PIA3: orchestration des flux tout au long de la vie des données

https://frama.link/ifb-ag20-mudis4ls

Noms

Orchestration des flux de données

Life cycle of the data – The typical stages

1. Production
(weeks - months)

National Infrastructures for Biology and Health (INBS)

France Génomique · France Médecine Génomique 2025 · aviesan · EMBRC France · Réseau Actine-Tuface · METABOHUB · Myobank · ProFI Proteomics · France-BioImaging · FLI France Life Imaging · FRISBI · phenomin · PHENOME EPPN · CONSTANCES

Preprocessing *at the source* of data production

2. Analysis
(typically 6 months - 4 years)

Bioinformatics services + R&D

ifb · elixir FRANCE

Compute, Storage, Databases, interoperability, training, support, ...

ELIXIR platforms and communities

Integrative bioinformatics

Mid-term storage (securing during the project)

3. Conservation
(5, 10, 20 years)

Data persistence
(regional + national centers)

GENCI · IDRIS · CINES · CC IN2P3 · TGCC

Data repositories
(national, international)

ENA European Nucleotide Archive · BioImage Archive · ProteomeXchange · PRIDE Archive PRoteomics IDEntifications Database · EMPIAR · MetaboLights · data INRAE · zenodo · Dataverse · cirad

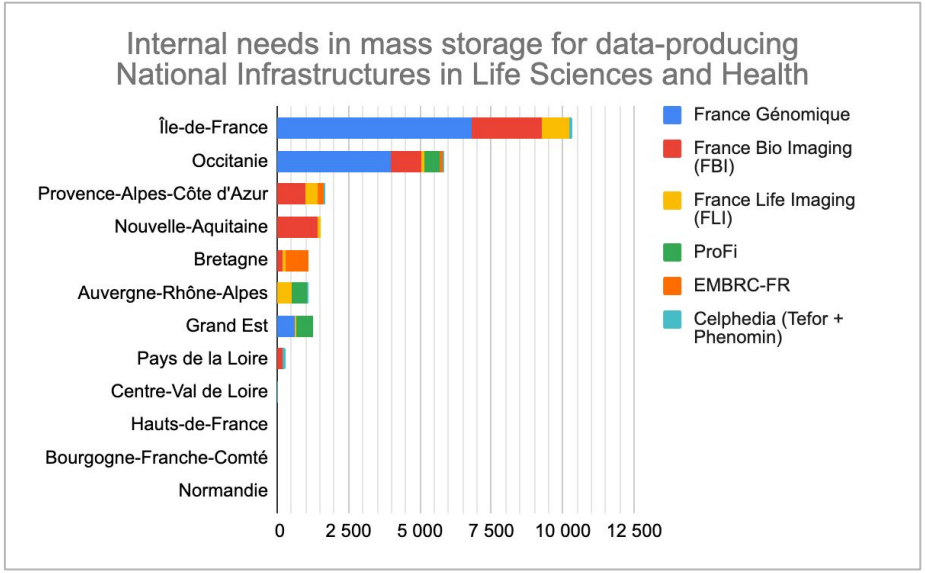
Other needs?

Cartographie des besoins de stockage des INBS

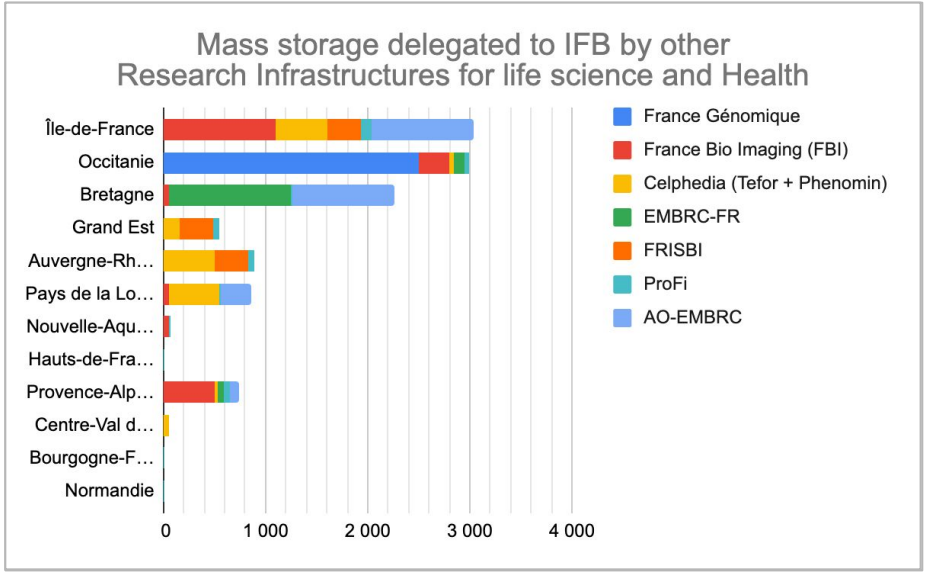
Nous avons demandé à chaque infrastructure nationale de biologie-santé **INBS** d'estimer la répartition par région

- Des besoins de stockage internes (numérique accolé aux machines de production ou autre raison)
- Délégués à l'IFB (notamment pour intégration avec autres types de données)

Besoins de stockage internes

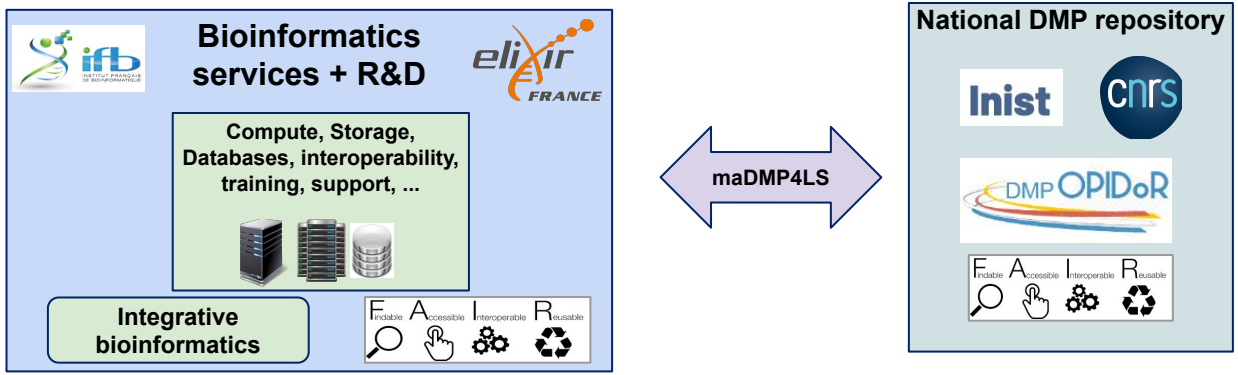


Besoins de stockage délégués à l'IFB



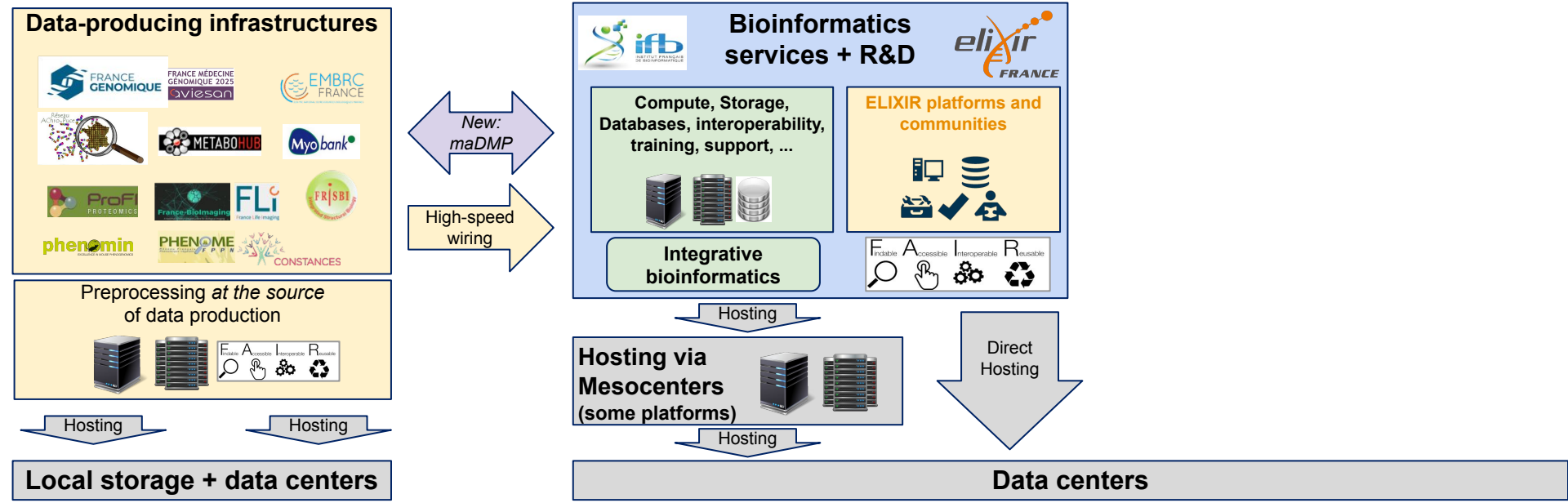
Orchestration du flux de la donnée tout au long du cycle de vie

- Projet-flash ANR **machine-actionable DMP for Life Sciences** (**maDMP4LS**),
  - allocation des ressources (espace-projet, volume, ressource calcul, accès collaborateurs) par accès programmatique aux DMP déposés sur OPIDoR (INIST)
  - mise à jour conjointe du DMP et des ressources au fil de l'évolution du projet
- **Gestion à chaque phase** des données et des métadonnées : production (collaboration avec autres infrastructures), analyse (serveurs IFB), conservation (centres de stockage), accès (dépôts nationaux et internationaux).
- Programmation des flux de données via le maDMP (extension du projet maDMP4LS) : des sites de production aux serveurs d'analyse, puis de ceux-ci vers les centres de stockage et vers les dépôts.



De la production à l'analyse des données

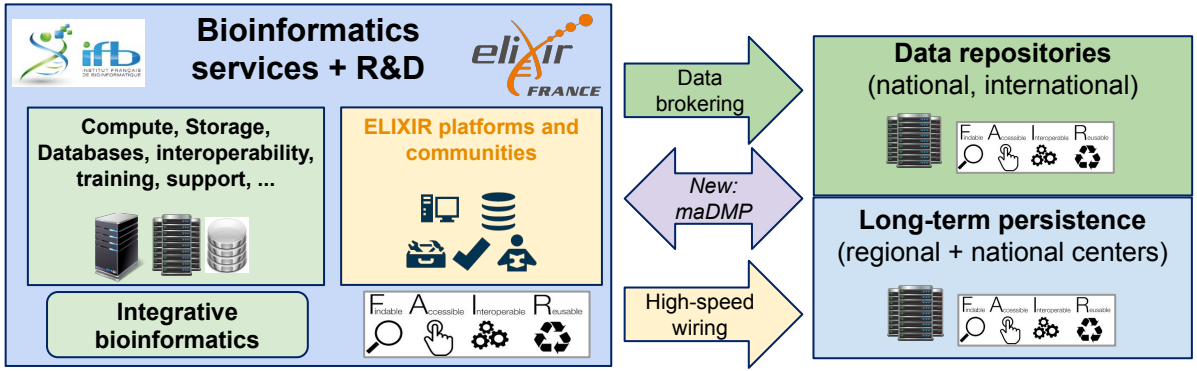
- Les infra productrices de données nécessitent du calcul et stockage local
- Offre IFB: environnements spécialisés pour l'analyse + intégration + interopérabilité des données
- Description, dès la conception du DMP, des modalités de transfert des plateformes de production/preprocessing vers les plateformes IFB pour l'analyse
- Automatisation de ces transferts (maDMP)



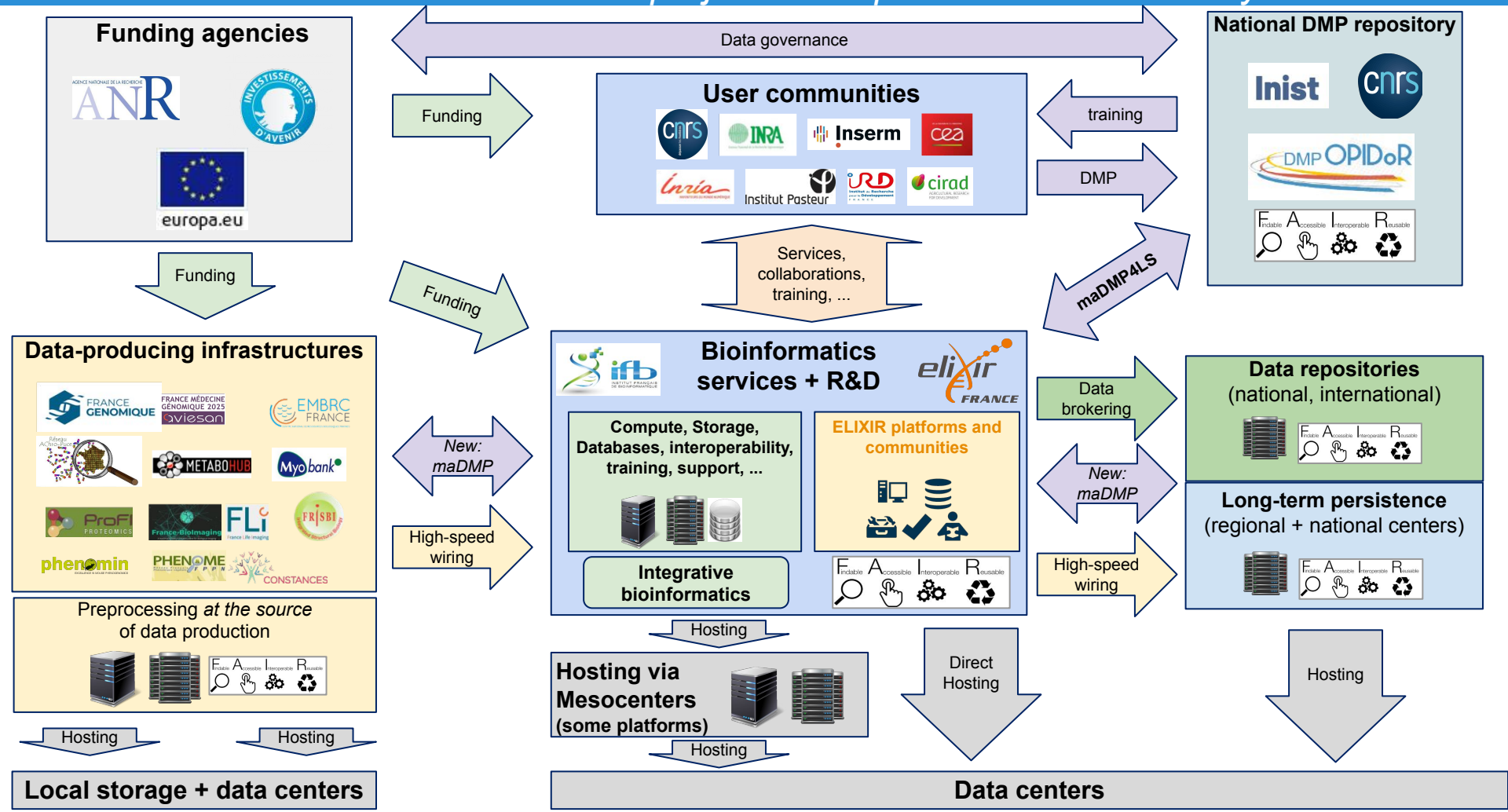
Conservation et accessibilité des données

- **Conservation des données**:
  - ni dans nos missions, ni dans nos projets
  - collaborer avec les data centres régionaux et/ou nationaux
  - Rôle de l'IFB: FAIRifier les données conservées
    - Formats standards
    - Annotation des données (métadonnées)
- **Dépôts**
  - Carence actuelle de solution au niveau national
  - Existence de quelques solutions institutionnelles (INRAE, CIRAD, IRD, CEA)
  - Dépôts internationaux spécifiques d'un type de données (séquences NGS, images, ...)

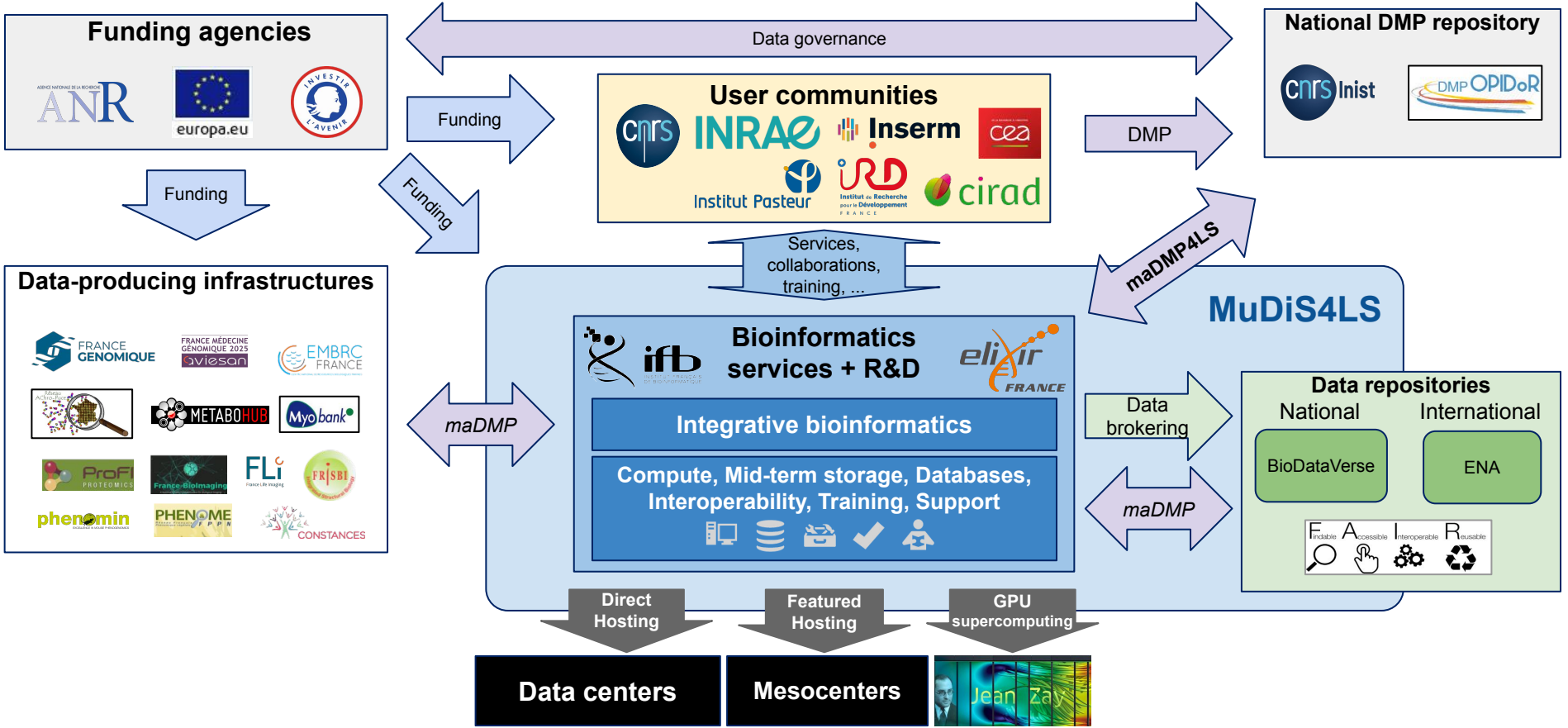
- **Data brokering**
  - projet-pilote avec European Nucleotide Archive (EBI) pour que l'IFB devienne courtier des données françaises. Extension envisagée à d'autres dépôts.
- **Elargissement et défis**
  - Extension aux autres dépôts internationaux ?
  - Traitement des données d'imagerie ?



At the cross-road of data flows – from project conception to results delivery



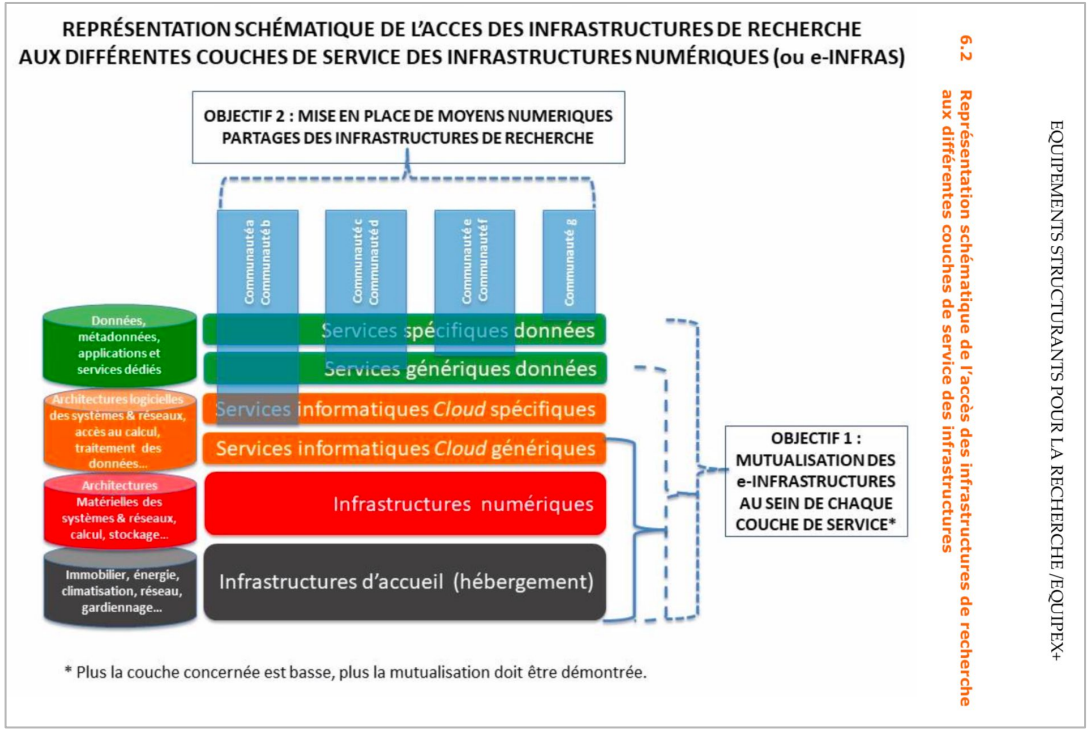
Orchestration des flux de données



Statistiques et figures

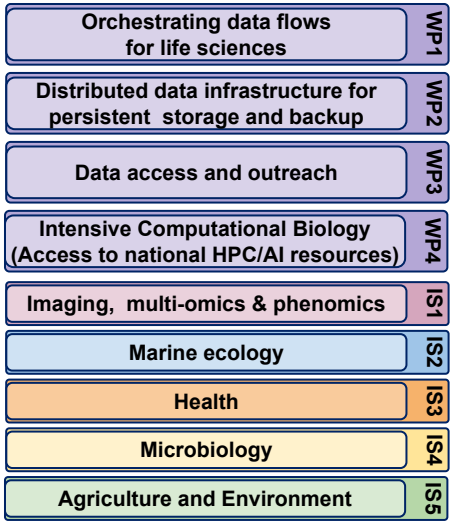
Eléments-clés du de l'AMI ESR/Exuipex+

- **Rationalisation** des moyens numériques nationaux
  - Actuellement : 30.000 salles de calcul en France (labos, instituts, mésocentres, centres nationaux)
  - Demain : 30 data centres nationaux et régionaux labellisés
- **Mutualisation** des moyens numériques et humains
- **Engagement des établissements partenaires** → recrutement de personnel permanent
- **Structuration des services**
  - Infrastructure d'accueil (bâtiments, climatisation, électricité, réseau)
  - Infrastructures numériques (moyens de stockage et calcul)
  - Services informatiques
  - Services données

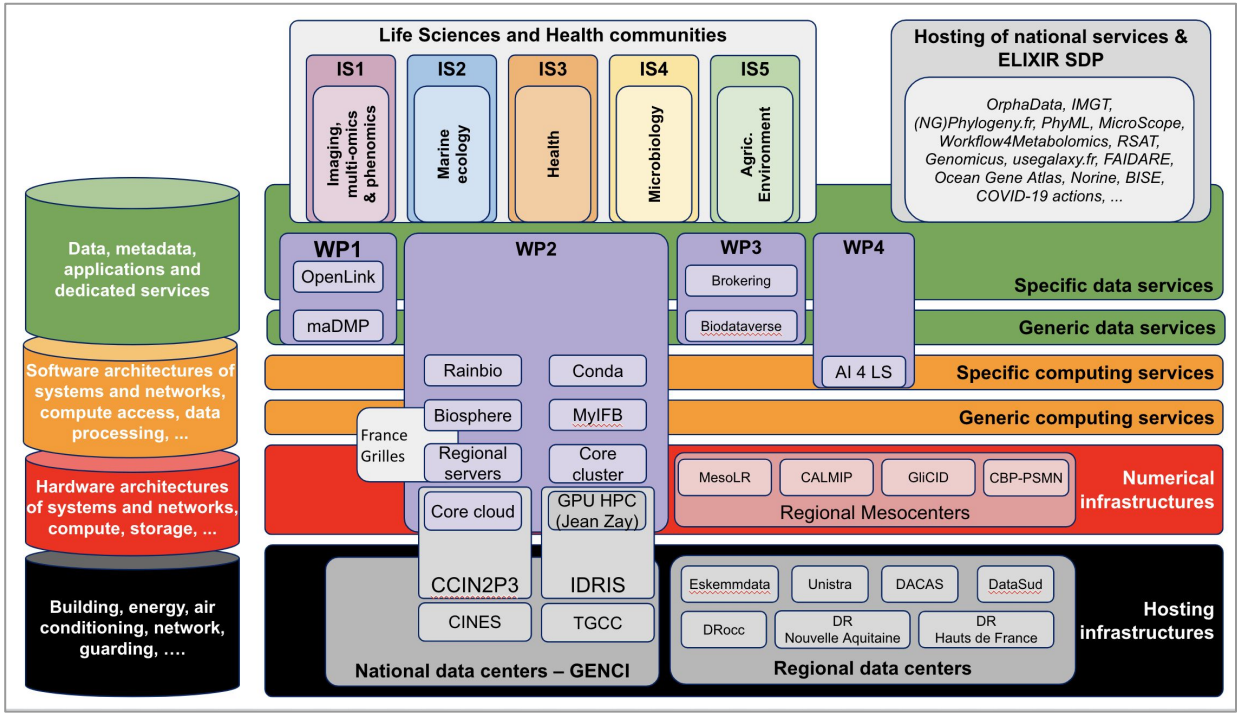


Key elements of the MUDIS4LS project

- 39 partner teams
- 14 partner organisms
- 4 national + 7 regional data centers
- 4 partner mesocenters
- 170 persons, 39 FTE;
  → 2.49 person•centuries
- Requested funding: 20M€
  - 13.70 M€ equipment
  - 3.53 M€ functioning
  - 1.28 M€ personnel
  - 1.81 M€ overheads
- 4 technological work packages (WP)
- 5 thematic implementation studies (IS)
- Engagement des tutelles : 7 postes permanents



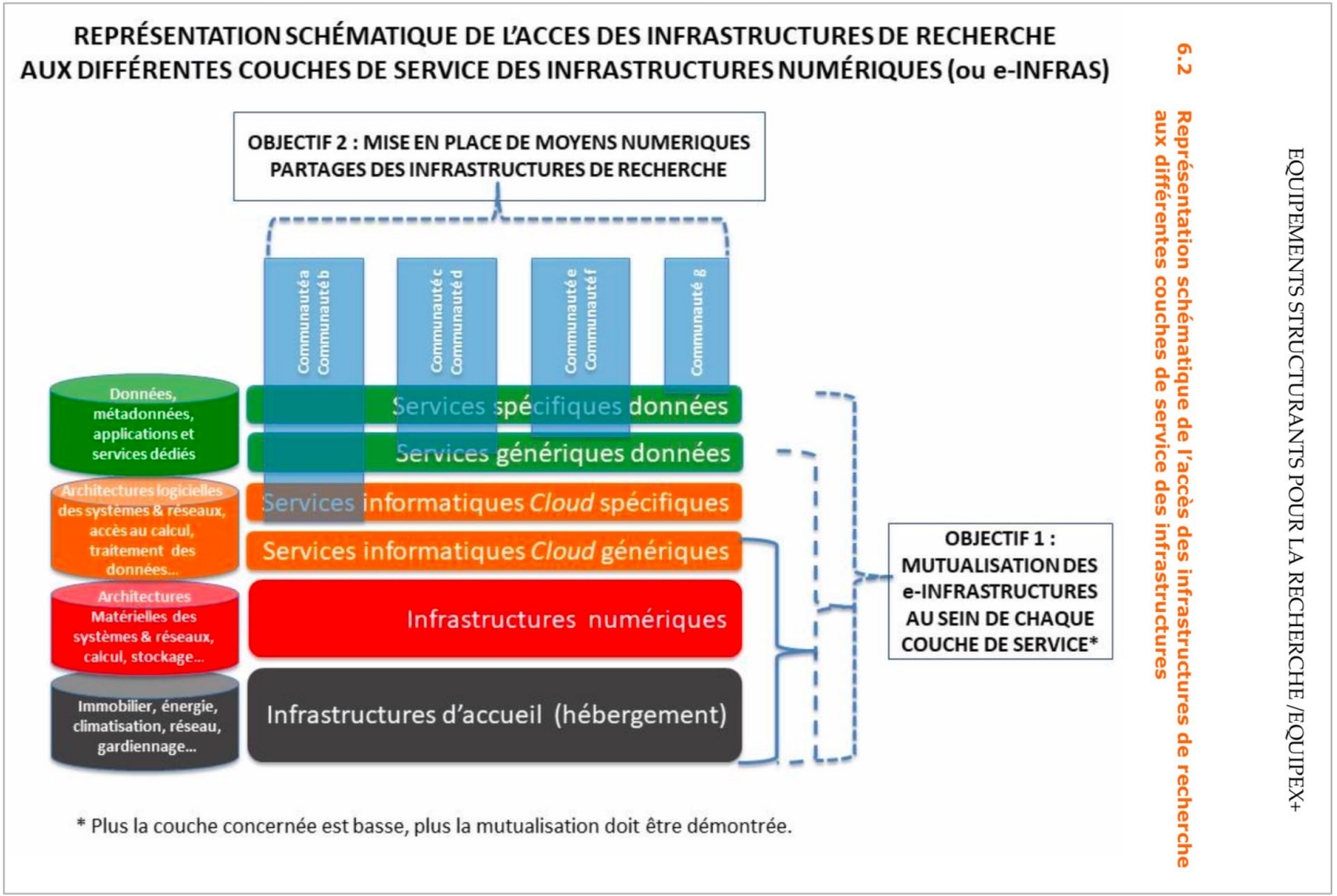
Mutualised Digital Space for Life Sciences (MuDiS4LS)



Récapitulatif total des demandes financières par destination

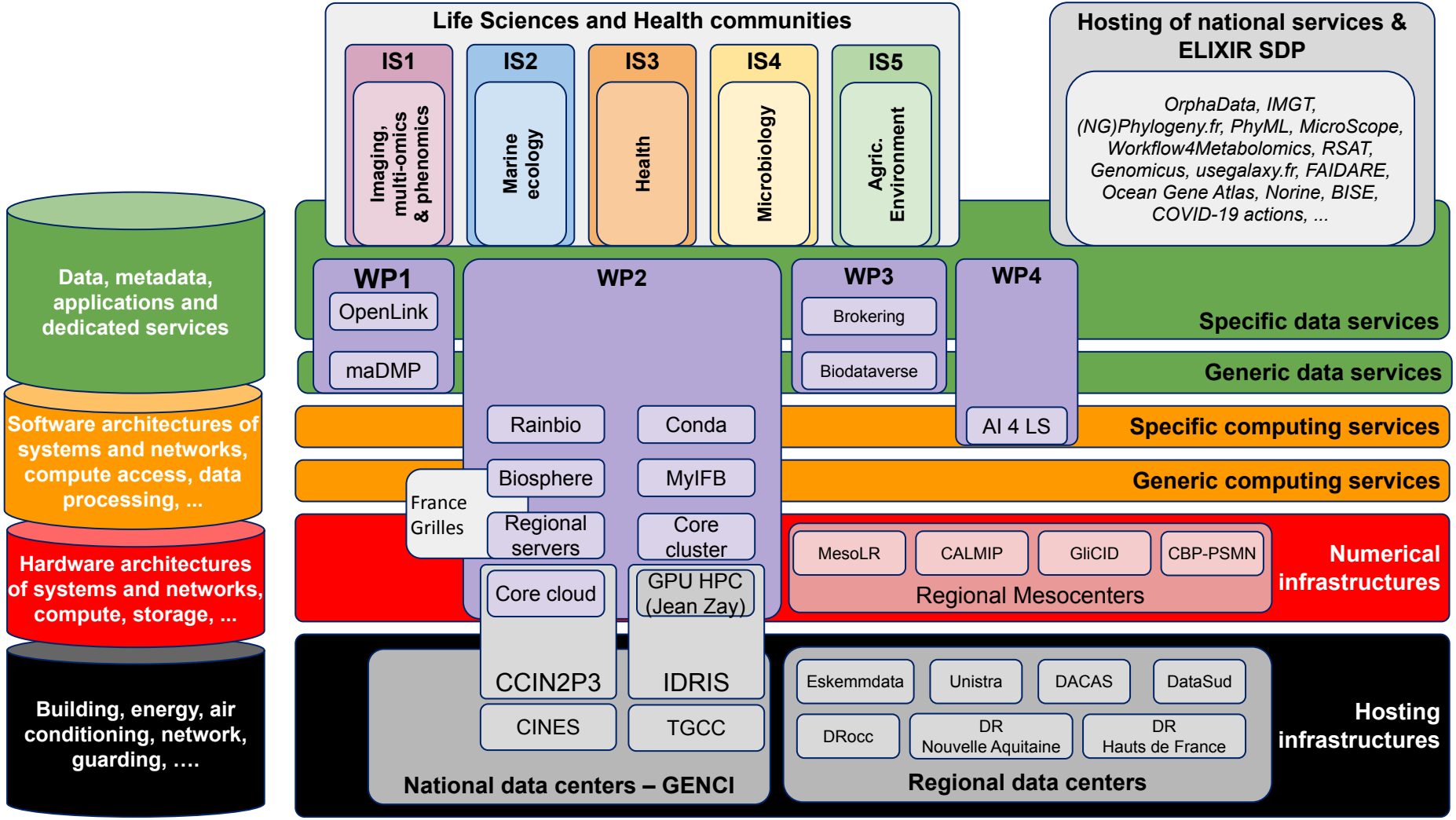
Description	Coût total	Aide demandée	Apport
Equipement	13 704 588.20 €	13 704 588.20 €	0.00 €
Personnel	14 285 017.04 €	1 278 609.60 €	13 006 407.45 €
Fonctionnement	3 533 523.90 €	3 533 523.90 €	0.00 €
Facturation interne	0.00 €	0.00 €	0.00 €
Frais de gestion	1 481 337.74 €	1 481 337.74 €	
Frais d'environnement	8 942 244.16 €		8 942 244.16 €
Total	**41 946 711.04 €**	**19 998 059.44 €**	**21 948 651.61 €**

E-infra layers



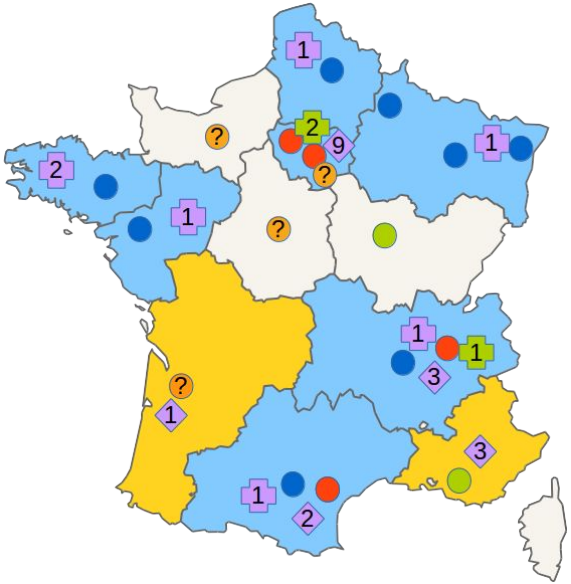
6.2 Représentation schématique de l'accès des infrastructures de recherche aux différentes couches de service des infrastructures

EQUIPEMENTS STRUCTURANTS POUR LA RECHERCHE /EQUIPEX+

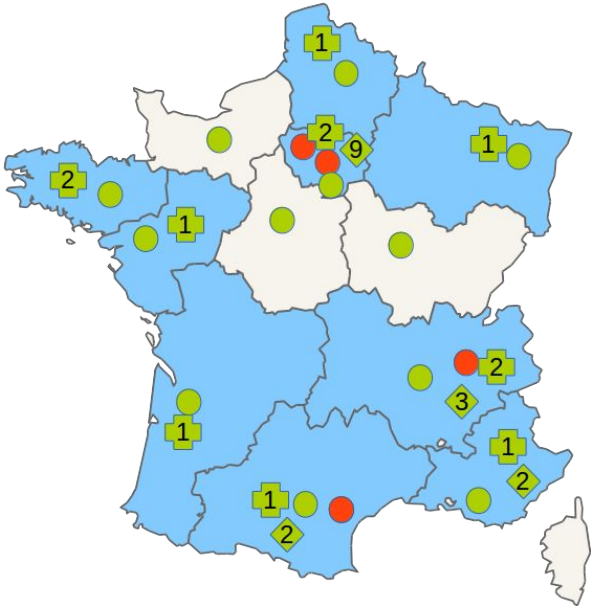

Life Sciences and Health communities
IS1
Imaging, multi-omics & phenomics
IS2
Marine ecology
IS3
Health
IS4
Microbiology
IS5
Agric. Environment
Hosting of national services & ELIXIR SDP
OrphaData, IMGT, (NG)Phylogeny.fr, PhyML, MicroScope, Workflow4Metabolomics, RSAT, Genomicus, usegalaxy.fr, FAIDARE, Ocean Gene Atlas, Norine, BISE, COVID-19 actions, ...
WP1
OpenLink
maDMP
WP2
WP3
Brokering
Biodataverse
WP4
AI 4 LS
Specific data services
Generic data services
Specific computing services
Generic computing services
Rainbio
Conda
Biosphere
MyIFB
France Grilles
Regional servers
Core cluster
Core cloud
GPU HPC (Jean Zay)
MesoLR
CALMIP
GliCID
CBP-PSMN
Regional Mesocenters
Numerical infrastructures
CCIN2P3
IDRIS
CINES
TGCC
National data centers – GENCI
Eskemmdata
Unistra
DACAS
DataSud
DRocc
DR Nouvelle Aquitaine
DR Hauts de France
Regional data centers
Hosting infrastructures
Data, metadata, applications and dedicated services
Software architectures of systems and networks, compute access, data processing, ...
Hardware architectures of systems and networks, compute, storage, ...
Building, energy, air conditioning, network, guarding, ...

NNCR anchoring in national and regional infrastructures

2020 status



2025 target



Regions

- Existing NNCR nodes
- Ongoing process to join NNCR
- No IFB platform

Regional status of data centers

- MESRI-labeled national data center
- MESRI-labelled regional data center
- Ongoing labelling project
- No known data center

NNCR status of IFB platform

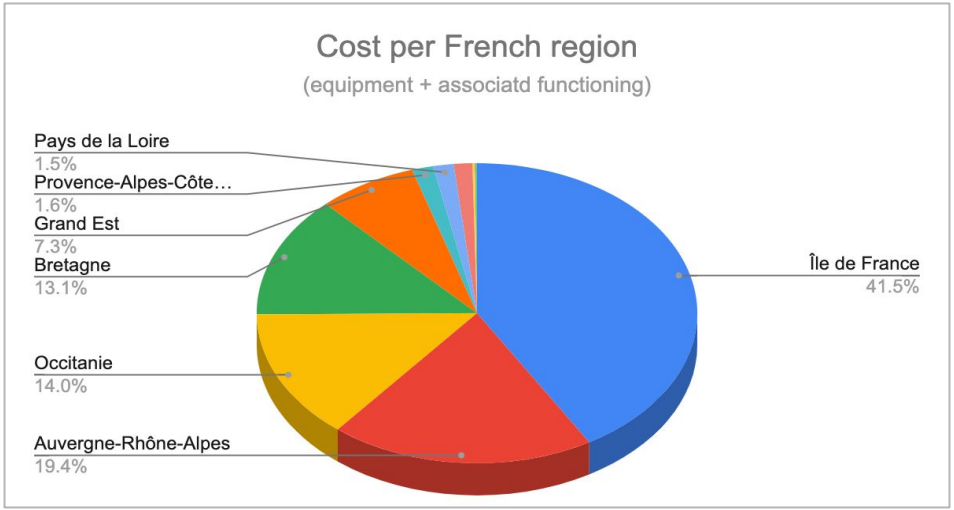
- NNCR
- Not NNCR

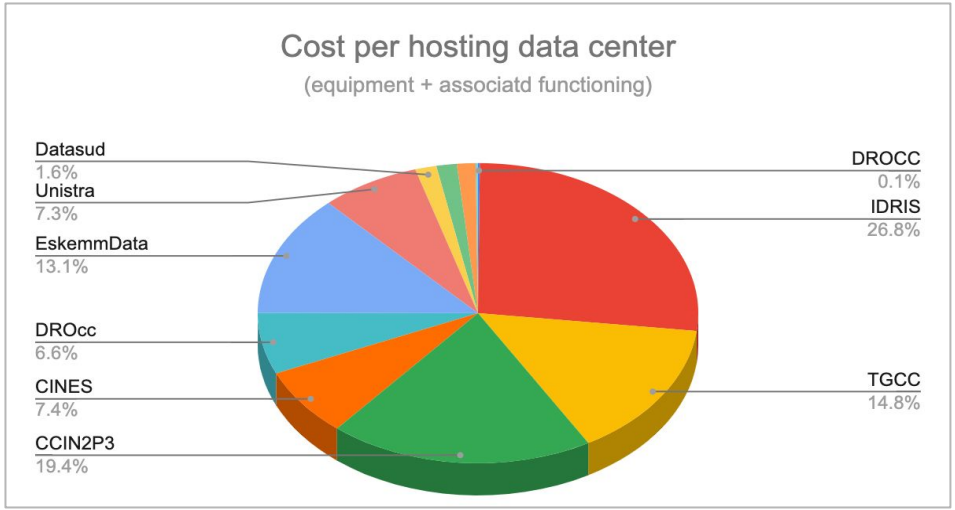
Anchoring status of IFB platforms

- Hosting in national or labelled regional center
- Self or other hosting

Regional distribution of equipment + hosting costs

Total cost (equipment + hosting costs): 17,471 k€

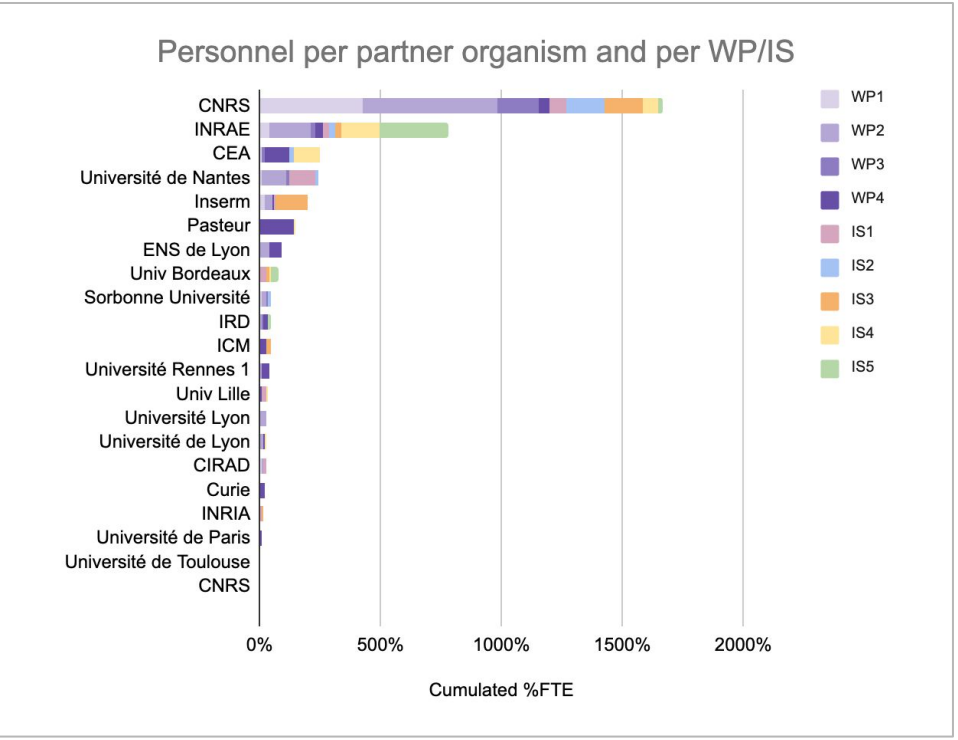


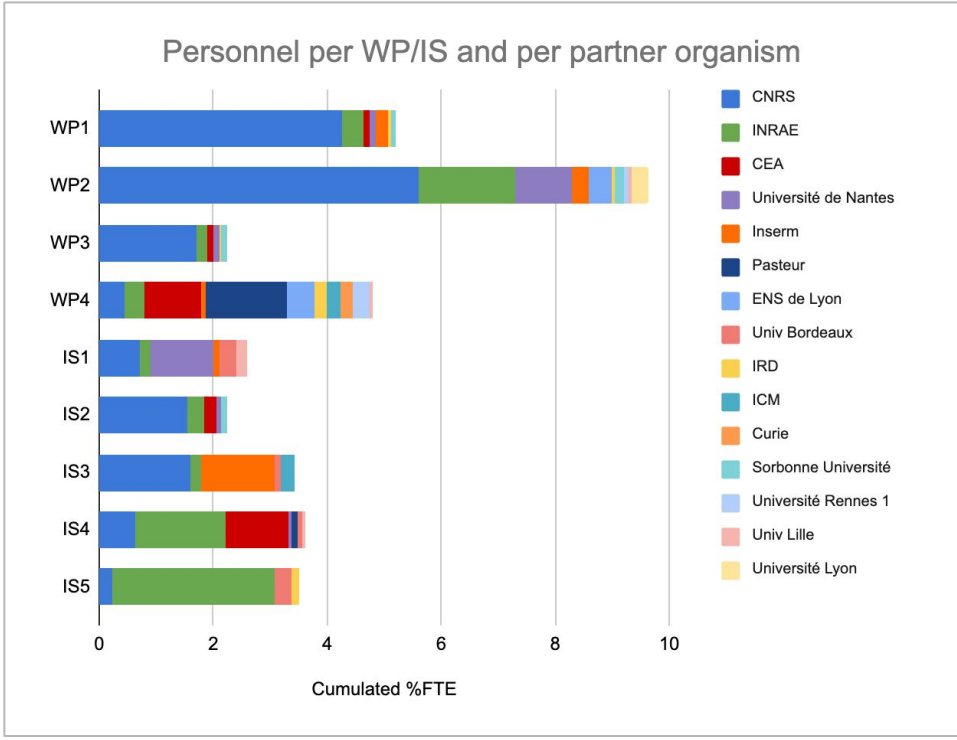


Distribution of the personnel by organism and WP/IS

Total human effort

- 170 persons
- 39 FTE
- 2999.28 person•months = 2.49 person•centuries





Répartition régionale du personnel

Total human effort

- 170 persons
- 39 FTE
- 2999.28 person•months = 2.49 person•centuries