INTRODUCTION AU PARTAGE DES DONNÉES DE LA RECHERCHE

SYLVIE ROUSSET

Directrice DDOR CNRS

Direction des Données Ouvertes de la Recherche

JSO 2022, 4ème édition

Au siège du CNRS

30 Novembre 2022

FAUT IL OUVRIR TOUTES LES DONNEES DE LA RECHERCHE?

Une donnée doit être ouverte ou protégée. L'ouverture des données s'entend selon l'expression **« ouvert autant que possible, fermé autant que nécessaire ».**

Toutes les données de la recherche n'ont pas vocation à être ouvertes ou divulguées. Il existe des exceptions évidentes telles que les données spécifiques à caractère confidentiel, que cela soit du fait de leur caractère personnel, pour des raisons de concurrence industrielle ou pour des intérêts fondamentaux ou réglementaires des États.

La décision d'ouverture ou de protection des données de la recherche doit être prise avec les services compétents du CNRS **:**

- les Services Partenariat Valorisation pour la propriété intellectuelle,
- la Délégation à la protection des données pour les données à caractère personnel et de
- la Direction de la sûreté pour les questions relatives à la souveraineté.

Les actions proposées ici traitent des données ayant vocation à être ouvertes, mais pas que !

DE QUOI PARLE T ON?

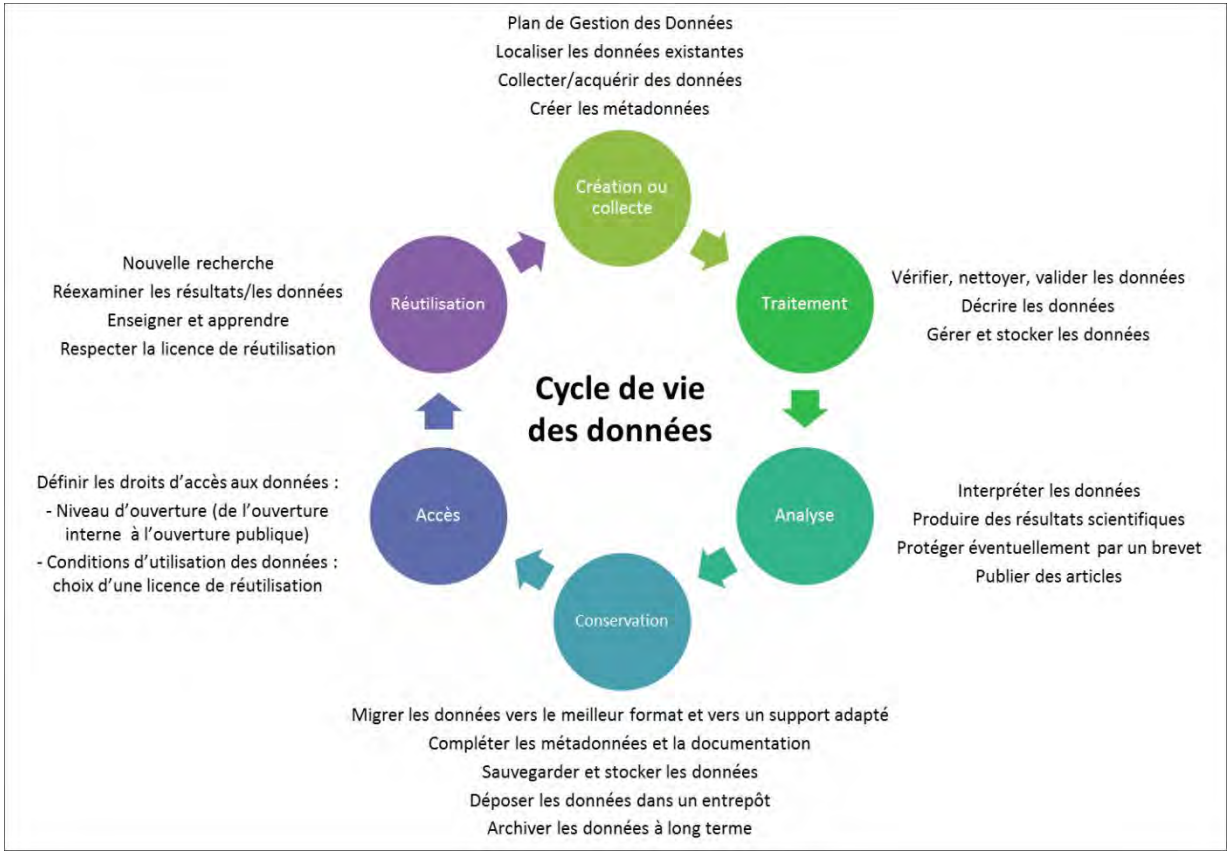
Définition des données de la recherche :
“information, en particulier des faits et des nombres, collectés afin d’être utilisés à des fins de raisonnement, discussion ou calculs.”

Trois types de données de la recherche :
- Les données brutes
- Les données “curées”
- Les données reliées aux publications scientifiques

Tous les formats des données de la recherche sont concernés :
y compris les textes, les documents, les logiciels, les algorithmes et les protocoles.

Ecosystème des données de la recherche :
Infrastructures numériques, Big Data, services et principes FAIR, Science Ouverte.

LE CYCLE DE VIE DES DONNÉES DE LA RECHERCHE



MOTIVATIONS

Rendre la recherche plus efficace et non redondante (pas de duplication inutile)

Assurer l'intégrité scientifique (reproductibilité et validation des résultats)

Etre en capacité de réutiliser les données même sans en être à l'origine

Croiser les données pour favoriser de nouvelles analyses, voire faire émerger de nouvelles thématiques

Satisfaire le cadre légal d'ouverture des données a priori : « Ouvert autant que possible, fermé autant que nécessaire » (loi pour une république numérique 2016)

Mutualiser et rationaliser les infrastructures informatiques, les moyens RH et identifier les nouveaux métiers (datastewarship …)

QUELQUES DÉFINITIONS

Entrepôt des données de la recherche

Service en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion des jeux de données de la recherche.

Stockage des données de la recherche

Déposer les données sur un support numérique pour les rendre accessibles. Cela peut être un ordinateur personnel, un disque partagé ou tout autre organe de dépôt. Le stockage permet d'assurer la continuité de l'exploitation sur du court terme.

Datacentre (datacenter)

Infrastructures matérielles (béton, alimentation électrique, climatisation, connexions réseaux, gardiennage, etc.) qui hébergent le matériel informatique, les moyens de calcul et les infrastructures de stockage de données.

Mesocentre

Un mésocentre est un ensemble de ressources humaines, matérielles et logicielles qui fournit à la communauté scientifique d'une même région un environnement scientifique et technique propice au calcul et au traitement de données.

DES CAS D'USAGES POUR PROMOUVOIR UNE CULTURE DE LA DONNEE

- 5 Equipes engagées : *Données Nucléaires (IN2P3), MESAdata (INP), Infranalytics (INC), Emergen (INSB), HiFiles4ML (INSIS))*
- Action initiée au dernier trimestre 2021 – 1ères réunions de travail début novembre 2021 – environ 4 réunions pour chaque cas d'usages

Principaux questionnements abordés dans les groupes

- Questions politiques *(incitation, obligation dépôt)*
- Nature des données à partager *(donnée évaluée, expérimentale, résultat de mesure, donnée associée à la publication ? Quelle est la donnée de référence ?)*
- Intérêt du partage *(des données brutes, données non attachées aux publications …)* et les Freins au partage *(RGPD, licences, embargos)*
- Volumétrie & sélection des données *(besoin de faire de la place sur les disques durs)*
- Stockage et sauvegarde des données *(lieux, moyens)*
- Sites de dépôt, entrepôts de référence *(nature, caractéristiques))*
- Interopérabilité des données et des formats *(besoin d'homogénéisation, de modèles communs, de solutions logicielles)*
- Besoin et pratiques de la communauté élargie *(intérêt de regrouper, fédérer des initiatives, synergies et convergences possibles)*
- DMP et Gestion des données *(instaurer les bonnes pratiques et se poser les bonnes questions)*

Bilan intermédiaire des 5 USE CASE – Juin 2022

Principaux travaux effectués et livrables

GT	Travail effectués et Livrables
Données Nucléaires	Rédaction du DMP (modèle IN2P3) pour stockage de leurs données au CC IN2P3
MESAdata	Définition d'un standard de métadonnées. Réseau METSA (>> centre de référence) Besoin d'une solution de stockage Dépôt Test DataVerse
Infranalytics	Solutions techniques et informatiques identifiées
Emergen	Rédaction de DMP et focus sur les questions juridiques
HiFiLES4ML	Expression des besoins et rédaction d'un cahier des charges pour le dépôt (et traitement) des données

Actions en cours et perspectives

Actions émergentes	Données Nucléaires	MESAdata	Infranalytics	Emergen	HiFiLES4ML
Identification de données à partager à préserver sur le long terme	X	X		X	X
Centralisation du stockage et/ou de la sauvegarde des données brutes dans un centre de calcul	X	X			X
Identification des plateformes collaboratives/ entrepôts thématiques nationaux et internationaux		X			X
Identification des référentiels à adopter : standards, ontologies, vocabulaires … (normalisation, structuration, interopérabilité)		X	X		
Gestion des données et rédaction de DMP (structure ou projet)	X	X		X	X
Définition d'une architecture logicielle pour coupler plusieurs bases de données et permettre leur interrogation			X		
Définition d'un schéma de métadonnées à partir de 3 jeux de données représentatifs des données produites au sein du réseau METSA		X			
Discussions en vue de lever certaines difficultés juridiques				X	

REALISATION D'UN ANNUAIRE OPIDOR CNRS

Identifier les dépôts de données et les services dont le CNRS est responsable ou auquel il participe

- Groupe de Travail créé en 2020 (pilotage Françoise Genova et Paolo Lai avec de nombreux représentants des instituts: *Mettre en œuvre l'Action 5 de l'Axe 2 de la Feuille de Route Science Ouverte du CNRS*
- Point de départ: Cat OPIDoR- le wiki des services dédiés aux données de la recherche, maintenu par l'INIST
- Les actions liées à la constitution de l'Annuaire
  - *Vérifier et compléter les informations de Cat OPIDoR*
  - *Identifier celles qui doivent apparaître dans l'Annuaire*
  - *Modifier éventuellement le modèle de données de Cat OPIDoR si des améliorations sont identifiées et travailler sur la visualisation de l'Annuaire*
- Les Instituts ont la responsabilité de l'identification et de la validation des entrepôts et services à faire figurer dans l'Annuaire pour obtenir une cartographie de qualité
- Le point de vue disciplinaire complété par la vision « top-down » des Infrastructures de Recherche de la Feuille de Route Nationale (implication du Comité TGIR)
- Travail de finalisation en cours avant insertion dans le Wiki CATOPIDOR général

LIEN VERS LA PAGE ANNUAIRE

https://cat.opidor.fr/index.php/CNRS_Donn%C3%A9es:Annuaire_des_entrep%C3%B4ts_et_services_de_donn%C3%A9es

OUTILS ET **SERVICES PROPOSES PAR L'INIST**

Accompagner la réalisation des **Plan de Gestion des Données** (PGD- OPIDoR)

Proposer **des Formations en ligne** via l'outil en ligne DORANUM

Fournir des **Identifiants pérennes** pour les jeux de données (DOI – Datacite – INIST)

Rédaction de plan de gestion de données

DMP OPIDoR

48 modèles dont **5** CNRS

URL https://dmp.opidor.fr

Depuis 2016

7 366 plans,
9 781 utilisateurs actifs,
50 établissements administrateurs

Outil open source développé par une communauté internationale

NEW

maDMP = **M**achine **A**ctionable DMP

- **réutilisation automatique d'informations** (projets ANR)
- **référentiels** (standards de métadonnées, entrepôts de données, terminologies…)
- **coûts associés à la gestion des données**

Modèle financeurs *ANR, Commission Européene, INCa*

Modèles institutionnels *: CEA, CIRAD, INRAE, UNISTRA, Institut Pasteur, Inserm …*

Modèles CNRS :

CC-IN2P3,
Projet PRESOFT – CC IN2P3
MASA,
PRODIG,
PACEA

cnrs Inist

DMP OPIDOR

DMP OPIDoR | DMP publics | Modèles de DMP | Aide | Plus | Français | Se connecter

Modèles de DMP

Modèles de DMP proposés par les financeurs ou par les organismes de recherche, disponibles dans DMP OPIDoR. Vous pouvez télécharger ces modèles et les recommandations associées, créer un plan à partir de ces modèles.

Rechercher

Nom du modèle	Nom de l'organisme	Type d'organisme	Description	Dernière mise à jour	Télécharger	Créer un plan
ANR - DMP template (english)	Agence nationale de la recherche (ANR)	Financeur	In line with its Open Science policy and the National Plan for Open Science, the French National Research Agency (ANR) requires all projects funded in 2019 onwards to produce a Data Management Plan (DMP). This move is intended to support European and international alignment efforts on the structure of open research data, and is guided by the principle: "as open as possible, as closed as necessary". In the interest of consistency, the ANR follows the recommendations of the Committee for Open Science (CoSO), which it has consulted on this matter. It has adopted the Science Europe DMP template, which aims to promote the international alignment of research data	16/05/2022		Connexion requise

https://dmp.opidor.fr/public_templates

DMP OPIDOR

DMP OPIDoR | DMP publics | Modèles de DMP | Aide | Plus | Français | Se connecter

CC-IN2P3 - DMP template (english)	CC-IN2P3 (Centre de Calcul - Institut national de physique nucléaire et de physique des particules du CNRS)	Institution	Model of data management plan proposed by the IN2P3/CNRS Computing Center. This model is generic and seeks to reinforce the research data life cycle management planning process. It does not refer to specific technologies or services. The main objective is to encourage a global reflection on the different aspects of research data management by means of a set of questions which are organized in 4 sections: • General description of the project • Dataset management • Legal and ethical framework • Long term storage and preservation.	16/05/2022		Connexion requise
Plan de gestion des données - Modèle CEA (FR)	CEA Commissariat à l'énergie atomique et aux énergies alternatives	Institution	*Les plans de gestion de données (PGD ou DMP) réalisés à l'aide de l'outil OPIDoR sont stockés sur des* ***serveurs extérieurs au CEA****. L'outil peut ne pas être adapté aux PGD de projets comportant des données à caractère confidentiel ou de certains projets internes.* **Choix du modèle de plan de gestion des données à faire en fonction du projet de recherche** • Si projet ANR, vous pouvez utiliser le modèle ANR	31/08/2022		Connexion requise

DORANUM

https://doranum.fr/

Formations Gestion de données

9 thématiques autour du cycle de vie des données de la recherche

+ de 100 ressources pédagogiques multimodales

Fiches synthétiques, Vidéos, Tutoriels,
Ressources interactives, Mini jeux, Quiz, Etc…

Ressources pédagogiques par disciplines scientifiques

Evolution

Parcours doctorants co-réalisé avec la DDE - UL

Gestion des données de recherche en Environnement

Parcours pédagogique 3 h PDF

DORANUM
THÉMATIQUES ABORDÉES

ENJEUX & BÉNÉFICES
Pourquoi partager les données ?
Qu'est-ce que l'Open Science ?

ASPECTS JURIDIQUES, ÉTHIQUES, INTÉGRITÉ SCIENTIFIQUE
Que puis-je partager, réutiliser ?
Quelles pratiques devrais-je respecter ?

PLAN DE GESTION DE DONNÉES
Pourquoi et comment rédiger un plan de gestion des données ?

MÉTADONNÉES
Comment décrire les données ?

IDENTIFIANTS PÉRENNES
Comment associer durablement des données à son auteur ?

DÉPÔT & ENTREPÔTS
Comment et où déposer mes données ?

STOCKAGE & ARCHIVAGE
Quelles données conserver à long terme et comment ?

DATA PAPERS & DATA JOURNALS
Comment publier mes données comme un article scientifique ?

ACCÈS & VISUALISATION
Où et comment extraire et visualiser les données qui m'intéressent ?

Attribution d'identifiants pérennes

Inist-CNRS = agence d'attribution DOI DataCite en France.

Depuis 2009

720 000 DOI attribués

263 735 pour des structures CNRS

Près de 175 utilisateurs

75 structures tutelle CNRS : laboratoires, centres de données, infrastructures de recherche

Entrepôt national Recherche Data Gouv

Animation du consortium Datacite France

1ère assemblée générale le 1er décembre 2022

SYNTHESE DES ACTIONS POUR LE PARTAGE DES DONNEES

Accompagnement de la Création de la Plateforme MESR Recherche.data.gouv
Construction de la plateforme inaugurée le 8 juillet 2022
(INIST, comité de pilotage RDG)
Ateliers de la donnée (soutien du CNRS comme partenaire)
Centres de ressources (OPIDOR, DORANUM, INIST et URFIST)
Centres de références thématiques (Humanum, progedo, CDS, Data terra, IFB)

Actions propres au CNRS (en cours)
Constitution d'un annuaire des services et entrepôts CNRS
Etude de Use-cases avec les scientifiques

QUELQUES GRANDS CHANTIERS EN COURS

Créer un espace institutionnel CNRS sur la plateforme RechercheDataGouv

Accompagner les scientifiques au dépôt de leur jeu de données

- INIST? Ateliers de la donnée en region? Administrateurs curateurs de l'espace institutionnel?

Créer des référentiels thématiques de métadonnées pour toutes les communautés scientifiques qui n'en ont pas encore

Comment l'organiser? Par labo? Par projets sciencitfiques? Par équipe de recherche? Par Infrastructure de recherche? Par communauté scientifique (quel grain?)

Disposer d'espace de stockage pour toutes les données de la recherche que l'on souhaite partager (par exemple les données liées aux publications mais plus largement) ? auprès d'un centre de calcul national ? auprès d'un data centre régional ? dans une infrastructure de recherche ? dans son labo ?

Quel modèle économique sous-tend ces questions ?

Comment répondre aux questions juridiques ?

MERCI DE VOTRE ATTENTION