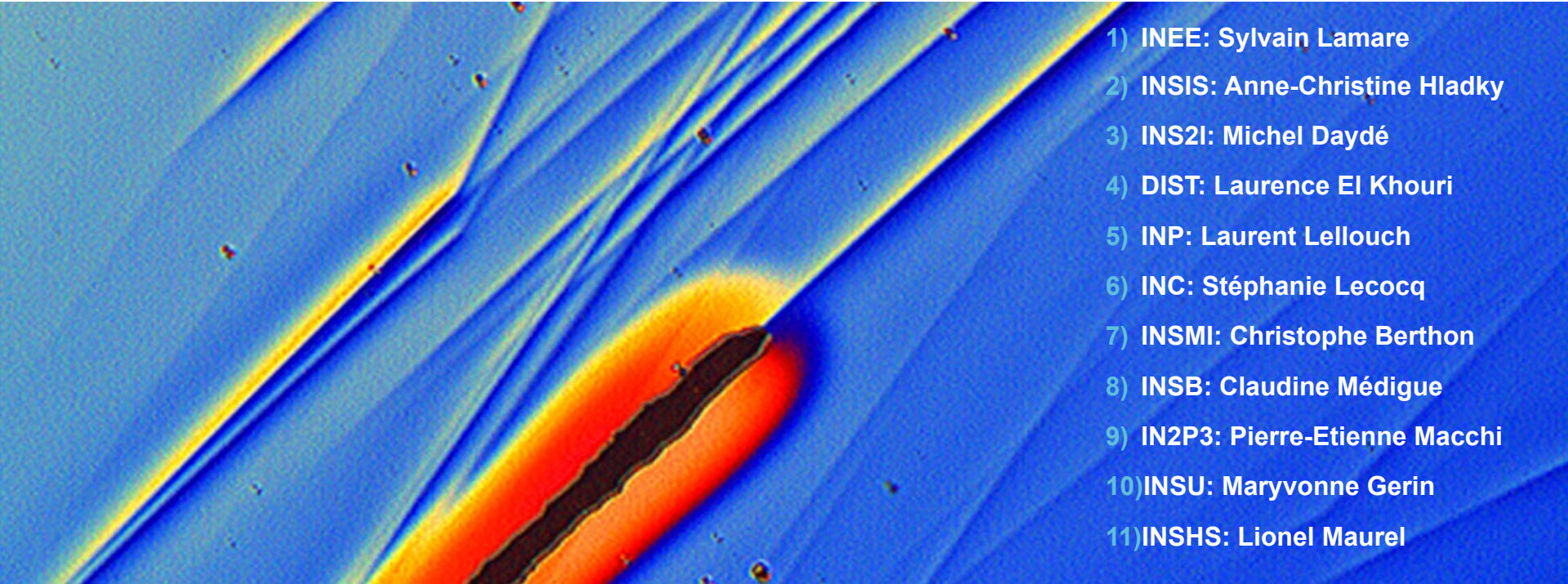
Atelier EOSC CNRS
« Qu'attend mon institut de l'EOSC? »



1) INEE: Sylvain Lamare
2) INSIS: Anne-Christine Hladky
3) INS2I: Michel Daydé
4) DIST: Laurence El Khouri
5) INP: Laurent Lellouch
6) INC: Stéphanie Lecocq
7) INSMI: Christophe Berthon
8) INSB: Claudine Médigue
9) IN2P3: Pierre-Etienne Macchi
10) INSU: Maryvonne Gerin
11) INSHS: Lionel Maurel

1 INSTITUT ECOLOGIE ET ENVIRONNNEMENT

SITUATION

1.Hétérogénéité des données (et des communautés scientifiques) :en fonction du domaine , nn fonction des technologies, en fonction des usages

2. Dispersion de la donnée : au niveau des Infrastructures, au niveau des organismes de recherche, a u niveau des laboratoires et projets

3. Interactions complexe entre les différents niveaux d'organisation du vivant , les facteurs abiotiques, les pressions d'origines variées

DEFIS

⇒ Mettre à la disposition de la communauté scientifique des voie d'accès simples et directes aux données/métadonnées

⇒ Rendre les données Facile à trouver, Accessibles, Interopérables et Ré-utilisables (FAIRisation des données de biodiversité)

⇒ Accompagner la structuration en cours des communautés scientifiques concernées, les sensibiliser aux plans de gestion des données.

Les données en Ingénierie et à l'INSIS dans la perspective EOSC

Caractéristiques

- *Équipes de recherche produisant chacune leurs propres données*, mais dans un cadre scientifique national bien structuré
- *Peu de pratiques de partage* de données entre les équipes
- *Peu d'organisation par communautés* : *e.g.* manque de formats standards, balbutiements en terme de plans de gestion de données (DMP)
- *Des données par nature de très grande taille* (>> O(To)): métrologies résolues dans les expériences ; simulations numériques spatio-temporelles ; capteurs en très grand nombre ; …
- Quelques initiatives de création de bases (mécanique des fluides), mais pas de dépôt systématique, et fonctionnalités d'échange limitées
- Forte utilisation des moyens de GENCI et PRACE, et infrastructures de type *Grid*

Apports possibles d'EOSC

- Incitation des communautés ingénierie à participer à des projets EOSC
- Aide à la création de formats et à l'utilisation des DMP, par l'exemple d'autres disciplines
- Structuration d'initiatives en terme de création de bases de données
- Infrastructures d'accueil de bases pour le dépôt et l'échange de données, stockage et pérennisation des bases
- Interaction avec des communautés interdisciplinaires (géophysique, matériaux, chimie, etc.)

QU'ATTENDS INS2I DE EOSC ?

- INS2I impliqué dans EOSC depuis son démarrage au travers du projet EOSCPilot
- Points d'intérêts :
  - Intégration des ressources de calcul HPC et relation EOSC avec EuroHPC et PRACE
  - Plateformes de données PerSCIDO (hébergé par GRICAD), OSIRIM, GALACTICA, … pour les recherches INS2I (réseaux sociaux, corpus, données médicales, imagerie, …)
  - Plateforme d'expérimentation à grande échelle SILECS (cloud, big data, IA, IoT, …)
  - Stockage sûr pour le partage d'algorithmes et de logiciels (Software Heritage d'INRIA)
  - Gestion et déploiement de services dans un environnement hétérogène distribué à grande échelle

EOSCpilot
The European Open Science Cloud for Research Pilot Project

DIST et l'EOSC – European Open Science Cloud

Proposer au niveau européen des services et infrastructures de science ouverte

- **Feuille de route science ouverte du CNRS : Accès ouvert aux productions scientifiques (publications, données), partage FAIR des données, internationalisation**
- **Développer une culture FAIR** de la gestion et du partage des données chez tous les acteurs du cycle de vie de la donnée.
- **Placer les instituts au centre de la stratégie**, car les pratiques des communautés scientifiques sont différentes d'une discipline à l'autre.
- **Aider les infrastructures de recherche** dans leur mise en place d'une politique de gestion des données
- **Soutenir et accompagner les entrepôts de données** y compris pour toutes données y compris celles de la longue traîne
- **Accompagner les chercheurs** dans les outils de gestion des données et le dépôt conjoint publication/données, notamment grâce à l'outil DMP OPIDOR pour remplir les plans de gestion des données, et l'attribution des DOI via Datacite (INIST)
- **Participer à la constitution d'une communauté française** autour du nœud français RDA

Institut de physique et EOSC

- **Pas de participation institutionnelle directe**

- **Implication au travers des TGIRs « Photon and Neutron (PaN) »** (voir présentation d'A. Götz)
  - PaNOSC (ESRF, ILL, INFRAEOSC-04-2018): FAIRiser les données de 6 IRs européens, développer/déployer des services pour les analyser et interfacer le tout avec EOSC
  - ExPaNDS (Soleil, INFRAEOSC-05-2019): déclinaison de PaNOSC pour 11 IRs nationaux

- **Quid des plateformes plus petites?**
  - Microscopes électroniques et sondes atomiques (METSA, fédération de 8 plateformes) produisent de plus en plus de données
  - e.g. NanoMAX (Tempos): croissance crystalline en haute résolution et en temps réel
    ➔ dizaines de To/j
  - Traitement et FAIRisation des données?

Si vous travaillez dans un labo de l'INP et vous créez des outils ou des solutions pour la science ouverte, contactez moi (ici ou à laurent.lellouch@cnrs-dir.fr)

- **Modèle possible pour accès à des infrastructures de pointe et pour dissémination des bonnes pratiques**
  ➔ solutions autour des centres nationaux (CC-IN2P3 pour HTC, IDRIS pour HPC et IA)
  ➔ expertise des IRs PaN en solutions de FAIRisation et d'analyse de données dans un contexte multidisciplinaire
  ➔ nouveaux AAPs MITI et instituts pour le développement de nouvelles techniques d'indexation, d'analyse, … de gros volumes de données?

- **E.g. projet ANR Flash porté par INC:** avec ILL, Soleil et IRs Infranalytics
  ➔ étendre système d'attribution de DOI de l'ILL à d'autres IRs, développer des DMPs modulaires et des métadonnées d'échantillon domaine-spécifiques

- **Pour équipes avec données moins importantes et/ou communautés moins organisées:** OSC national ou CNRS avec outils simples de FAIRisation serait très utile

- **FAIRisation requiert infrastructures et ressources financières et humaines: qui paye quoi?**

INC AND EOSC

INC EOSC project participation

Data sharing and data reusage are relatively new concepts in the chemistry communities.

Right now, INC is not involved in any EOSC project.

INC EOSC interests

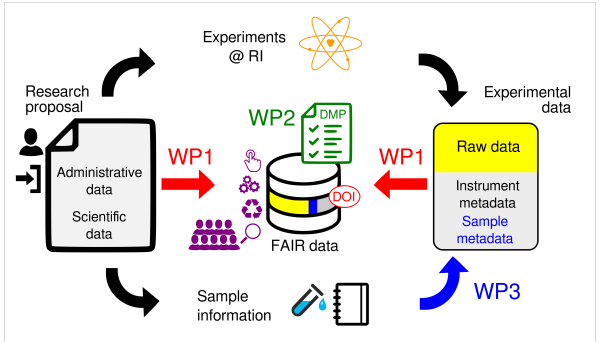
- FAIRisation of data,
- creation of trusted repositories,
- interoperability and interfaces between existing repositories,
- seamless access to HPC systems for simulations

A project of data policy implementation is currently under way at the level of three Research Infrastructures.

Objectives : FAIRisation of the data produced through:

- **the establishment of DMPs,**
- **the allocation of persistent identifiers to datasets,**
- **the standardization of sample metada.**

This work is conducted jointly with ILL and Soleil.



EOSC pour l'Insmi

- Pourquoi, comment ?

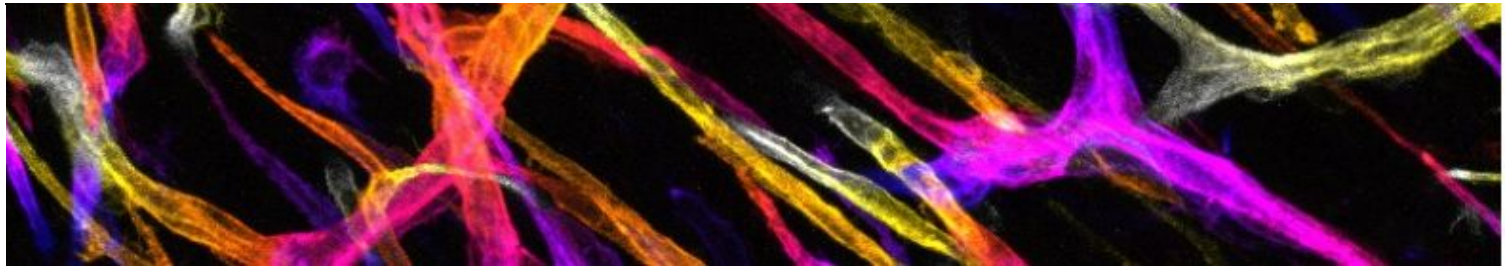
Insmi, acteur des données

FAIRisation, changement de paradigme

- Qu'est-ce qu'une donnée FAIR ?
- Pourquoi une donnée FAIR ?
- La mise en œuvre.

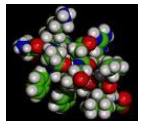
Institut des sciences biologiques

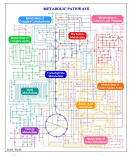
Décrypter la complexité et la diversité du vivant



- **Volumétrie**, en particulier des données de séquençage (pré- et post-NGS) et imagerie
- **Multiplicité des sources**
  - Données produites partout et souvent éloignées des centres de traitement
- **Transfert des données** production->serveurs->utilisateurs
- **Diversité**
  - Multi-omiques (génome, transcriptome, protéome, métabolome, métagénomique)
  - Images
  - Structures macromoléculaires
  - Réseaux biomoléculaires
- **Complexité**
  - représentation des connaissances biologiques
- **Connection à d'autres types de données**
  - Données de santé, phénotypes, écotypes, climat, ...
- **Médecine personnalisée**
  - Couplage données omiques / données de patients
  - Protection des données « sensibles »


Genomics

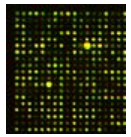

Structural biology

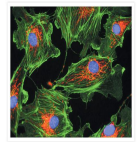

Proteomics Metabolomics


Bibliographic


Environmental

Intégration de données hétérogènes


Transcriptomics



Biological and medical images


Cohorts Biobanks

EOSC & l'INSB

- **Données FAIR**:
  - **Trouvabilité :** identifiants
  - **Accessibilité** : immédiate, sécurisée (« trusted repositories »)
  - **Interopérabilité** : identifiants, références croisées, formats standards, vocabulaire contrôlé, métadonnées, …
  - **Ré-utilisabilité** : « figer » également l'environnement logiciel (containers, VMs)
- **Data Management Plan**
- **Partage de services** (en particulier, catalogue d'outils)

=> Coordination de la participation à EOSC à travers IFB

L'IN2P3 et l'EOSC

L'IN2P3 participe à des :

collaborations internationales

de durée longue

de plus en plus productrices de données

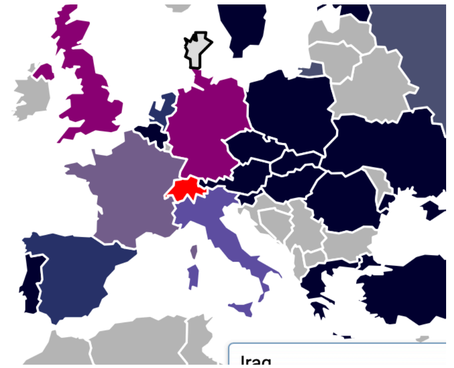
➡ infrastructures de traitement de données **distribuées**

➡ accès à long terme aux données

➡ solutions soft et hard évolutives

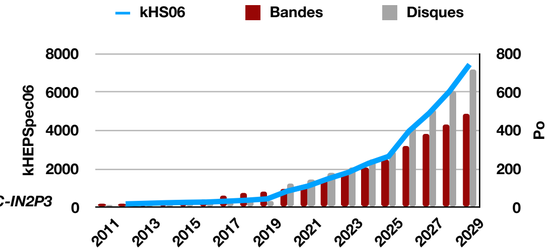
➡ technologies novatrices (grid, virtualisation…)

- Principal contributeur à France Grilles
- Représentant français à EGI



4° fournisseur d'heures CPU normalisées en 2019

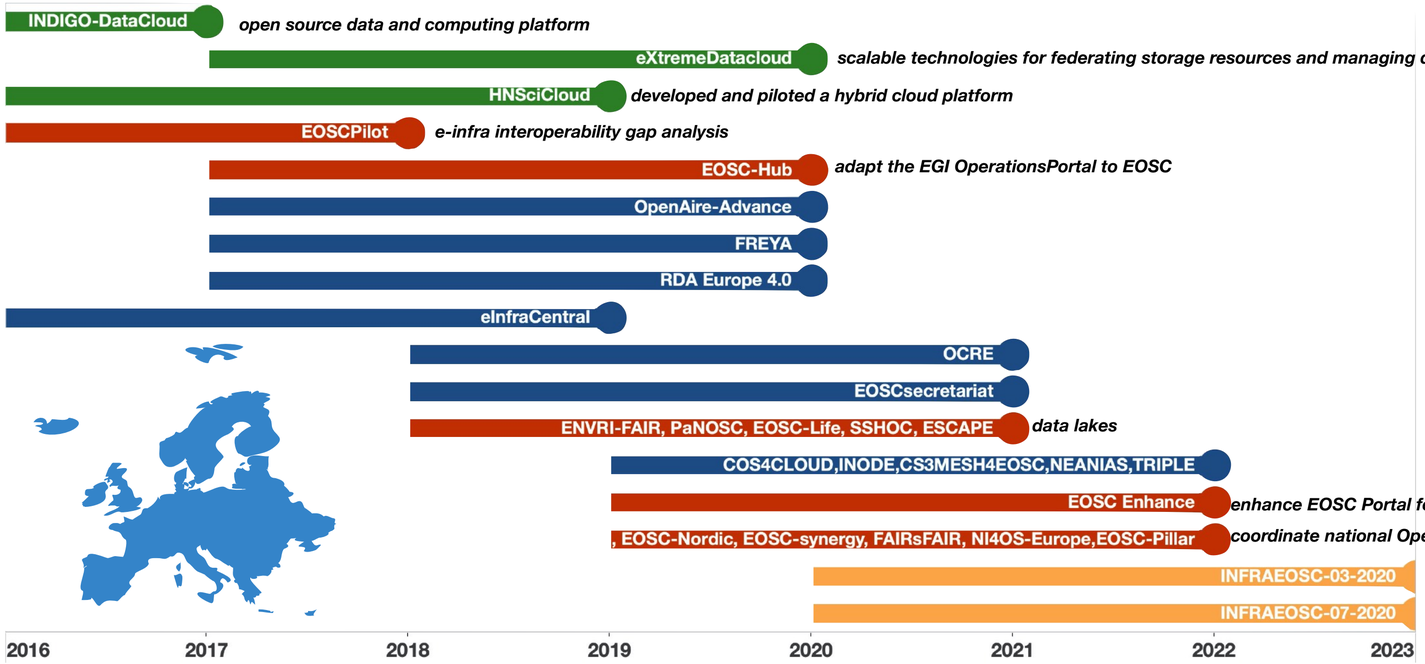
Nouvelles expériences (LSST, Euclid, Spiral, HL-LHC, Dune…) **décuplent** les besoins :



Evolution du volume de données stockées au CC-IN2P3

+ Open Science

➡ Participation aux projets européens



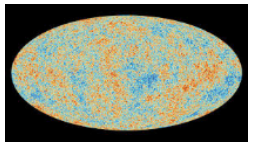
Objectifs partagés avec l'EOSC :

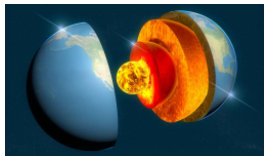
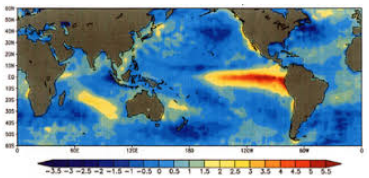
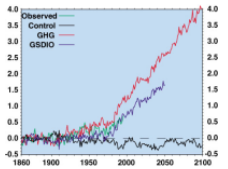
➡ Interopérabilité des données, des infrastructures

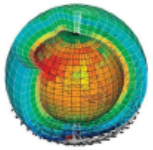
➡ accès transparent et aisé à des infrastructures hétérogènes

Timeline des projets européens relatifs à l'EOSC
en vert : projets avec participation IN2P3 dont l'appel ne faisait pas référence à l'EOSC mais dont les résultats ont eu une incidence
en orange : projets avec participation IN2P3 dont l'appel ne faisait pas référence à l'EOSC
en orange léger : futurs projets avec participation IN2P3

INSU

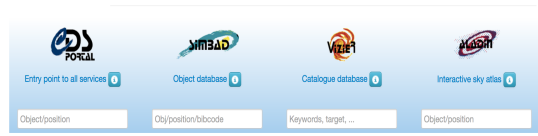





- Forte expérience de la production, mise à disposition des données, et fourniture de services
  - 2 centres de données nationaux thématiques CDS et Data Terra comme fer de lance des actions nationales
  - Acteurs intégrés aux systèmes européens et internationaux de distribution de données
  - Equipes intégrées multi-compétences (Recherche – Documentation – Informatique)
  - Précurseur du Contexte FAIR avec l'Observatoire Virtuel Astronomique et la fédération de centres de données interopérables en sismologie
  - Passage à la Certification des centres de données (déjà plusieurs centres WDS/CTS; >10 certifications planifiées avant 2022).
- Données hétérogènes, complexes, volumes & flux en très forte augmentation
  - Observations sol, mer, air et spatiales – long terme : TGIR et IR, Services nationaux d'Observations maintenant et dans le futur (SKA), services internationaux (p ex magnétisme …)
  - Modèles et codes communautaires
  - Prélèvements in situ: carottes, échantillons
- Forte participation aux projets liés à EOSC :
  - RDA; Clusters ENVRI + ENVRIFAIR, ASTERICS & ESCAPE ; PHIDIAS ; EOSC-Pillar; ESFRI: AENEAS, ACTRIS, ICOS, EPOS, EMSO ..
- Rôle attendu de l'EOSC sur :
  - Un modèle de plateforme logicielle de services distribués de calcul et de données
  - Accélérer la logistique des données tout au long des workflows entre infrastructures périphériques et centralisées
  - Validation, homogénéisation, et support des standards issus des communautés scientifiques
  - Un cadre et des moyens en support au maintien et évolution de codes, librairies et logiciels libres (p. ex. software heritage)
  - Interface avec les systèmes de partage de données multi-sources existants (OV)
  - Renforcer la visibilité et le support pour l'utilisation de services standardisés de calcul et de données par les communautés scientifiques
  - Renforcer les ressources et les services fournis par un continuum d'infrastructures (calcul, stockage, réseau)
  - Soutenir et agréger des environnements humains multi disciplinaires en support
  - Accélérer la convergence HPC-HPDA et l'utilisation des méthodes de type AI (Machine Learning et Deep Learning)

Centre de Données astronomiques de Strasbourg
Strasbourg astronomical Data Center

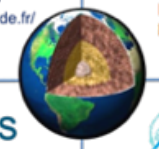


DATA TERRA

Data & Services Hubs

ForM@Ter
Solid Earth
Data and services hub
https://en.poleterresolide.fr/

Theia
Pôle Thématique Surfaces Continentales
Land Surface
Data and services hub
http://www.theia-land.fr/en



ODATIS
Ocean
Data and services hub
http://www.odatis-ocean.fr/en/

AERIS
Atmosphere
Data and services hub
http://en.aeris-data.fr/

Les services EOSC doivent répondre aux besoins et à la diversité des pratiques de recherche des communautés scientifiques.
Ils doivent être pérennes, flexibles et robustes