Institut national de physique nucléaire
et de physique des particules

Calcul, données et science ouverte
Rencontre équipes physique nucléaire

11 juillet 2022

Sabine Crépé-Renaudin

Calcul et données à l'IN2P3

Les données et les logiciels, des éléments centraux de nos sciences

- données de simulation, d'expérience (brutes, reconstruites ou d'analyse), métadonnées, log, publications…
- logiciels pour les manipuler

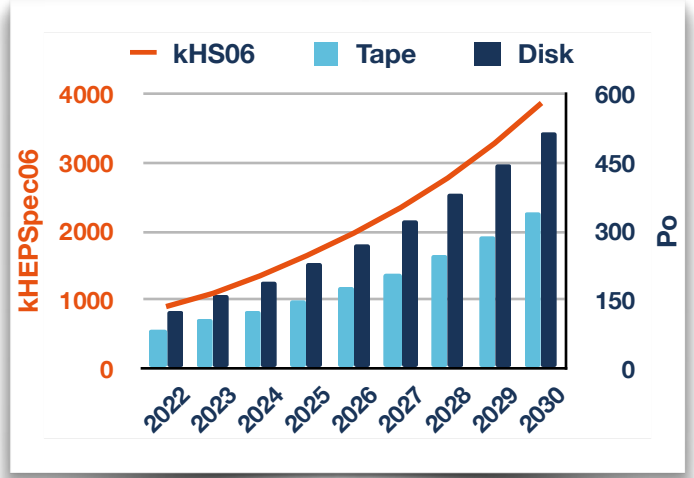
Des demandes en informatique en forte croissance pour les prochaines années

- prochaines générations d'expériences (HL-LHC, LSST etc)

➜ 1 exaoctet de données à l'IN2P3 d'ici 2028

- problématiques différentes en fonction des thématiques
  - grandes quantités de données distribuées, images et grandes bases de données, données multiples d'expériences distribuées

➜ Importance de gérer l'ensemble de ces données avec rigueur et méthode et avec les bons outils



Le contexte de la science ouverte

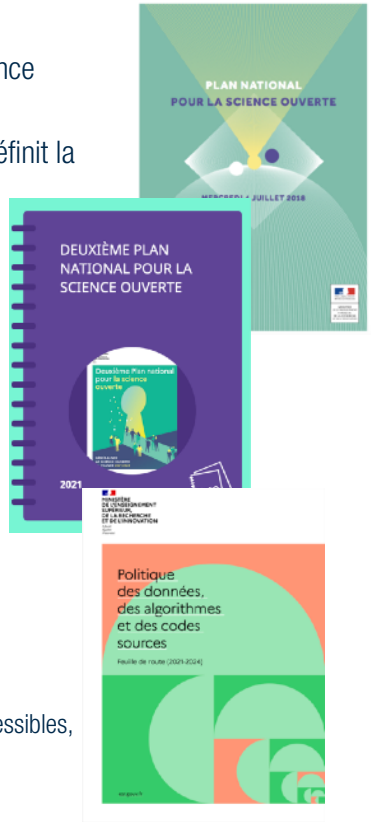
Contexte général

Contexte international

- UNESCO – Recommandation sur la science ouverte, OCDE – Recommandation sur les données de la recherche, G7 Open Science Working Group
- L'Union Européenne demande l'ouverture des publications et des données des recherches qu'elle finance; depuis 2021, elle définit la science ouverte comme un critère d'excellence scientifique.
- Promotion de l'EOSC – European open science cloud
- Groupes de travail internationaux : RDA, GO FAIR...

Contexte national

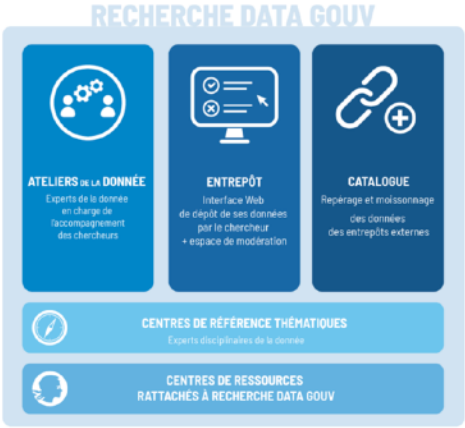
- Loi pour une République numérique (2016)
- Plan national pour la science ouverte (2018)
  - Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics.
  - Créer la fonction d'administrateur des données et le réseau associé au sein des établissements.
  - Créer les conditions et promouvoir l'adoption d'une politique de données ouvertes associées aux articles publiés par les chercheurs.
- deuxième Plan national pour la science ouverte (2021)
  - Mettre en œuvre l'obligation de diffusion des données de recherche financées sur fonds publics
  - Créer Recherche Data Gouv, la plateforme nationale fédérée des données de la recherche
  - Promouvoir l'adoption d'une politique de données sur l'ensemble du cycle des données de la recherche, pour les rendre faciles à trouver, accessibles, interopérables et réutilisables (FAIR)
- Feuille de route 2021-2024 (2021)



Actions du ministère

Organisation

- COSO : Comité pour la science ouverte présidé par la DGRI
- Secrétariat permanent pour la science ouverte (SPSO)
- Groupement d'intérêt scientifique Fonds national pour la science ouverte (GIS FNSO) => financement
- https://www.ouvrirlascience.fr/accueil/



Plateforme Recherche.data.gouv

- Entrepôt
  - catalogue + dépôt des données (si pas d'entrepôt thématique national ou international)
  - basé sur dataverse, piloté par l'INRAE, 1,1M€ 13 ETP pour 2021, V1 pour T2 2022 + accompagnement chercheur 3M€
  - chaque établissement pourra créer son espace institutionnel au sein de la plateforme => portail CNRS
- Ateliers de la donnée
  - Objectif : structurer au plan territorial le réseau des ateliers de la donnée, labelliser les initiatives nouvelles ou existantes, accompagner dans leur réflexion les établissements pendant les phases de maturation de leurs propositions
  - Ateliers = point d'entrée en proximité des équipes de recherche sur toute nature de besoin relatif à la donnée.
- Centres de référence thématiques : le CNRS se positionne pour l'organiser
  - Objectif : définir les référentiels thématiques et disciplinaires

Organisation des Journées européennes de la science ouverte (OSEC)

- https://osec2022.eu/ 4-5 février 2022 àParis (Online)

OSEC 2022
Open Science European Conference
4-5 février 2022
Académie des sciences, Paris - France

FRANCE22

Actions du CNRS

Feuilles de route

- Feuille de route du CNRS pour la Science Ouverte (Novembre 2019)
- Officialisation du Plan Données de la Recherche du CNRS 2020

Création de la DDOR : Direction des Données Ouvertes de la Recherche (novembre 2020)

- Fusion de la DIST et de la mission MICADO sur les aspects calcul et données

Journées Science Ouverte au CNRS

- 2020 : bilan science ouverte au CNRS
- 2021 : autour de l'évaluation => abandonner les évaluations quantitatives (nbre public, hindex) et aller vers évaluation qualitative

Création d'un annuaire des dépôts et des services de données du CNRS

- en construction à partir Cat Opidor le wiki des services dédiés aux données de la recherche, maintenu par l'INIST
- contribution via les instituts du CNRS
- dans l'annuaire les entrepôts et services de données auxquels le CNRS participe, concept étendu pour intégrer nos données ouvertes à l'international
- objectif : affichage et base de données pour les chercheurs

Cas d'usage par institut pour aider les thématiques éloignées avec spécialistes de l'INIST et DDOR

- pour mieux comprendre les besoins concrètement et aider au démarrage

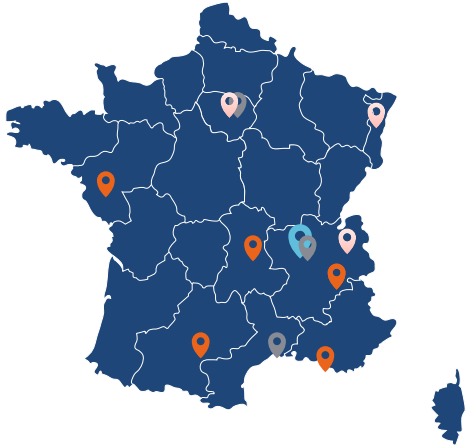
Institut national de physique nucléaire et de physique des particules

À l'IN2P3

Des infrastructures et un savoir faire

Des infrastructures

- Centre national CC-IN2P3, Tier1 de WLCG
- Les Tier2s de WLCG, certains sont aussi des centres régionaux liés en général aux universités



Expertises dans les laboratoires et au CC-IN2P3

- Expertise dans le référencement, la gestion, l'accès et le stockage des grandes quantités de données avec des technologies diverses.
- Expertise dans la mutualisation de l'infrastructure, des ressources et services de stockage
- Expérience dans le stockage de masses de données pendant de longues périodes (plus de 25 ans)
- Pas (peu) de données ouvertes directement stockées dans l'institut

Expertises via les collaborations internationales

- chaque collaboration établie sa politique de gestion et d'ouverture des données
- les personnels de l'IN2P3 participent aux développements des outils associés

Les outils pour les données

Stockage des données et calcul

- disques, bandes
- stockage de masse et stockage distribué
- HTC, GPU, HPC, cloud

Référencement des données

- expertises en IST
- expertises en bases de données
- gestion des métadonnées des expériences : ami

Distribution des données et accès aux données

- Intergiciels de gestion des calculs et des données
  - Rucio, Panda, DIRAC
- transferts et accès aux données xrootd, dpm, FTS…

Logiciels

- gitLab
- licences

Cycle de vie des données

- Plan de gestion des données
- défini et disponible pour le CC sur DMPOpidor/INIST et RDMO

Ouverture des données

- Gestion des embargos
- Entrepôt de données
  - pour leur publication et leur accès ouvert
- Documentations et logiciels pour leur réutilisation
- Archivage sur le temps long

Politique de gestion des données

Pour les nouvelles expériences : information du DAS thématique et du DAS calcul et données

- ➜ description des besoins
- ➜ chiffrage (CC)
- ➜ décision de la direction des contributions au traitement des données de l'expérience

Rédaction du DMP

- pour réfléchir dès le départ au cycle et aux besoins
- document évolutif à remettre à jour chaque année
- définit les données, la façon dont elles sont décrites, qui en a la responsabilité, les besoins de stockage et de traitement et leur devenir :
  - combien de temps on garde quelle donnée
  - faut-il pouvoir les relire/réutiliser dans le temps ➜ archivage (garder à la fois les données et les logiciels et s'assurer qu'on peut les utiliser)
  - ouvrir ou pas ? quelle données ? où ?

Stocker au moins une copie des données au CC

- au CC demande annuel des ressources (stockage et calcul)
- mise à jour annuel des projections des besoins

Données ouvertes

De quoi parle-t-on ?

- de bonnes pratiques pour la gestion des données sur tout le cycle

  ➜ collecte ➜ description/identifiant DOI ➜ référencement ➜ stockage ➜ traitement et analyse ➜ effacement ou archivage ➜ ouverture… ou pas

Quels besoins à l'IN2P3 ?

- Bonne expertise et mise en place de la première partie du cycle

  ➜ à étendre à l'ensemble des thématiques

➜ Recenser les expertises sur la fin du cycle dans l'institut

  - ex Virgo, Auger, LHC, autres ?
  - recenser les outils, les entrepôts existant utilisés

➜ Évaluer les besoins sur l'ensemble du cycle et particulier sur les entrepôts, l'archivage et l'ouverture (la mise à disposition) des données

➜ Mise en place des outils nécessaires

➜ Définition et application de la politique de données ouvertes de l'IN2P3

  ➜ Chargé de mission données ouvertes

Données et physique nucléaire

Objectifs de la réunion

Comprendre

- vos usages/pratiques en matière de gestion et traitement des données
- vos développements
- vos besoins

Prendre connaissance et échanger autour de

- vos projets
- vos idées

Pour à terme construire une stratégie pour atteindre nos objectifs en matière d'ouverture des données

- pour la physique nucléaire
  - développement des outils et des pratiques
  - opportunité projet EuroLabs + FITS
- pour l'ensemble des thématiques de l'IN2P3

des questions ?