QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

19ème ÉCOLE INTER-ORGANISMES
QUALITÉ ET RESPONSABILITÉ SOCIÉTALE EN RECHERCHE ET EN ENSEIGNEMENT SUPÉRIEUR

du 13 au 15 Septembre 2021

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

Alain RIVET, responsable Qualité – Système d'information, CERMAV-CNRS

Contexte national

Nombreuses initiatives nationales :

- le “Plan National pour la Science Ouverte” (2018),
    - Annoncé par Frédérique Vidal, le 4 juillet 2018, rend obligatoire l’accès ouvert pour les publications et pour les données issues de recherches financées sur projets.

- 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.

Pyramide des données

Tutoriel Form@doc (Schéma adapté de Report on integration of data and publications. Opportunities for Data Exchange (Reilly S. et al., 2011))

Croissance des données scientifiques

Forte croissance de la production de données scientifiques

- La plupart des disciplines se sont mises a produire massivement des données
- Riches en information car structurées suivant un plan de recherche et une démarche scientifiques
- Englobent des connaissances uniques « Time stamped »

Les données numériques deviennent un enjeu majeur de la recherche

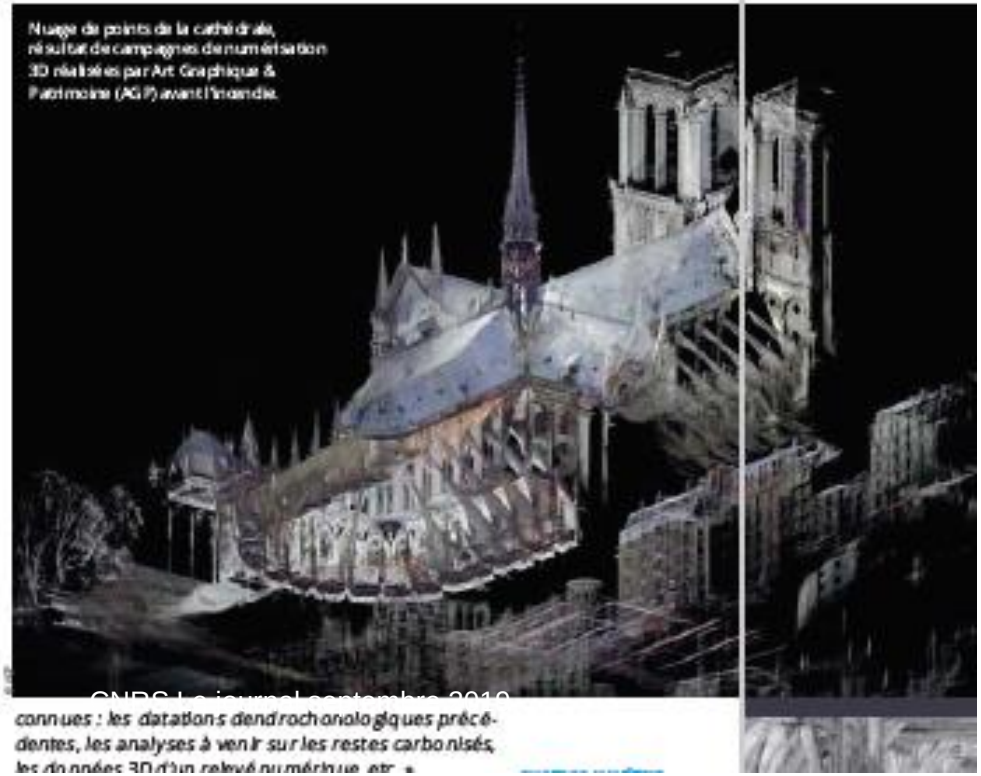
Notre dame

Notre-Dame et son double numérique

Créer une sorte de Google Earth de la cathédrale de Paris, tel est l'ambitieux projet d'une équipe de chercheurs. Objectif : regrouper au sein d'une plateforme collaborative la totalité des connaissances passées et à venir sur l'édifice. Explications.

« La restauration de Notre-Dame de Paris sera un chantier historique. » Et Livio de Luca, directeur du laboratoire Modèles et simulations pour l'architecture et le patrimoine (MAP)[1], et lauréat 2019 de la médaille de l'innovation (lire p. 10), ne souhaite pas que cette histoire se perde à nouveau dans les flammes. *« Nous allons créer un système d'information intégrant toutes les données scientifiques et techniques sur la cathédrale »*, explique-t-il.

Le groupe de travail qu'il coordonne s'apprête ainsi à en réaliser une sorte de « double numérique » rassemblant tout ce que l'on sait de l'édifice, des croquis de construction jusqu'au relevé 3D de son état actuel, et qui sera capable d'intégrer toute information à venir. En fait d'une simple réplique en images de synthèse, il s'agit plutôt de construire une base de données et de connaissances inédite. *« Grâce à elle, poursuit-il, les différentes équipes du chantier pourront partager leur expertise et prendre in fine de meilleures décisions. »*



Nuage de points de la cathédrale, résultat de campagnes de numérisation 3D réalisées par Art Graphique & Patrimoine (AGP) avant l'incendie.

FÉDÉRER LES DONNÉES

Traitement permettant de rassembler plusieurs banques de données hétérogènes via un modèle commun aux yeux de l'utilisateur.

connues : les datations dendrochronologiques précédentes, les analyses à venir sur les restes carbonisés, les données 3D d'un relevé numérique, etc. »

CNRS Le journal septembre 2019

Publications à long terme

C. Diaconu, Fredocs, Aussois, 2013

Charte de déontologie

« La description détaillée du protocole de recherche dans le cadre des cahiers de laboratoire,-ou de tout autre support, doit permettre la traçabilité des travaux expérimentaux »

« Tous les résultats bruts (qui appartiennent à l'institution) ainsi que l'analyse des résultats doivent être conservés de façon à permettre leur vérification. »

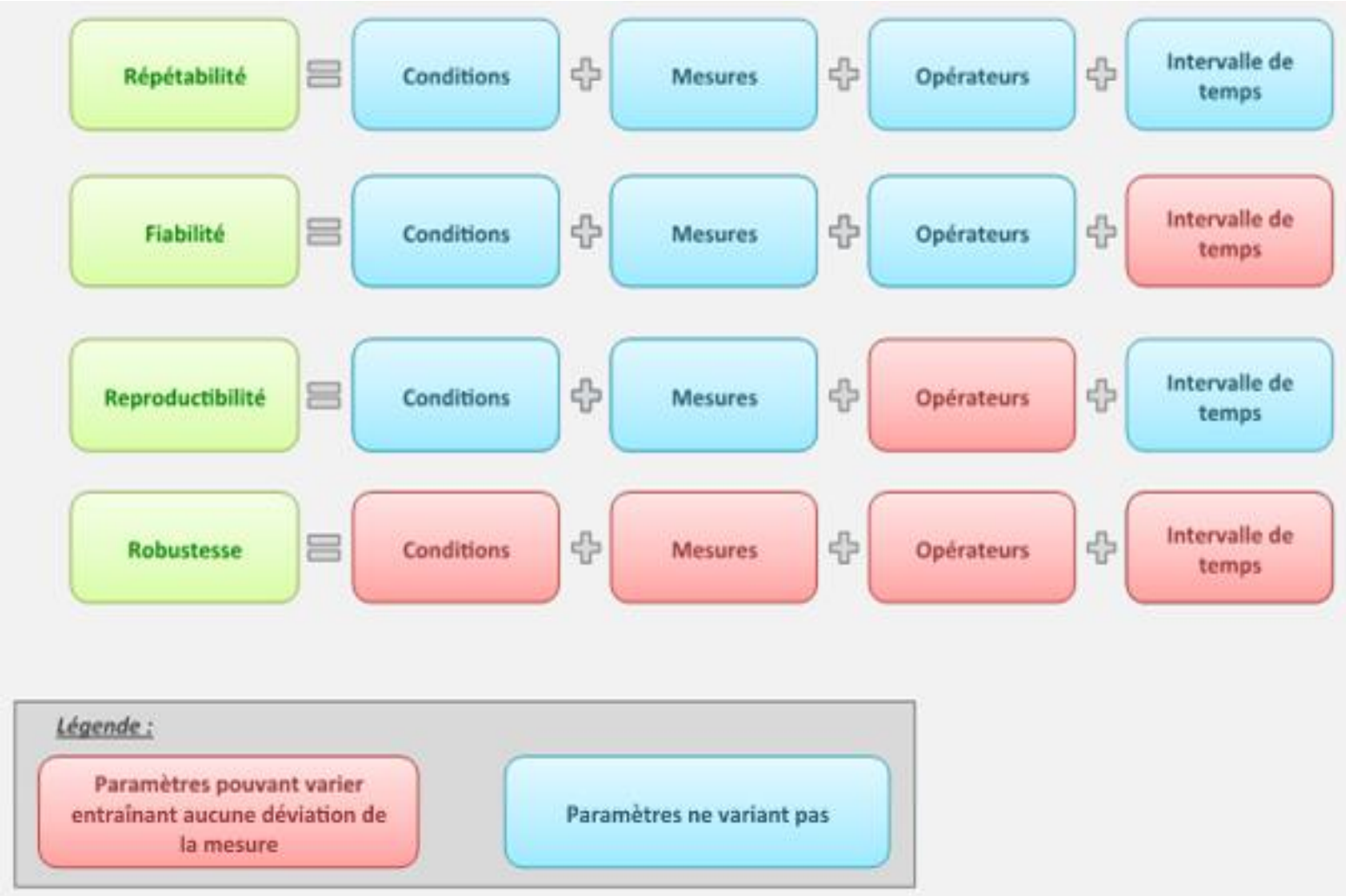
https://www.hceres.fr/fr/CharteFrancaiseIntegriteScientifique

Qualité des données

Ouverture des données oui mais…

des données de qualité

Qualité intrinsèque des données



Management de la qualité : l'assurance d'une recherche robuste et fiable, Réseau Inserm Qualité

Qualité des jeux de données

Plusieurs éléments permettent d'évaluer le niveau de qualité d'un jeu de données :

- Des éléments sur les données elles-mêmes et leur structure : format de fichier, structure du fichier, contenu.

- Des éléments attestant du potentiel de réutilisation et de croisement des données :
  - Le respect de standards, référentiels et schémas déjà établis ;

- Des éléments qui accompagnent les données :
  - documentation claire et rigoureuse, gestion des versions et des mises à jour des données…

Atelier données

Une initiative des réseaux...

... accompagnée par la MITI...

... complétement en ligne avec le plan données de la recherche du CNRS.

Plate-forme réseaux de la MITI

Les différents réseaux métiers et technologiques de la MITI ont pour point commun de fédérer une population autour d'un métier ou d'une technologie.

Les réseaux et la plateforme en quelques chiffres :

- 20 réseaux nationaux et près de 13000 adhérents (dont ~30% de chercheurs et chercheuses),
- 59 réseaux régionaux associés,
- ~50 journées thématiques nationales ou ateliers techniques par an,
- ~30 actions nationales de formation par an

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 avons voulu témoigner des activités de soutien des réseaux, et fournir les meilleures pratiques du moment en matière de gestion des données.

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

Lettre de mission du GT Atelier Données

- Construire et diffuser une vision transversale de la gestion des données afin d'enrichir la pratique de chaque réseau dans le domaine des données et permettre le développement de la complémentarité entre réseaux.
- Valoriser l'apport des expériences et expertises « métier » dans une vision transversale de gestion de données dans les réseaux technologiques et scientifiques de la MI
- Sensibiliser les communautés professionnelles de l'appui à la gestion des données (organisation de journées thématiques par exemple) ;
- Identifier les problématiques concernant les « data » dans chaque réseau (livrables à définir).
- Mettre en commun et partager de nouvelles pratiques en réseau et au sein de chaque réseau.

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.

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 : collecter

Rechercher dans ce livre ...

3.2.5. Les cahiers de laboratoire

L'ensemble des données produites par la recherche doit être répertorié et enregistré dans l'objectif d'une réutilisation potentielle. Nous disposons pour ce faire d'un certain nombre de supports comme les cahiers de laboratoire. 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 de ses inventeurs. Les plaquettes du réseau CURIE "Le cahier de laboratoire national : Pourquoi l'utiliser ?" et "Le cahier de laboratoire national : Comment l'utiliser ?" présentent des recommandations sur la bonne gestion de ce dernier.

Alain Rivet positionne le cahier de laboratoire comme un outil de gestion des données de la recherche :

> *Cahier de laboratoire et gestion des données de la recherche*
>
> *Alain Rivet, CERMAV*
> *Atelier Dialog'IST « Rendre FAIR les données, mais quelles données préserver ? », réseau Renatis, 2020*

Le guide : Préserver et archiver

Archiver

> L'archivage consiste à ranger un document dans un lieu où il sera conservé pendant une période plus ou moins longue et d'y associer les moyens pour réutiliser les données : la réutilisation se faisant en ajoutant de l'intelligence à la sauvegarde. Le contenu des documents archivés n'est pas modifiable. Par contre le contenant (format) des documents archivés peut être modifié (pour éviter l'obsolescence logicielle).

Le terme archive est défini par le législateur : *les archives sont l'ensemble des documents, y compris les données, quels que soient leur date, leur lieu de conservation, leur forme et leur support produits ou reçus par toute personne physique ou morale, et par tout service ou organisme public ou privé dans l'exercice de leur activité* (art. L. 211-1 du code du patrimoine). Les données de la recherche entrent pleinement dans le périmètre des archives.

Pour en savoir plus sur le statut des archives scientifiques du CNRS et sur leur délai de conservation, nous vous conseillons ces deux documents :

Textes réglementaires et durée de conservation

Marie-Laure Bachèlerie, DAJ-CNRS
Séminaire « Archivage Numérique des Données de Recherche ».

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

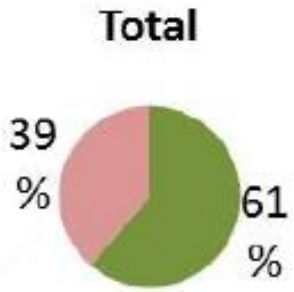
Sauvegarder, préserver, sécuriser l'information et, voire archiver les données sont des phases essentielles de la gestion rigoureuse des données.

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.

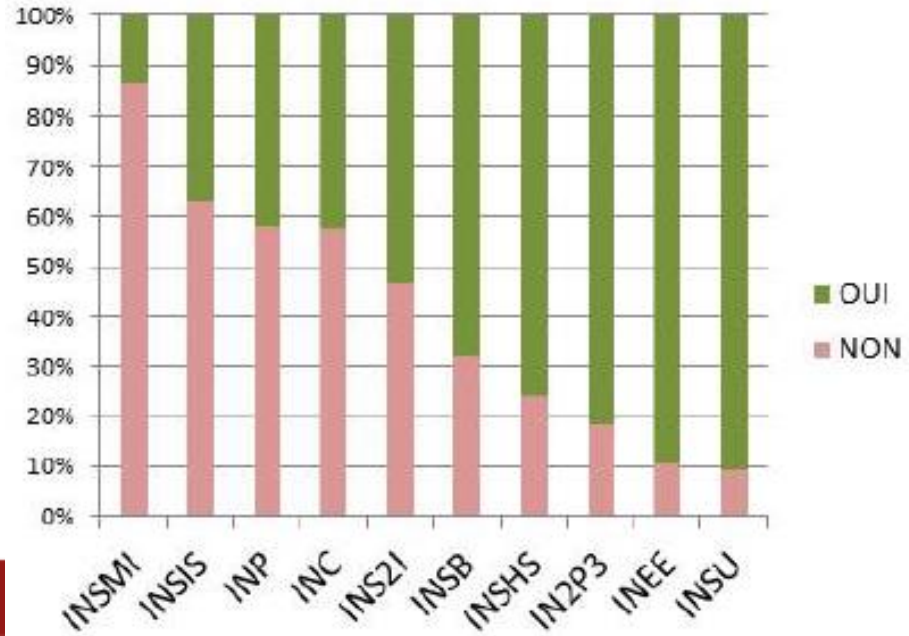
Cette étape nécessite une *phase de sélection des informations pertinentes (validées, utiles…)*, tout en se préoccupant de leur exploitation future à travers les *problématiques de durée de vie, de confidentialité et de sécurité des données*.

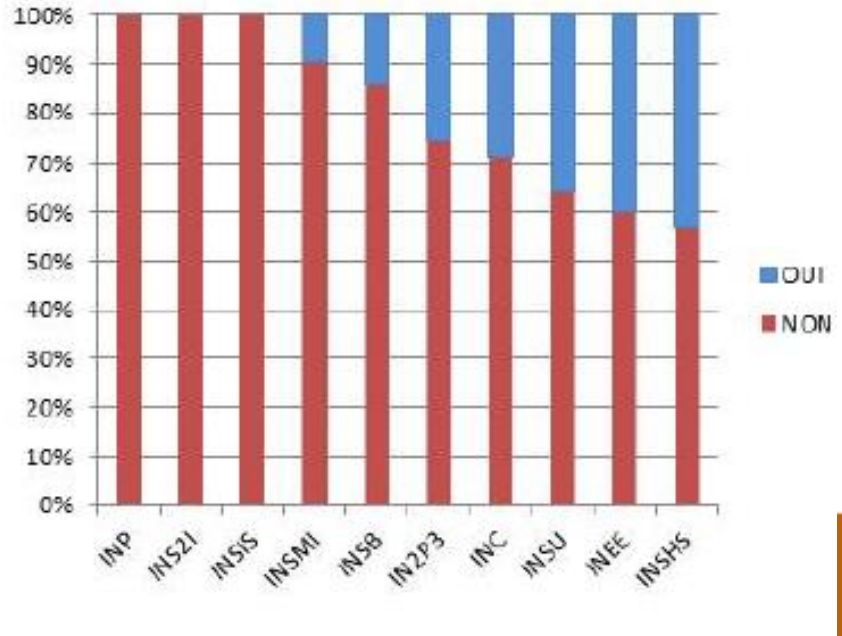
Enquête DIST (mars 2015)

"PAP 3 47- Les recherches conduites dans votre laboratoire produisent-elles des données de la recherche nécessitant des pratiques de gestion"



Enquête DIST: avez-vous une activité relative aux données de la recherche ?
OUI:33% NON : 67%





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

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

Site Web : **https://gt-atelier-donnees.miti.cnrs.fr/**

Contact : **donnees-inter-reseaux@services.cnrs.fr**

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, d'autres entités 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 ».

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 : réseau RELIER
- Pierre Navaro : réseau Calcul

Edition Web

- Pierre Navaro : réseau Calcul