Direction des données ouvertes de la recherche
CNRS - DDOR

Journées Calcul-Données 2022

Direction des données ouvertes de la recherche du CNRS

Premier bilan et perspectives

Denis VEYNANTE

1 CONTEXTE : Science ouverte

Enjeu scientifique fort pour la gestion et le partage des données

- Recherche plus efficace et non redondante *(pas de duplication inutile)*
- Intégrité scientifique *(reproductibilité et validation des résultats)*
- Capacité de réutiliser des données *(même sans en être à l'origine)*
- Croiser les données *(nouvelles analyses voire nouvelles thématiques)*

Satisfaire un cadre légal :

- Loi pour une république numérique (2016)
  - *Données « ouvertes autant que possible, fermées autant que nécessaire »*

Mutualiser et rationaliser

- Infrastructures informatiques
- Moyens humains
- Identifier les nouveaux métiers *(« data stewardship »)*

European open science cloud

- au niveau européen, AISBL dont le CNRS est membre (2020),
- au niveau français un « EOSC français » (DGRI)

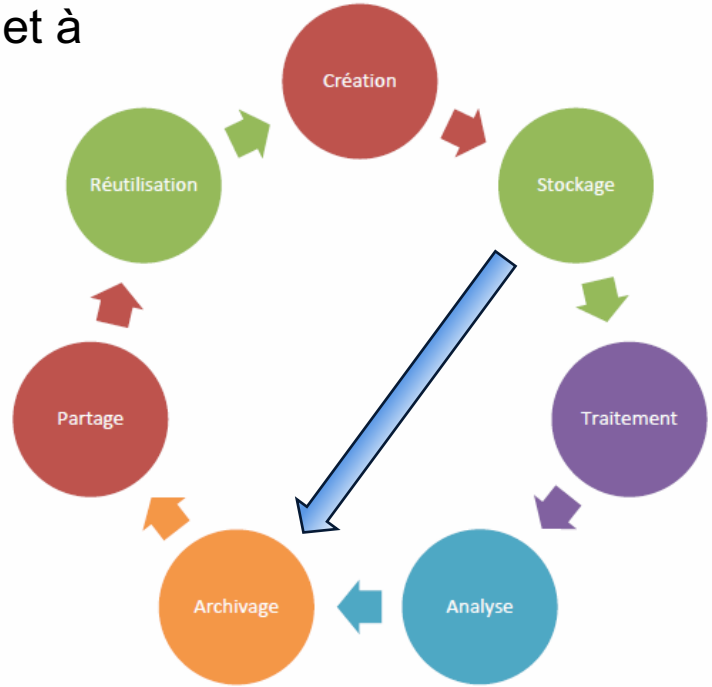
1 CONTEXTE : défis et ambitions

Findable Accessible Interoperable Reusable

Construire un écosystème propice au partage et à la réutilisation des données

Les données au CNRS : un « produit » complexe

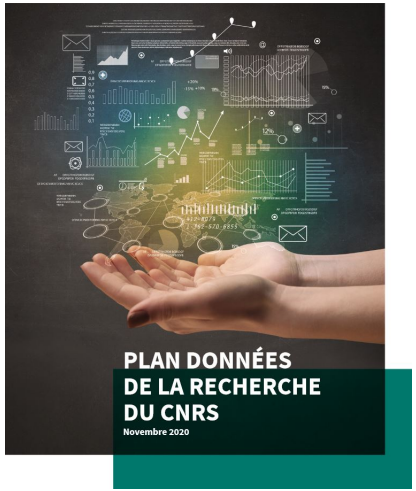
- Maturité différente selon les disciplines
- Pratiques diverses et variées
- Large champ d'investigation
- Aspects divers :
  - *Disciplinaires et organisationnels*
  - *Technologiques*
  - *Juridiques*



Le cycle de vie de la donnée

1 CONTEXTE : un plan d'action à déployer

- Favoriser l'émergence de bonnes pratiques
- Favoriser l'émergence de nouveaux outils
- Apporter des réponses en termes de ressources humaines
  - *Evolution et gestion des carrières*
  - *Nouveaux métiers*
- Développer la formation



Plan "Données de la recherche" du CNRS
Novembre 2020

Feuille de route science ouverte du CNRS
Novembre 2019

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

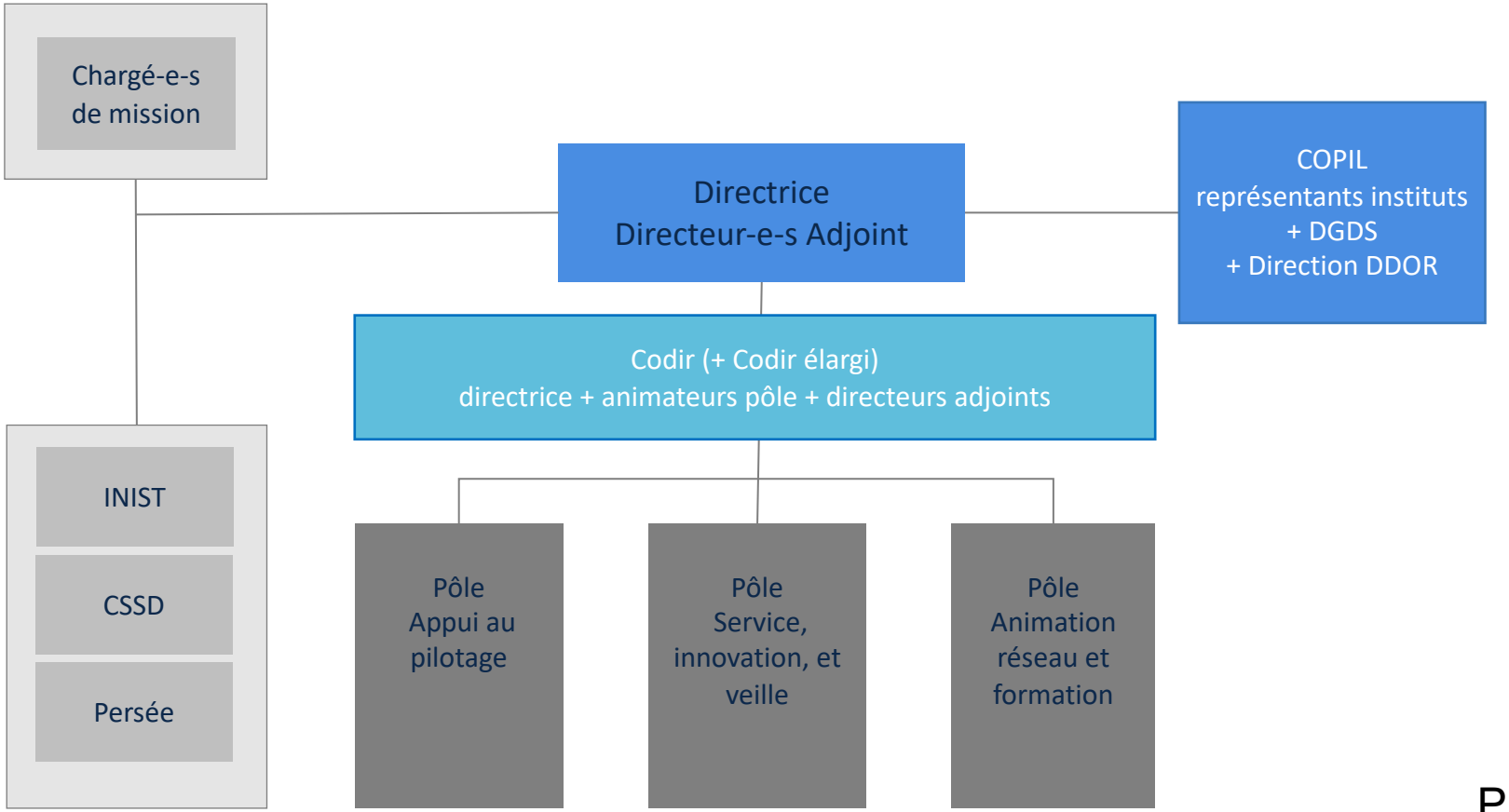
Mais deux structures pour couvrir ces sujets au CNRS

- 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)*

2 Une direction fonctionnelle des données ouvertes de la recherche : organisation de la DDOR



2 DDOR : différents comités

Comité de pilotage
- Directeur général délégué à la science
- Représentants des 10 instituts *(niveau direction adjointe scientifique)*
- Direction DDOR

Comités de direction
- « Restreint » : Directrice, directeurs adjoints, animateurs de pôles
- « intermédiaire » : « restreint » + directrices INIST, CCSD et Persée
- « élargi » : « intermédiaire » + directeurs CC-IN2P3 et IDRIS

Groupes de travail
- Données
- Publications
- Infrastructures numériques
  - Directeurs centres nationaux CC-IN2P3 et IDRIS
  - Directeur Maison de la Simulation
  - Directeurs CALMIP (Toulouse) et GRICAD (Grenoble)

Quelques actions

3 Feuille de route du CNRS pour la Science Ouverte

- 100% de publications en accès ouvert et ré-utilisables
- Des données de la recherche FAIR-isées
- Développer et promouvoir les outils pour l'analyse et la
- fouille des textes et des données
- L'évaluation individuelle des chercheurs

- Site WEB https://www.science-ouverte.cnrs.fr
  - Feuille de route Science Ouverte du CNRS
  - CNRS Roadmap Open Science

OSEC 2022
Open Science European Conference

4-5 février 2022, Paris

3 Une évaluation plus qualitative et moins quantitative

Quatre principes pour l'évaluation individuelle

1. Ce sont les résultats eux-mêmes qui doivent être évalués, et non pas le fait qu'ils aient pu être publiés dans une revue prestigieuse ou autre média réputé.

2. Expliquer la portée et l'impact des productions citées, ainsi que la contribution personnelle apportée. L'exhaustivité de la liste des productions est inutile.

3. Tous les types de production peuvent contribuer à l'évaluation : données sous-tendant la publication, code source, preprints, data papers…

4. Toutes les productions citées dans les dossiers d'évaluation doivent être accessibles dans HAL ou éventuellement dans une autre archive ouverte*.

*Trois exceptions à cette règle sont recevables :

1. Les résultats trop récents peuvent être sous embargo. Auquel cas ils doivent quand même avoir été déposés dans HAL, avec une durée d'embargo ne dépassant pas ceux prévus par la loi (6 mois en STM, et 12 mois pour les SHS). Ils sont alors fournis par un lien privé dans HAL (ou alors dans le dossier).
2. Pour les recrutements, cette règle ne peut pas être absolue pour les candidats exerçant à l'étranger dans des institutions étrangères ou internationales, ou des institutions privées.
3. Le type de production peut ne pas être accepté dans HAL.

3 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
Identifier les ressources, technologies et moyens existants

PARTICIPATION AUX PROJETS RECHERCHE DATA.GOUV
Comité de pilotage ateliers de la donnée, centres de référence thématique, centres de ressources

3 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

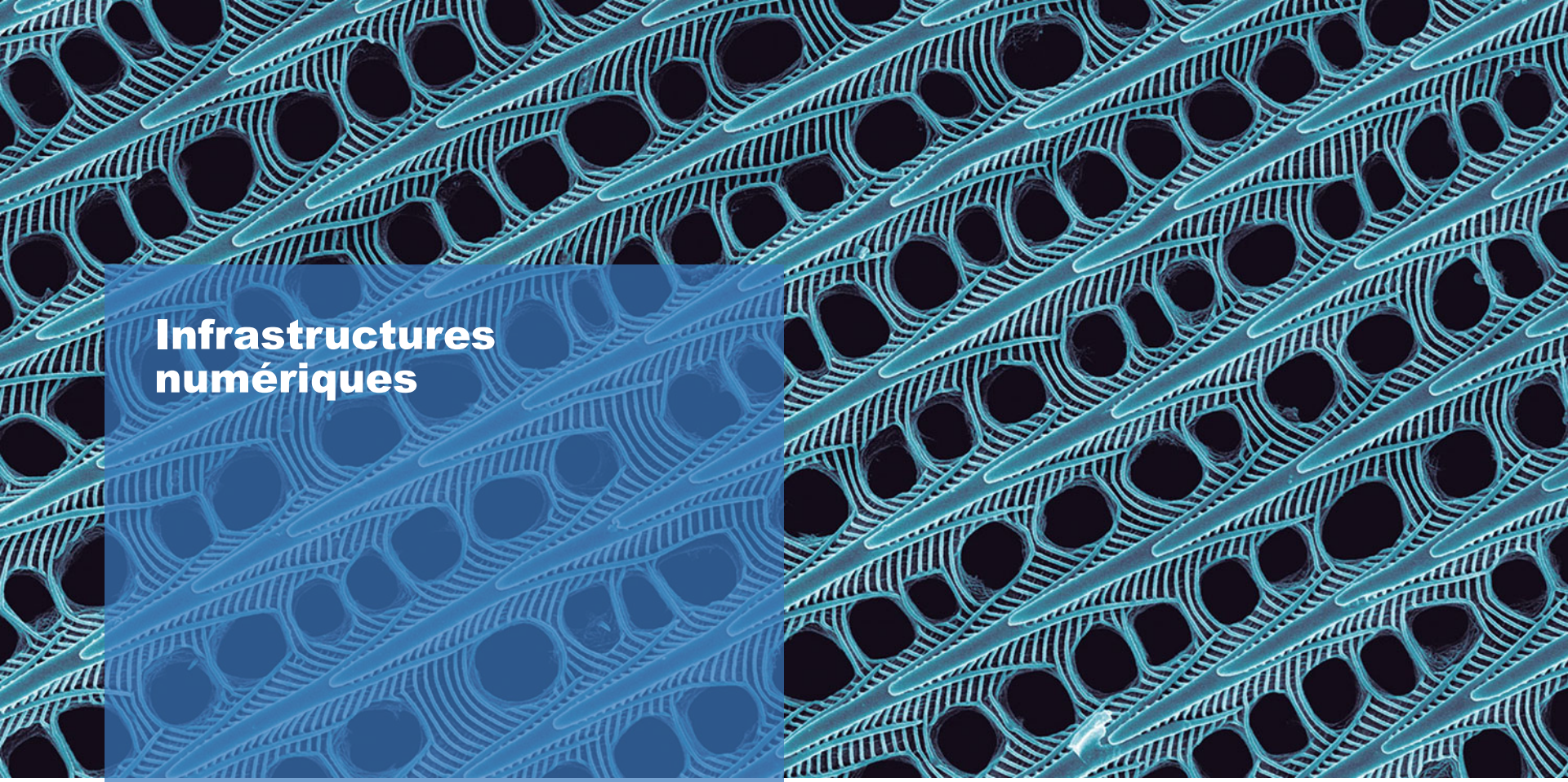
Quelques projets en matière de données

Recherche.data.gouv (suite)

Réflexion autour du PGD Structure

Compétences et Formation (projet EOSC)

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

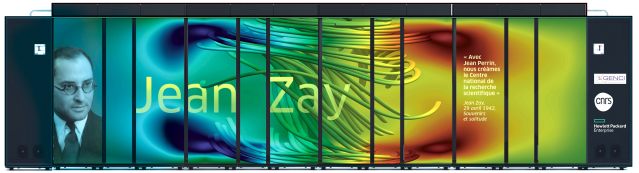


Infrastructures numériques

P

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)



- **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)

4

Définir une politique du CNRS vis-à-vis des demandes d'UAR

(qui arrivent en ordre dispersé, par des filières variées…)

Motivations des partenaires :

- Politique : tutelle nationale pour fédérer les acteurs locaux
- Reconnaissance
- Bénéficier d'un soutien CNRS en moyens financiers et surtout humains

Intérêt pour le CNRS :

- Infrastructures utilisées par nos laboratoires
- Contribuer à la rationalisation et la mutualisation des infrastructures informatiques
  - Réduire le nombre de salles machines et les moyens significations dans les laboratoires
  - Réduction des coûts afférents
- Réduire l'empreinte environnementale associée (salles à l'état de l'art)
- Optimiser les moyens humains affectées à la gestion opérationnelle des machines
  - Au profit d'un support de haut niveau aux utilisateurs, vraie valeur ajoutée

4 Question : faut-il labelliser ? Et si oui quoi ?

Mésocentres / datacentres / centres de données ?

- La distinction ne survivra pas forcément à terme :
  - Mésocentre et centre de données pourraient devenir
    - Des services du datacentre
    - Des centres ou réseaux de compétences qui n'opèrent pas les machines

- Quelles missions ?
  - Mésocentres aujourd'hui essentiellement orientés calcul intensif… mais évoluent !
  - Les datacentres couvriront toutes les infrastructures données de la recherche et IA, y compris l'hébergement physique des « entrepôts de données »

Science ouverte

- « Ateliers de la donnée » *(plateforme recherche.data.gouv)*
- Référencement et mise à disposition des données ?
  - Mésocentre = producteur et outil de traitement de données
  - Infrastructures de stokage
- Publications ?

Infrastructures numériques / compétences ?

4 Question : faut-il labelliser ? Et si oui quoi ?

Quels sites ? A quelles conditions ?

Etat des lieux en cours…

- Présence « significative » du CNRS
- Stratégie lisible et partagée
  - Rationalisation / mutualisation
  - Participation à la gouvernance
  - Engagements des partenaires en moyens, y compris RH
  - Modèle économique pérenne
- Concertation avec la DGRI
  - Labellisation datacentres régionaux basée sur une trajectoire
  - Projets plus ou moins avancés en pratique
- En fonction des moyens que le CNRS pourra y consacrer…

4 Inquiétude : énergie

Etat des lieux

- Faillite Hydroption *(fournisseur CNRS)*
  - Reprise du contrat par EDF à un tarif supérieur
- Actions de l'Etat :
  - Augmentation quota ARENH *(« accès régulé à l'électricité nucléaire historique »)*
  - Réduction de la CSPE pour 2022 de 22.5 € à 0.5 € / MWh
  - A priori, hausses contenues et financements assurés pour 2022
- Demande de réduction de consommation de 10 %
  - *« Sans compromettre la recherche »*
  - Arrêts partiels des machines cet hiver ?

Perspectives 2023

- Facteur de 3 à 5 sur le prix de l'électricité
  - 400 € / MWh TTC (aujourd'hui 80 € / HT)
- Hausse non financée aujourd'hui
- Problème qui dépasse très largement les infrastructures numériques
  - TGIR : CERN, Soleil, …

4 Inquiétude : énergie

A plus long terme ?

- Evolution des tarifs ?
- Machine Exascale dimensionnée à 20 MW à échéance 2025 – 2026 ?
  - A comparer à la puissance installée du parc GENCI actuel d'environ 4 MW.

Solutions ?

- A court terme :
  - Réduction de la consommation (arrêts partiels des machines)
  - Financements supplémentaires / « bouclier tarifaire » ?
- A plus long terme
  - Rationalisation des infrastructures *(datacentres à l'état de l'art)*
  - Optimisation des codes
  - …

Coûts et financements de la science ouverte?

- Infrastructures numériques, espaces de stockage, …
- Modèle économique ?

5 En route vers exascale…

Machines de classe pré-exascale et exascale

- Accès à des ressources européennes significatives
- Machines massivement accélérées (GPU, …)
- Nécessité de portage / adaptation / réécriture des codes
  - Besoins ressources humaines, non financées aujourd'hui

PEPR NumPEX

- Direction INRIA, CNRS et CEA
- Recherche « amont » en support exascale
  - Bibliothèques, algorithmes, …
  - Pas support au portage !!!
- Présentation au prochain Forum ORAP
  - Fin novembre, Maison de la Simulation *(date non finalisée)*