# INTRODUCTION SUR LES ACTIVITÉS DU CNRS SUR LES DONNÉES

SYLVIE ROUSSET

Directrice Données Ouvertes de la Recherche

DDOR-DGDS, CNRS

RDA France - Plénière

12 octobre 2023

# CONTEXTE : SCIENCE OUVERTE

https://www.science-ouverte.cnrs.fr/

**Continuum calcul intensif / traitement et curation des données / documentation et référencement des données / mise à disposition des données / publications**

**Mais deux structures pour couvrir ces sujets au CNRS**

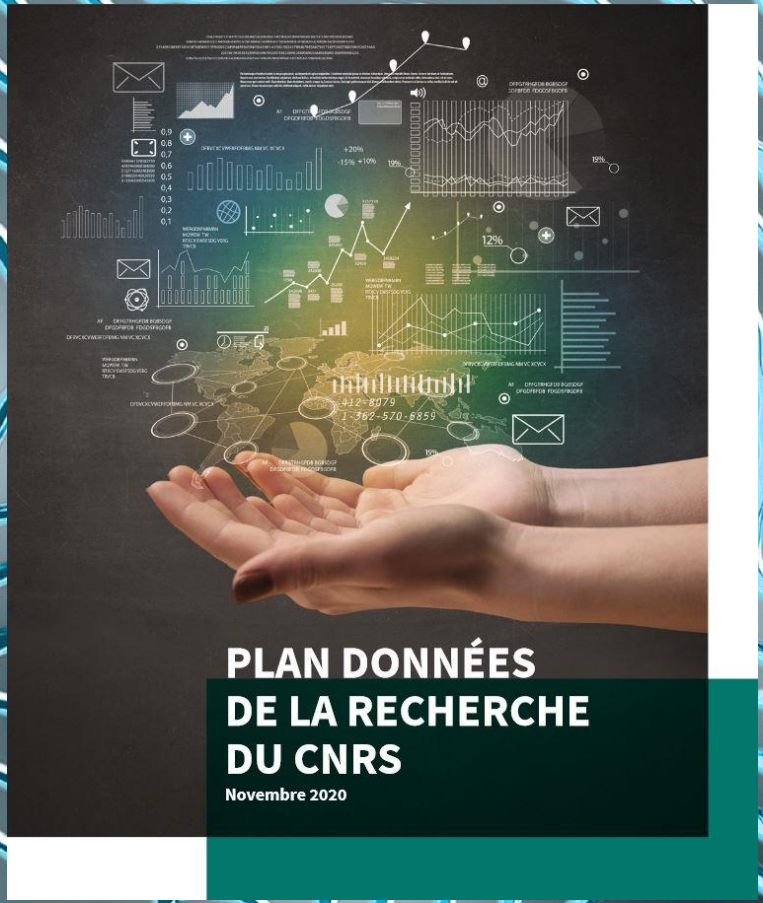
- Direction de l'information scientifique et technique (DIST)
- Mission calcul – données, créée en 2015 *(principalement axée « calcul intensif »)*

**Création de la Direction des données ouvertes de la recherche (DDOR)**

- Fusion de la DIST et de Micado au 1er novembre 2020
- Direction fonctionnelle *(pérennise le rôle de la mission calcul-données)*

# LE PLAN DONNEES DE LA RECHERCHE DU CNRS

Lancé en novembre 2020



# LE PARTAGE DES 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
- la Direction de la sûreté pour les questions relatives à la souveraineté.

# 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 (data stewardship …)**

# PLAN D'ACTIONS : VALORISER LES DONNEES

Promouvoir une « culture de la donnée », ainsi que les infrastructures, les outils et services, en incluant la question des moyens (humains et financiers) et de la formation

Travailler avec les instituts en prenant en compte les spécificités disciplinaires et l'existant, et en s'appuyant sur une gouvernance transverse des données

Soutenir les communautés scientifiques pour définir les éléments spécifiques de la gestion de leurs données à FAIRiser

Encourager le dépôt des données dans des entrepôts thématiques en priorité

Valoriser les données en lien avec les publications, et les « data papers »

# 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)**

# Quelques actions récentes

# QUELQUES EXEMPLES DE RÉALISATION EN MATIÈRE DE DONNÉES

**ENQUÊTE DDOR SUR LES PRATIQUES DE SCIENCE OUVERTE**
*Connaître les pratiques de stockage et d'archivage des données*

**CAS D'USAGE**
*Ouvrir la réflexion sur le partage des données, accompagner, analyser les difficultés*

**ANNUAIRE DES ENTREPÔTS ET SERVICES CNRS**
*Identifier les services et infrastructures dont le CNRS est responsable ou auxquels il participe*

**DATACENTRES, MÉSOCENTRES & SCIENCE OUVERTE**
*Ressources, technologies et moyens existants*

**RECHERCHE DATA.GOUV**
*Comité de pilotage ateliers de la donnée, centres de référence thématique, centres de ressources*

# INSTRUCTION DE CAS D'USAGE FAIRISATION DES DONNÉES

**Instruire tous les aspects de la mise à disposition de données :**

- Identification des données *(brutes, traitées, …)*
- Plan de gestion des données
- Métadonnées
- Identification des entrepôts possibles
- Aspects juridiques
- Proposer une solution
- …

**Une dizaine de cas sélectionnés :**

- Représentatif de différentes thématiques
- Maturité et besoins très différents
  - *Communautés très structurées, y compris au niveau international*
  - *Communautés qui commencent à envisager la question*

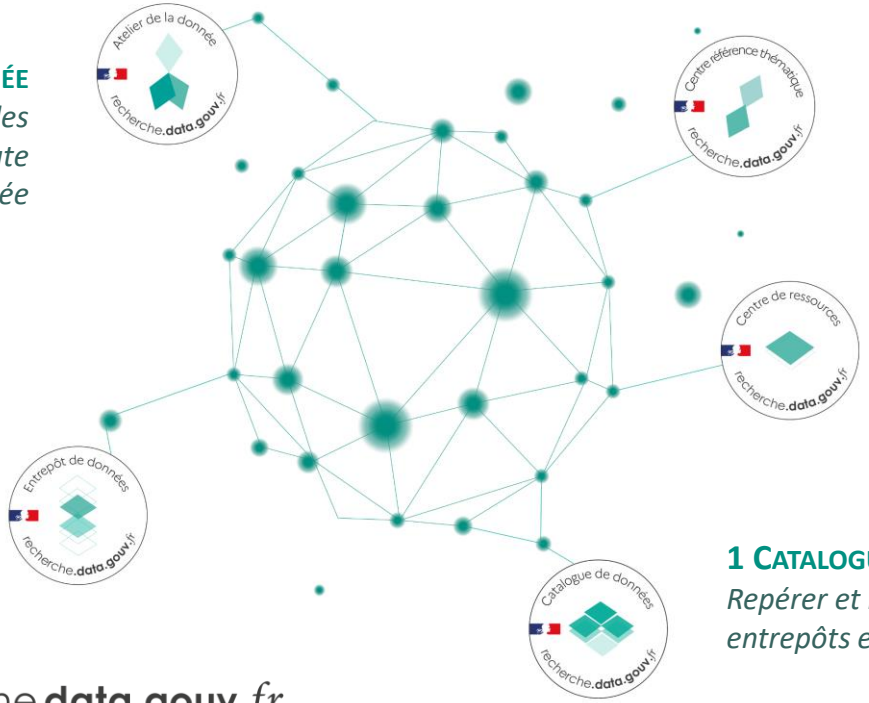
**Instruction tri-partite** :

- Equipe scientifique productrice des données
- Equipe INIST
- Suivi par équipe DDOR

# PLATEFORME NATIONALE POUR LE PARTAGE DES DONNEES

## Un écosystème au service du partage et de l'ouverture des données de recherche

**19 ATELIERS DE LA DONNÉE**
*Expertise généraliste en proximité des équipes de recherche pour toute question relative à la donnée*




**6 CENTRES DE RÉFÉRENCE THÉMATIQUES**
*Expertise par domaine scientifique*

**4 CENTRES DE RESSOURCES**
*Pour soutenir les ateliers et capitaliser leurs pratiques*

**1 ENTREPÔT DE DONNÉES**
*Offre mutualisée pour tous les établissements pour le dépôt et la publication des données*

**1 CATALOGUE DES DONNÉES**
*Repérer et moissonner les données des entrepôts externes de confiance*

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

recherche.data.gouv.fr

cnrs

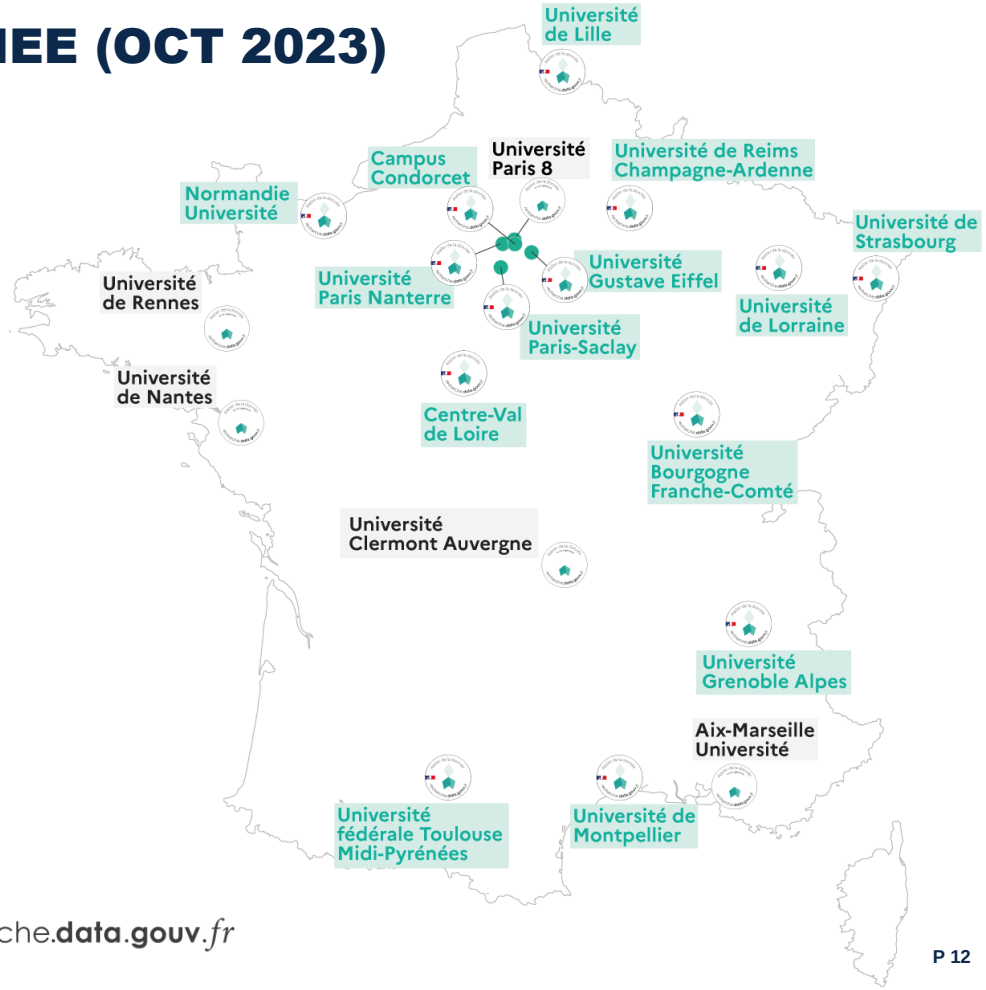
# 19 ATELIERS DE LA DONNEE (OCT 2023)

## Point d'entrée généraliste en proximité des équipes de recherche pour toute question relative à la donnée

Mutualiser les services et les compétences des établissements à l'échelle d'un territoire

*Développement progressif des ateliers de la donnée*

- *Au rythme de leur conception par les établissements*
- *Au fil des appels à manifestation d'intérêt successifs (3 par an d'ici fin 2023)*
- *14 ateliers labellisés*
- *5 ateliers en trajectoire de labellisation*




RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

recherche.data.gouv.fr

# 6 CENTRES DE RÉFÉRENCES THÉMATIQUES

***Expertise par domaine scientifique***

Définition et diffusion des **bonnes pratiques et des standards internationaux de gestion, traitement et diffusion des données**
par domaine scientifique

**Premier paysage proposé :**
**infrastructures de recherche ayant une activité structurante de gestion et diffusion de données pour leur communauté scientifique**

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

DATA TERRA

EDS

Huma-Num IR*

PRO GE DO
DATA INFRASTRUCTURE

PNDB Pôle National de Données de Biodiversité

Centre référence thématique recherche.data.gouv.fr

RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

recherche.data.gouv.fr

cnrs

# QUELQUES PROJETS EN MATIÈRE DE DONNÉES


Recherche.data.gouv

PGD CNRS


Compétences et Formation (projet EOSC)


Réflexion autour d'une solution pour les données non couvertes par Recherche.data.grouv

# NOUVEAUX OUTILS GENERIQUES AU CNRS

- ✓ **Catalogue Entrepôts et Services Données CNRS**
- ✓ https://cat.opidor.fr/index.php/CNRS_Donn%C3%A9es_de_la_Recherche_:_Catalogue_des_entrep%C3%B4ts_et_des_services

- ✓ **« CNRS Research Data »** : un Entrepôt CNRS sur la plateforme RechercheDataGouv pour le stockage de données génériques qui ne peuvent pas être stockées dans des entrepôts thématiques https://entrepot.recherche.data.gouv.fr/dataverse/cnrs

- ✓ **« DMP – CNRS »** modèle de DMP (PGD) en préparation sur OPIDOR

- ✓ Présence forte dans **l'écosystème RechercheDataGouv**

- ✓ Travail en cours sur la question du **stockage des données**

# L'ESPACE CNRS SUR LA PLATEFORME NATIONALE : CNRS RESEARCH DATA

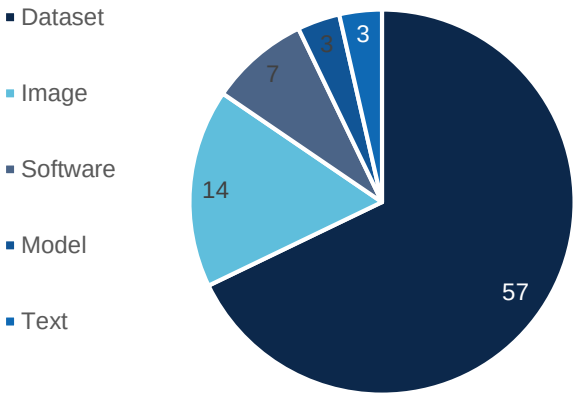
# Espace institutionnel CNRS Research Data

- **Proposer un espace aux scientifiques pour déposer leurs données et les rendre accessibles lorsqu'il n'existe pas d'entrepôts thématique**
- **Contribuer à l'écosystème Recherche Data Gouv en créant un espace institutionnel CNRS.**
  - Le CNRS est déjà présent via les Centres de Ressources (DORANum, OPIDoR) et les Centres de Références Thématiques (CDS, HUMA-NUM, IFB, etc.)
- **Mutualiser les efforts en évitant le déploiement et la gestion d'un entrepôt des données propre au CNRS.**

# Etat des lieux au 30/08

- **Ouverture de CNRS Research Data le 29 juin 2023**
- **61 jeux de données publiés**
  - Chimie, physique, sciences de l'ingénierie et des systèmes, sciences de l'Univers, sciences de la vie, etc.
  - 6024 fichiers déposés
- **3 jeux de données en cours de curation**
- **2 collections créées :**
  - ICMCB (Bordeaux)
  - OSUG (Grenoble) – 4 sous-collections
- **2 collections en cours de mise en place**
  - Institut Néel (Grenoble)
  - LULI (Palaiseau)

**Types de données**



# UN MODELE DE PGD/DMP

# CNRS

# Un PGD CNRS : Pourquoi? Comment?

- Volonté de proposer un modèle de PGD CNRS
  - Dans le cadre général de la mise à disposition de l'espace CNRS Research Data et de la plateforme Recherche Data Gouv
  - Répondre au besoin d'accompagnement des équipes scientifiques
  - Afficher la politique du CNRS en matière de données de recherche
- Un travail fait en collaboration étroite avec l'INIST
  - Choix de départ du modèle structuré Science Europe
  - Ajout et adaptation de recommandations CNRS
  - + Rédaction d'étiquettes RGPD par le Service Protection des Données (SPD)
- Une première version disponible pour examen et test
  - Version en ligne : https://opidor-preprod.inist.fr/ .

# Plan du PGD : les questions à se poser

## 1. Description des données et collecte ou réutilisation de données existantes

- 1.1 Description générale du produit de recherche
- 1.2 Est-ce que des données existantes seront réutilisées ?
- 1.3 Comment seront produites/collectées les nouvelles données ?

## 2. Documentation et qualité des données

- 2.1 Quelles métadonnées et quelle documentation (par exemple mode d'organisation des données) accompagneront les données ?
- 2.2 Quelles seront les méthodes utilisées pour assurer la qualité scientifique des données ?

## 3. Exigences légales et éthiques, code de conduite

- 3.1 Quelles seront les mesures appliquées pour assurer la protection des données à caractère personnel ?
- 3.2 Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?
- 3.3 Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

## 4. Traitement et analyse des données

- 4.1 Comment et avec quels moyens seront traitées les données ?

## 5. Stockage et sauvegarde des données pendant le processus de recherche

- 5.1 Comment les données seront-elles stockées et sauvegardées tout au long du projet ?

## 6. Partage des données et conservation à long terme

- 6.1 Comment les données seront-elles partagées ?
- 6.2 Comment et quelles données seront-elles conservées à long terme ?


LA CONTINUITÉ
CALCUL – DONNEES
AU SEIN DE LA DDOR

# Stockage des données à partager

- **Pas d'offre générique CNRS stockage / mise à disposition des données**
  - Quelques solutions « ad hoc » pour certaines communautés (Climeri, …)
- **Instruction d'une offre pour répondre aux besoins non couverts par RDG**
  - Volumétrie
  - Capacité de traitement associée
  - Etendre les missions de centres existants
- **Rationalisation des infrastructures numériques**
  - Datacentres labellisés
  - Réduction des coûts et de l'empreinte environnementale
  - **Ne pas développer sa propre solution !!!**
- **Coûts et financements de la science ouverte ?**
  - Modèle économique ?

# Datacentres et Mésocentres

## CNRS opérateur de deux des quatre datacentres d'envergure nationale


- **IDRIS** (Orsay) – calcul intensif
  - Opère le calculateur Jean-Zay financé par GENCI
  - + partie dédiée aux recherches en intelligence artificielle
  - Projet CLUSSTER
  - Hébergement (mésocentre Paris-Saclay, données CLIMERI, IFB, …)

- **CC-IN2P3** (Lyon)
  - Traitement de données massives pour les activités IN2P3
  - Hébergement : DSI CNRS, HAL, HumaNum, BBEES, …

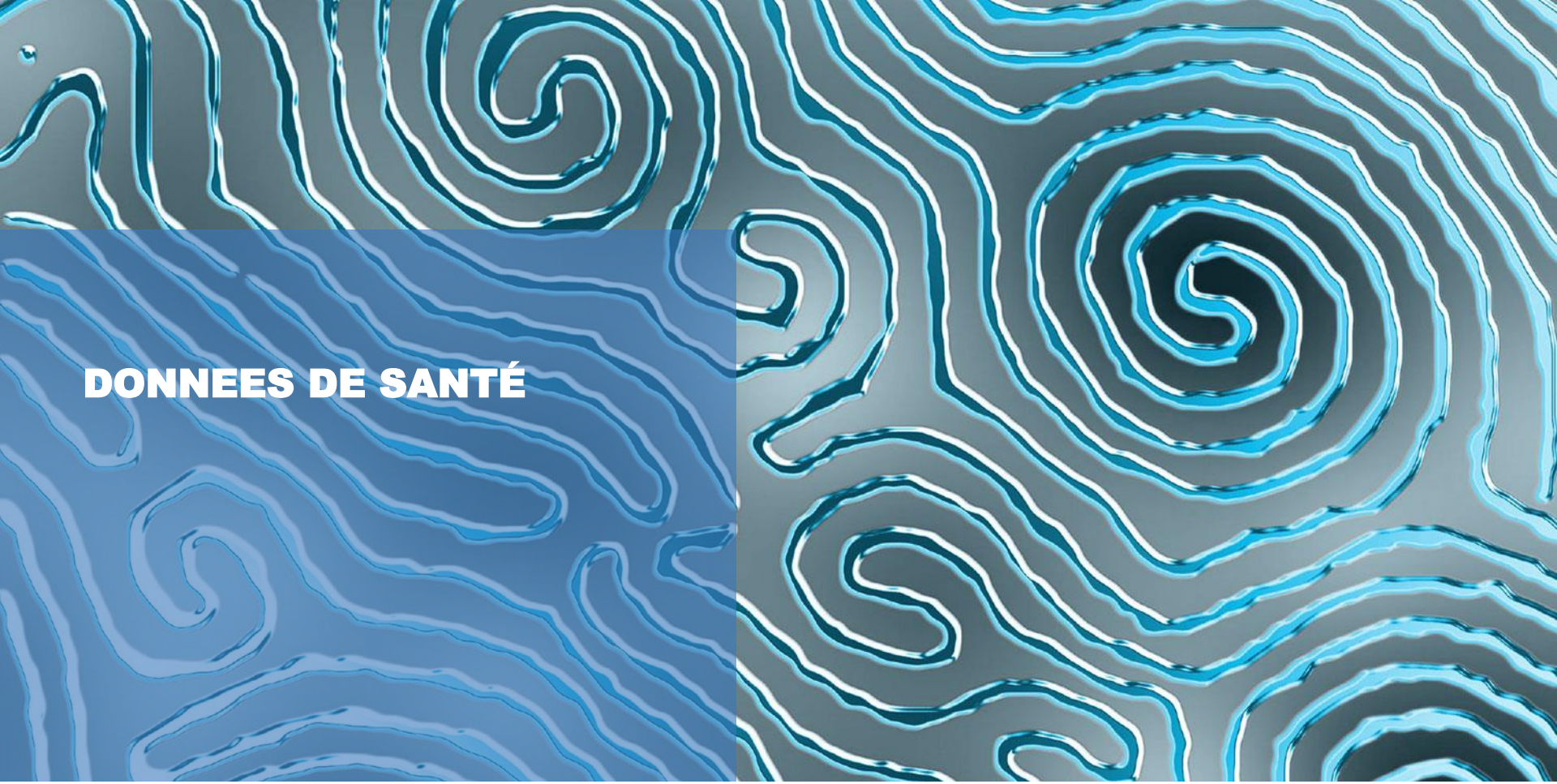
**Projet Equipex+ FITS** :

- Offre de services calcul / données pour les TGIR
- Basée sur CC-IN2P3 et IDRIS + partenariat GENCI
- Plusieurs « use cases » : Soleil, HL-LHC, LSST, IFB, France-Grille

**Deux mésocentres sont rattachés au CNRS (UAR)**

- CALMIP (Toulouse)
- GRICAD (Grenoble)

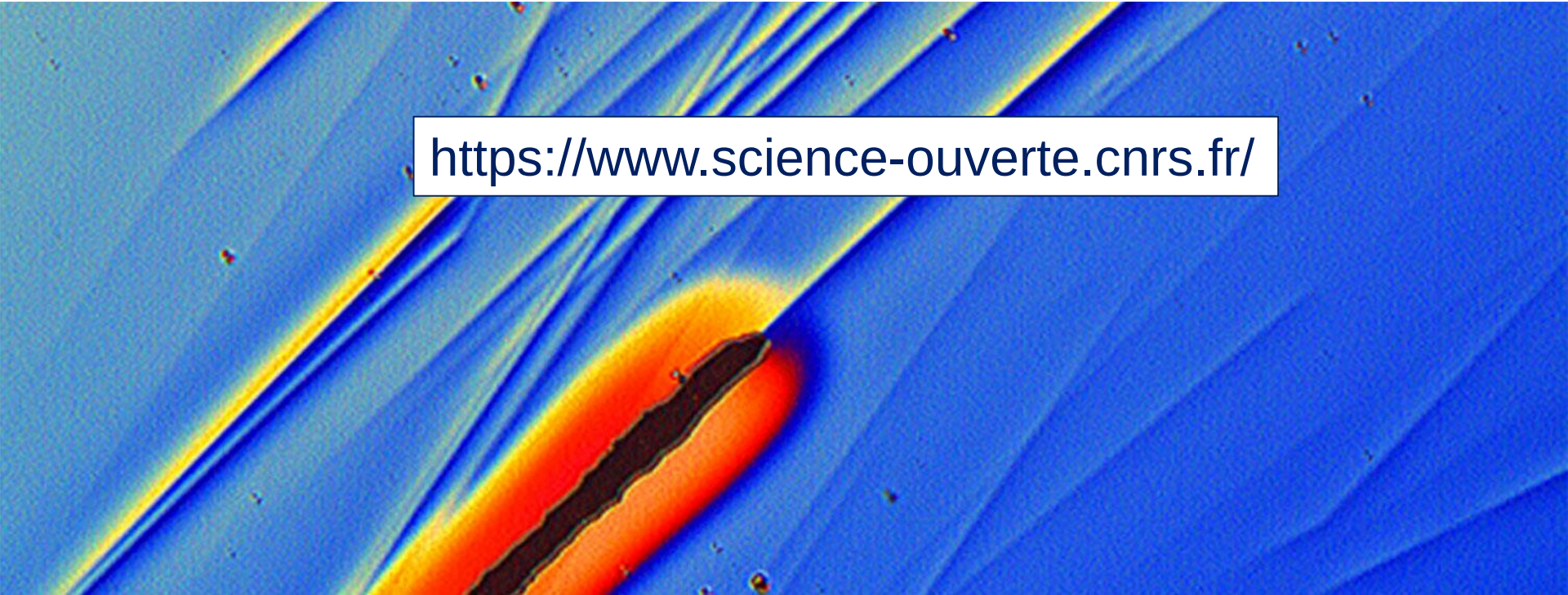
+ Demandes d'association d'autres mésocentres


DONNEES DE SANTÉ

# Données sensibles / données de santé

- **GENCI peut accueillir des projets utilisant des données sensibles**
- **Le CNRS bénéficie d'un accès permanent à la base SNDS**
  - Données de la CNAM (remboursements SS) + quelques bases supplémentaires
  - Procédures d'accès simplifiées
    - Instruction interne au CNRS en un mois (délégation à la protection des données)
    - Formation obligatoire
    - Pas de déclaration CNIL
- **Le CNRS est membre du Health Data Hub**
  - Offres ?
  - Porteur d'un projet de partage des données de santé au niveau européen
- **Habilitation à l'hébergement de données de santé**
  - Conditions très exigeantes
  - Pas de centres nationaux ni régionaux habilités
    - Quelques utilisations de données de santé via les CHU
  - Instruction du sujet si indispensable (un centre ?)

# MERCI DE VOTRE ATTENTION



https://www.science-ouverte.cnrs.fr/