'aux espaces disques et aux applications sur des serveurs où les traitements pourront être réalisés, avec la documentation adaptée ;

– d'utiliser des protocoles si possibles normalisés ou standardisés pour présenter les données brutes et les dispositifs d'acquisition (capteurs) et les rendre interopérables ;

– de mettre en place une gestion et conduite de projets pour faire travailler ensemble les différents acteurs intervenant dans la chaîne de collecte : électroniciens, informaticiens, chercheurs, etc. ;

– de disposer de cahiers de laboratoire, tablettes de terrain ou supports divers pour consigner les relevés et métadonnées observées ;

– de définir le stockage nécessaire à la collecte de données.

4.4 Traiter les données

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

Cette partie du guide est structurée en différentes sections décrivant cette préparation des données :

– préparer les fichiers de données, en vue de leur analyse, en utilisant des formats interopérables ;

– utiliser des infrastructures logicielles (*framework*) d'intégration de données, lorsqu'elles sont hétérogènes ;

– mettre en place et utiliser des plateformes de gestion de données locales, en vue de leur analyse ;

– vérifier et s'assurer de la qualité des données.

4.5 Analyser

Derrière le terme « Analyser », on entend l'extraction de l'information des données, le plus souvent par l'utilisation de puissance de calcul. Cela recouvre de nombreux types de techniques (calcul intensif, traitement statistique, *machine learning*, visualisation, etc.), et nécessite des plateformes adaptées.

Cette étape du cycle de vie de nombreuses données impose que ces données soient exploitables, c'est-à-dire bien organisées, dans des formats adaptés à l'analyse envisagée, de façon à pouvoir leur appliquer des traitements automatisés.

Plusieurs événements récurrents, annuels ou bisannuels, comme les JCAD (Journées Calcul et Données), les JDEV (Journées du DEVeloppement logiciel) par exemple, auxquels participent activement les réseaux métier, proposent de nombreuses interventions sur ces différentes thématiques, allant de la description des plateformes aux outils disponibles, en passant par l'organisation des développements et la reproductibilité, détaillée dans la section Reproductibilité du guide. Ils incluent aussi très souvent des retours d'expérience particulièrement riches.

4.6 Préserver et archiver

Cette sixième étape du cycle de vie rend compte de l'importance de préserver et archiver les données sur le long terme. On s'attache, dans cette partie, à bien définir et clarifier les termes, réfléchir aux données pertinentes à préserver et voir quelles solutions s'offrent à nous.

Préserver, sécuriser l'information et sauvegarder, voire archiver les données sont en effet des phases essentielles de la gestion rigoureuse des données, mais il n'est pas toujours aisé de faire la distinction entre ces notions et d'utiliser le bon terme et la procédure associée. De plus, préserver pour un usage futur dont on ignore le plus souvent les caractéristiques est compliqué. Des conseils appropriés pour sélectionner les données à préserver et pour mener à bien la préservation d'objets numériques sont présentés, accompagnés par divers retours d'expérience.

4.7 Publier et diffuser

Cette dernière étape d'un projet de recherche représente en quelque sorte la finalité de toute politique de gestion de données puisqu'elle vise, dans un contexte de science ouverte, à publier et diffuser les données de manière à ce qu'elles soient accessibles et réutilisables selon des formats et des processus interopérables.

L'accompagnement des réseaux métier s'exerce sur le processus de publication des données dans des entrepôts ou des plateformes techniques, pour en permettre l'accès, ainsi que sur la documentation des données avec des métadonnées descriptives provenant de vocabulaires contrôlés et de leurs formats d'exploitation pour en assurer la réutilisabilité. Ainsi, les réseaux travaillent sur l'ensemble des informations (données, métadonnées, modes opératoires, échantillons, publications, visualisation et interfaces graphiques) nécessaires à la mise en œuvre des supports de diffusion et de valorisation les plus 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 lors du dépôt dans des entrepôts de données.

La dernière question à se poser concerne la fin de vie de ces données. Rappelons qu'au sein d'un établissement public, les données et documents produits dans l'exercice des activités du personnel constituent des données et documents relevant du Code du Patrimoine et appartiennent ainsi à l'établissement. De ce fait, Il existe des durées officielles établi par la section des archivistes en universités, rectorats, organismes de recherche et mouvements étudiants (AURORE) au sein de l'association des archivistes français à travers un « référentiel de gestion des archives de la recherche.

5 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 et d'Instituts du CNRS. Les réseaux métier sont en effet particulièrement actifs et investis dans la veille technologique et la diffusion de savoirs nécessaires pour une gestion *FAIR* des données. Grâce à leurs actions, ils constituent le relais nécessaire pour diffuser les bonnes pratiques utiles pour le travail dans les laboratoires. En ce sens, à travers les multiples séminaires et formations organisés, ils sont vecteurs de bonnes pratiques et de diffusion du savoir dans la gestion des données de la recherche.

La vie des données ne doit pas s'arrêter avec le projet car l'objectif est que ces données puissent être réutilisées dans le temps, parfois même dans des domaines autres que celui dans lequel elles ont été produites à l'origine. Les étapes du cycle de vie permettent alors d'assurer aux données les meilleures conditions à leur bonne utilisation, à leur archivage pérenne ainsi qu'à leur réutilisation pour d'autres besoins et d'autres projets.

L'originalité de ce document est qu'il rassemble la majeure partie de la production des réseaux métier de ces dernières années, dans le cadre de la gestion des données de la recherche. Ce document est donc un condensé des actions menées autour de la gestion des données de la recherche. Il est le fruit d'un travail collaboratif qui a consisté à collecter, sélectionner et mettre à disposition des ressources vers les actions phares des réseaux métier, enrichi d'informations et de conseils. 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 et accompagner les personnels de la recherche en charge de la gestion des données.

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 métier et qui s'inscrivent dans la politique nationale liée à la science ouverte. Il sera complété au fil du temps par d'autres thèmes et actions d'intérêt organisées par les réseaux. Il est clair désormais qu'il faut considérer la gestion des données comme une tâche à part entière dans les projets de recherche, et 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, afin que ces données puissent être accessibles, bien documentées, réutilisables et donc réutilisées dans le cadre de la science ouverte.

Cette présentation se veut donc une invite à prendre connaissance et à s'approprier le « guide de bonnes pratiques sur la gestion des données de la recherche ».

6 Bibliographie

[1] Hadrossek C., Janik J., Libes M., Louvet, V., Quidoz,M-C., Rivet, A., Romier G, . Guide de bonnes pratiques sur la gestion des données de la recherche. 2021 : https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html.

[2] Ministère de l'Enseignement Supérieur de la Recherche et de l'Innovation. Deuxième plan national pour la science ouverte (2021-2024) : https://www.ouvrirlascience.fr/wp-content/uploads/2021/06/Deuxieme-Plan-National-Science-Ouverte_2021-2024.pdf

[3] Orange C. Eyraud P. Morlock E. Renault S. Martin C. Rivet A. Band-Foissac O., Libes M. Quidoz M-C. Romier G. Cadiou A. Bonnet L. ; Cartographie de l'action des réseaux en matière de gestion des données de la recherche. 2017 : https://gt-atelier-donnees.miti.cnrs.fr/download/GTInterreseaux-CartoSyntheseV6-optimise.pdf

Les réseaux métier

Atelier-Données : https://gt-atelier-donnees.miti.cnrs.fr

Calcul. Réseau Calcul : https://calcul.math.cnrs.fr/

Devlog. Réseau du développement logiciel : http://devlog.cnrs.fr/

Qe. Réseau Qualité en Recherche : https://qualite-en-recherche.cnrs.fr/

rBDD. Réseau Bases de Données : https://rbdd.cnrs.fr/

Renatis. réseau des professionnels de l'Information Scientifique et Technique : https://renatis.cnrs.fr/

*Resinfo. Fédération des réseaux métiers d'**A**dministrateurs **S**ystèmes et **R**éseaux dans l'Enseignement et la Recherche : https://resinfo.org/*

Medici. Réseau des métiers de l'édition scientifique publique : http://medici.in2p3.fr/

SIST. Séries Interopérables et Systèmes de Traitements : https://sist.cnrs.fr/

DDOR. Direction des Données Ouvertes de la Recherche : https://www.science-ouverte.cnrs.fr/ddor-cnrs-direction-des-donnees-ouvertes-de-la-recherche/

INIST. Institut de l'Information Scientifique et Technique du CNRS : https://www.inist.fr/