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

v1.0 - Janvier 2021

https://mi-gt-donnees.pages.math.unistra.fr/guide

Vous êtes invités à poser vos questions sur le chat, nous y répondrons à la fin de l'exposé

Originalité

L'originalité de ce guide réside dans son application aux données de la recherche sous l'angle des pratiques de différents métiers de la recherche :

- Il fournit un point de vue transversal à travers une compilation de diverses pratiques métiers. Il présente :
  - les nombreuses **actions de formation** ou de sensibilisation des réseaux ;
  - les compétences et expertises développées issues de **pratiques standardisées** qui font leurs preuves sur le terrain ;
  - des recommandations et des solutions techniques et organisationnelles grâce à la **veille technologique et juridique** réalisée très régulièrement.

- Il traduit les efforts et le soutien mis en place par les membres des réseaux, dans la gestion et la valorisation des données scientifiques.

Origine

Ce guide est la production du groupe de travail inter-réseaux « Atelier Données » : groupe composé (en 2016) de plusieurs réseaux de la "Mission pour les Initiatives Transverses Interdisciplinaires" (MITI), et de réseaux d'Instituts du CNRS :

- Calcul : réseau pour la communauté du calcul
- Devlog : réseau national des développeurs en logiciel
- Medici : réseau des métiers de l'édition
- QeR : réseau Qualité en Recherche
- rBDD : réseau Bases de données
- Renatis : réseau des professionnels de l'information scientifique
- Resinfo : réseaux des administrateurs systèmes et réseaux
- INIST-CNRS : Institut de l'Information Scientifique et Technique
- SIST : réseau INSU des gestionnaires de données environnementales
- DDOR-CNRS : Direction des données ouvertes de la recherche

Une initiative des réseaux MITI

De par leurs missions, les membres des réseaux répondent aux besoins des communautés scientifiques :

- participent à la réflexion et à la mise à disposition des outils, méthodes et infrastructures en matière de gestion et de partage des données scientifiques,
- conseillent et mettent en place de bonnes pratiques,
- organisent des formations et journées d'études.

A travers ces actions nous voulons témoigner des activités de soutien des réseaux, et fournir les meilleures pratiques du moment en matière de gestion des données.

Contexte national

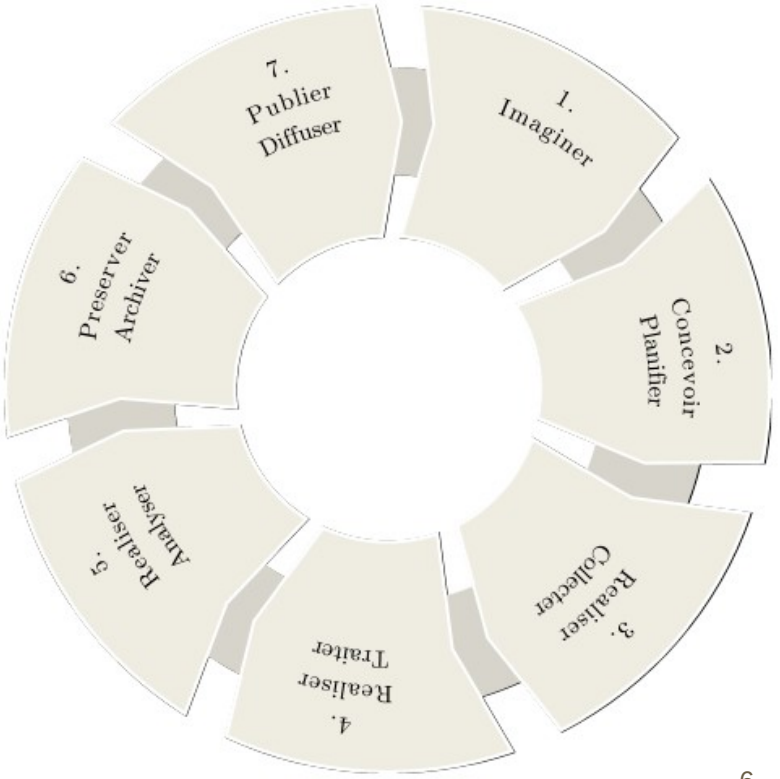
Ces initiatives et actions des réseaux concordent avec les initiatives nationales :

- le “Plan National pour la Science Ouverte” (2018),
- le Noeud National RDA-France (2018),
- la Feuille de route du CNRS (2019),
- le “Plan Données de la recherche du CNRS » (2020) qui fournissent les orientations en matière de gestion “FAIR” et d’ouverture des données.

Contenu : cycle de vie des données

Pour adopter un point de vue commun aux différents métiers et activités de nos réseaux :

- Nous nous basons sur le cycle de vie des données :
  - le cycle de vie des données représente un cadre structurant et fournit un vocabulaire commun
- Le guide fournit une lecture nouvelle des actions des réseaux, enrichie des approches complémentaires des pratiques des différents réseaux



Le guide

2.3.2. Créer un plan de gestion de logiciel

Les logiciels sont aussi des données, un peu particulières et qui méritent donc un modèle approprié de plan de gestion : le plan de gestion de logiciel. Le projet PRESOFT propose un modèle adapté à la fois au logiciel et au contexte de la recherche en France. Après une présentation de ce contexte, du modèle et de la procédure associée, les apports de PRESOFT sont détaillés. À noter que le modèle proposé par PRESOFT s'étend sur l'ensemble de la « vie » du logiciel depuis l'idée, avec les documents préparatoires, jusqu'à la préservation (sous toutes ses formes) et qu'il prend en compte toutes les formes de financement (projets, stages...). Le modèle est disponible sur DMP OPIdOR.

Plans de gestion de logiciels

Geneviève Romier, Vincent Breton, CNRS-IN2P3
JCAD 2019

2.3.3. Retour d'expérience

Afin de conclure ce tour d'horizon des plans de gestion de données, ce retour d'expérience relatif au domaine de la biodiversité vous permettra de mieux comprendre comment utiliser les plans de gestion de données comme un véritable outil de gestion qui va bien au-delà du document administratif nécessaire à la validation du projet.

Du Plan de Gestion des Données au Datapaper : suivi des données scientifiques tout au long de leur cycle de vie.

HEINTZ, Wilfried, INRA Dynafor
SIST 2018

Annexes au guide

Pour faciliter la consultation du guide les différentes sections sont complétées par deux sections annexes qui traitent des sujets transverses à toutes les étapes du cycle de vie de la données :

- les infrastructures
- la reproductibilité

Conclusions

- Ce guide n'est pas exhaustif puisqu'il est le reflet des sujets abordés dans le cadre des actions des réseaux impliqués dans la rédaction du guide

- Enrichissez le guide des pratiques métiers d'autres réseaux : nous invitons d'autres réseaux à nous rejoindre et participer à la prochaine version … en apportant leurs pratiques métiers dans le cadre de la gestion des données

- Rejoignez et participez aux activités des réseaux : le blog RH du CNRS en recense un certain nombre dans son billet « Evoluer, échanger, innover : les réseaux professionnels du CNRS ».

Conclusions

- Ce guide vise à améliorer les pratiques de gestion des données de la science pour :
  - garantir l'intégrité scientifique et la traçabilité de la recherche produite,
  - rendre accessible, partager, permettre la reproductibilité et la réutilisation des données de la recherche : données FAIR

- Les réseaux apportent un fort soutien basé sur une expérience de terrain pour atteindre ces objectifs

contact : ***contact-guide@services.cnrs.fr***

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

v1.0 - Janvier 2021

Merci pour votre attention et place aux questions

https://mi-gt-donnees.pages.math.unistra.fr/guide

1. Imaginer - Préparer

“Imaginer” est la première étape de notre cycle de vie des données.

- phase *préparatoire* qui correspond à *l’identification des problématiques techniques et juridiques* associées à la gestion des données

- L’apport des réseaux est ici important en termes de croisement des disciplines et des métiers pour *apporter un éclairage global et répondre au mieux aux besoins des communautés scientifiques :*
  - s’informer, comprendre pour anticiper et envisager le déroulement d’un projet.
  - connaître les *contraintes et opportunités, les outils et infrastructures disponibles, les politiques d’accompagnement, les acteurs, les réglementations en vigueur ou encore les compétences et expertises à acquérir.*

2. Concevoir - Planifier

Dans cette étape, on *définit les tâches à accomplir pour réaliser le projet de recherche, élaborer un planning, rechercher d'éventuels partenaires et financements, et élaborer les spécifications nécessaires*

Pour ces travaux de conception et de planification, les réseaux *apportent un appui sur la gestion et les méthodologies de conduite de projet*, et conseillent et mettent en place *des outils pour assurer* ***l'interopérabilité*** *des systèmes mis en oeuvre :*

- Recommandations et des retours d'expérience pour ***commencer la rédaction de plans de gestion de données (DMP)***
- I*dentification des infrastructures adaptées au projet* (fonctionnalités, capacités et services fournisseur du service)
- *Mise en place du mode de collecte et de stockage* afin d'organiser la traçabilité en amont, traçabilité qui permettra de garantir la réutilisation des données

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, avec leurs métadonnées descriptives.

Il s'agit donc, dans cette phase :

- de *travailler sur les processus d'acquisition des données* obtenues : capteurs environnementaux, instruments, sondages, modèles numériques
- d'assurer la traçabilité des données : cahiers de laboratoires, tablettes de terrain…
- de rendre ces données « FAIR » en les décrivant et en y associant des métadonnées, en *utilisant des normes et des standards (thésaurus, vocabulaire contrôlés…) afin que les données soient interopérables*
- se prémunir des pertes, en stockant et sauvegardant les données

4. Traiter

Cette phase correspond au *prétraitement des données brutes issues des acquisitions et des collectes.*

Il s’agit souvent de :

- *regrouper, choisir, qualifier les données pertinentes* puis les *transformer dans des formats standards interopérables*, et les préparer en vue de leur analyse ultérieure.
- Utiliser des infrastructures logicielles , services d’intégration de données (“*framework*”), 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

5. Analyser

L’étape d’analyse des données correspond à *l’extraction de l’information des données traitées.*

Cela recouvre de nombreux types de techniques : *calcul intensif, traitement statistique, machine learning, visualisation* …, ce qui peut nécessiter également des plateformes de traitement adaptées.

Cette étape du cycle de vie *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.

6. Préserver - Archiver

Sauvegarder, préserver, sécuriser l'information et, voire archiver les données sont des phases essentielles de la gestion rigoureuse des données.

Les notions de *stockage*, de *sauvegarde* et d'archivage ainsi que les actions de *préservation* et de *pérennisation* revêtent des *notions et des sens et des pratiques différentes* que nous explicitons dans le Guide.

Cette étape nécessite une *phase de sélection des informations pertinentes (validées, utiles…),* tout en se préoccupant de leur exploitation future à travers les *problématiques de durée de vie, de confidentialité et de sécurité des données.*

7. Publier et Diffuser

Cette étape consiste 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 s'exerce sur :

- l*e processus de publication des données dans des "**catalogues**", des "**entrepôts**" ou des plateformes techniques, pour en permettre l'accès*,
- la documentation des données avec des métadonnées descriptives provenant de vocabulaires contrôlés et de leurs formats d'exploitation pour en assurer la réutilisabilité.
- 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
- L*'identification des données via des **identifiants pérennes**, lors du dépôt dans des entrepôts de données.*
- La publication de "*Datapaper*" pour valoriser et expliciter en détail les données

Le paysage des infrastructures destinées à la recherche scientifique est vaste :

- *Infrastructures de données et de **stockage,** “data lake”, Infrastructures de **calcul**, Infrastructures de **travail collaboratif,** Infrastructures **logicielles*** , etc...

Cette section du Guide est destinée à la *présentation d’infrastructures européennes, nationales, thématiques dans différents domaines.*

Ces infrastructures font régulièrement l’objet de présentations qui permettent de comprendre leur organisation, leur mode de fonctionnement, et de suivre leurs évolutions.

Le Guide fournit les bons pointeurs et points d’entrée sur *EOSC, France Grilles, GENCI, Mésocentres, Cines, Infrastructure de recherches, Data Terra, etc.*

Annexe - Reproductibilité

Cette annexe, transversale à toutes les parties du Guide, rassemble des exposés concernant la reproductibilité et la répétabilité des mesures, des expériences, des calculs, etc.

on y aborde :

- les enjeux et les défis liés à la reproductibilité : La confiance dans les résultats de la recherche repose, entre autres, sur le fait que les expériences ou les calculs soient reproductibles.
- L'utilisation d'Environnements de travail qui favorisent et permettent la reproductibilité
    - bibliothèque Python Execo et le logiciel OpenMole, Guyx, Jupyter , etc.

Crédits

Auteurs

- Christine Hadrossek : DDOR
- Joanna Janik : DDOR
- Maurice Libes : réseau SIST
- Violaine Louvet : réseau Calcul
- Marie-Claude Quidoz : réseau rBDD
- Alain Rivet : réseau QeR
- Geneviève Romier : réseau rBDD

Relecteurs

- Pierre Brochard : réseau DevLog
- Dominique Desbois : réseau DevLog
- Emilie Lerigoleur : réseau SIST
- Caroline Martin
- Pierre Navaro : réseau Calcul

Edition Web

- Pierre Navaro : réseau Calcul