PIA3 / Appel à manifestations d'intérêt ESR-EquipEx+
/ Call for proposals ESR-EquipEx+

2020

MuDiS4LS

Document de soumission / *Submission form*

Acronyme du projet / *Project acronym*	MuDiS4LS
Titre du projet en français	Espaces numériques mutualisés pour des données FAIR en biologie-santé
Project title in English	Mutualised Digital Spaces for FAIR data in Life and Health Science
Project manager	Scientific manager: **Jacques van Helden** (Professeur AMU, co-direction IFB) Technical managers: **Julien Seiler** (Ingénieur de recherche IGBMC CNRS, co-responsable Core Cluster IFB) and **Gildas Le Corguillé** (Ingénieur d'étude, Station Biologique de Roscoff, Sorbonne Université, co-responsable Core Cluster IFB)
Requested funding	**19,996,388.02 €** TVA non récupérable incluse
Leading institution	Centre National de la Recherche Scientifique (CNRS) EPST
Institution managing the fundings (see definitions here after), to be completed if different from the project leading institution	
Axe / *Axis*	☑ **Axe 1 : numérique** ☐ Axe 2 : générique
Champ(s) scientifique(s) du projet / *Scientific field(s) of the project*	☐ Sciences de la Matière et de l'Energie ☐ Sciences du Système Terre-Univers-Environnement ☑ **Sciences de la Vie et de la Santé** ☐ Sciences du Numérique et Mathématiques ☐ Sciences Sociales et Humanités
Ce projet est-il la suite, pour tout ou partie, d'un (ou plusieurs) projet financé dans le cadre du PIA ?	☐ Non ☑ **Oui :** IFB, Institut Français de Bioinformatique (PIA2)
Ce projet est-il partie prenante d'un projet d'Idex/Isite ?	☑ **Non** ☐ Oui :

Liste des établissements partenaires (voir définition ci-après) / *List of partner institutions (see definition hereafter)*

Nom de l'établissement d'enseignement supérieur / *Name of academic institution*	**Statut / *Legal status***
Université de Bordeaux	EPSCP
Université de Lille	EPSCP
Université de Nantes	EPSCP
Université de Lyon	EPSCP
Sorbonne Université	EPSCP
Université de Rennes 1	EPSCP
Université de Paris	EPSCP
Centre Informatique National de l'Enseignement Supérieur (CINES)	EPA
Nom de l'organisme de recherche / *Name of research organisation*	**Statut / *Legal status***
CNRS, Centre National de la Recherche Scientifique	EPST
INRAE, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement	EPST
Inserm, Institut National de la Santé et de la Recherche Médicale	EPST
CEA, Commissariat à l'Energie Atomique et aux Énergies Alternatives	EPIC
IRD, Institut de Recherche pour le Développement	EPST
Institut Pasteur	Fondation de recherche
Institut Curie	Fondation Privée reconnue d'Utilité Publique
Institut du Cerveau	Fondation Privée reconnue d'Utilité Publique
CIRAD	EPIC

Table of contents

Abbreviations

CPER	Contrats du Plan Etat-Régions	Regional - National co-funding of infrastructures
CPU		Central Processing Unit
EOSC		European Open Science Cloud
FAIR		Findable, Accessible, Interoperable, Reusable
GPU		Graphical Processing Unit
HDS	Certification Hébergement de Données Santé	French regulation for Health Data Hosting
HDH		Health Data Hub project
HPC		High-Performance Computing
IFB	Institut Français de Bioinformatique	French Bioinformatics Institute
INBS	Infrastructures Nationales Biologie-Santé	National Infrastructures for Life Sciences and Health
(ma)DMP	Plan de Gestion des Données (automatisable)	(machine-actionable) Data Management Plan
NNCR		National Network for Computing Resources
PIA	Programme d’investissements d’avenir	Funding for national infrastructures

SUMMARY

Context. Almost all the domains of life sciences rely on a diversity of high-throughput technologies that produce exponentially growing amounts of data, which need to be stored, processed, transferred, integrated, secured, properly annotated with metadata and made available within repositories in order to operate a concrete change towards open science. Since 2012, the "Programme d'Investissements d'Avenir" funding has provided life scientists with a diversity of cutting-edge data-producing technologies (sequencing, proteomics, metabolomics, imaging, 3D structure). The Institut Français de Bioinformatique (IFB) was commissioned to develop a national network of bioinformatics resources to support life and health sciences, and to deploy innovative resources to address the challenges of integrative bioinformatics. IFB federates 30 facilities ensured by 20 member and 10 associated teams ("platforms of services" in the French context) that provide well-structured and recognized services including compute and storage facilities, methodological developments, software implementation, software packaging and deployment, virtualisation, database design and curation, systems interoperability, scientific support, training, and innovative approaches for integrative bioinformatics. IFB covers all the domains of expertise in bioinformatics as well as its application to different scientific and technological sectors: fundamental biology, agriculture, health, environment. As the national bioinformatics infrastructure, it stands at the cross-road between the various data types produced by high-throughput technologies. IFB is also the French node of the European Bioinformatics infrastructure ELIXIR, which federates similar services from 23 European countries.

Scientific objectives. The main goal of MuDiS4LS (Mutualized Digital Space for FAIR data in Life and Health Sciences) is to develop a framework that will rely on the national and regional data centers to enable scientists controlling the flow of biological data, from their origin (data-producing national infrastructures) to their public release in national or international repositories, while ensuring their mid-term securing during the intermediate phases of analysis and exploitation.

The basis of MuDiS4LS will be the **National Network of Computing Resources** (**NNCR**) launched by IFB in 2017 with a mutualized task force contributed by permanent staff from 9 platforms located in 6 regions of France. Our objective is for the 20 member platforms of the IFB to collaborate in the operation of the NNCR, making it possible to extend its scope of action to 9 regions of France.

The **MuDiS4LS** project revolves around 5 strategic axes:

1. **Promoting Open Science,** by developing resources to make life and health science data and software **Findable, Accessible, Interoperable and Reusable** (**FAIR**). This will include drafting the Data Management Plans (**DMPs**) at the initial conception of research projects, and evolving these DMPs from a static document to a programmatically accessible tool, that will be executed (**machine-actionable DMP**) in order to streamline data fluxes between different hosting places (data production, fast storage during the analysis phase, mid-term securing, long-term deposits). The DMP will also be articulated with the **electronic laboratory book**, which provides life scientists with a user-friendly, powerful interface to log their research, from experiments at bench to computational analyses. We will deploy **data brokering** services to help life scientists depositing their data in national and international repositories

in well-standardized formats. Besides, IFB will also foster the development of a **FAIR culture** by training life scientists to access, share and use science data.

2. **Streamlining a secured, national computational infrastructure** by extending the NNCR to all the regions where IFB has platforms, and anchoring it in the labeled data centers, where every newly acquired equipment will be installed. We will also develop synergies and collaborative projects with two national computing centers to address the needs of specific communities (e.g. hosting of health and personal genomics data at CINES, supercomputing for Artificial Intelligence at IDRIS).
3. **Mutualizing data spaces** on NNCR nodes that will collect data produced by other national infrastructures, enabling integrative analyses.
4. **Strengthening the mutualization of human resources and expertises** by extending the inter-platform task forces already developed by IFB since 2017.
5. Reinforcing the involvement of France in **international projects** and taking a leadership position in the European Open Science Cloud (**EOSC**) by building on these consolidated national resources.

Implementation. The project will be implemented in four technological work packages (Orchestrating data flows for life sciences, Data life-long secured storage and backup, Data access and dissemination, Access to supercomputing resources).

Targeting life sciences communities. The adoption of the MuDiS4LS computing resources by the French life science communities will be catalyzed by 5 implementation studies that will address strategic research fields and will serve as starting points to launch innovative projects in those domains:

1. IS1. Integration and FAIR sharing of imaging and multi-omics data.
2. IS2. Marine biology data integration and dissemination.
3. IS3. Bioinformatics solutions to handle health data.
4. IS4. FAIR integration and sharing of new data deluge in microbiome research.
5. IS5. Integration and FAIR sharing of genetic and multi-omics data for agriculture.

Importantly, these implementation studies will drive the technological choices of the work packages, and provide realistic study cases to evaluate the adequacy of the infrastructure to address the end-user needs.

1. Project description

1.1. Scientific and technological scope of the Programme

1.1.1. Motivation

Most domains of life sciences rely on a diversity of high-throughput technologies producing exponentially growing amounts of data, which need to be properly collected, annotated, integrated, processed, secured and shared. The ability to harness this new data flow, which is essential for a transition to open science, critically depends on our ability to deploy the appropriate hardware infrastructure.

Life sciences are suffering from a crisis of reproducibility, and studies show that a significant proportion of the published results cannot be replicated by other groups (Baker, 2016). The main issue in this crisis is the accessibility and quality of research data, protocols and methods. The Covid-19 publications crisis (Mehra *et al.*, 2020) is emblematic of how non-reproducible publications can undermine public confidence in research and demonstrates that producing and sharing reliable data is of paramount necessity.

Addressing these challenges requires the establishment of an efficient organisation dedicated to the management and processing of research data. Thanks to its combined expertise in data science and information technology and its tight connections with French life sciences communities, the IFB has essential assets to design and deploy such an organisation.

Integrative bioinformatics is taking a pivotal role, by allowing researchers to make sense of massive amounts of heterogeneous data. Imaging and Next Generation Sequencing are particularly greedy in storage resources. Data produced by other technologies require less important storage spaces but their integration also crucially depends on the ability to share analysis environments and to ensure interoperability between heterogeneous data types. Moreover, making sense of the data strongly depends on the quality of annotations (metadata), the persistence of identifiers, and on all conditions defined in the FAIR principles (Wilkinson *et al.*, 2016). Based on the previous experience acquired through IFB pilot projects in integrative bioinformatics, which involved 17 national research infrastructures and national cohorts, the present project will address the integrative bioinformatics challenge. Thus, we aim at providing life sciences and health research communities with mutualized services, supported by infrastructures tailored to cope with the needs in computation, storage, security, transfer, and management of digital data, from their production to their storage in public repositories, including analysis, valorization and dissemination. National and regional infrastructures will thus be continuously upgraded to fit the evolution of life sciences and confront the tremendous increase of the demand for computing, storage and data transfer.

The project also aims at engaging life scientists to adopt information technologies for their own research laying the groundwork for operational open science, and fostering innovation.

The implementation of the project will be driven by selected use cases in health, environment, agronomy and microbial biotechnologies, which will serve as starting points to launch several innovative projects in those strategic research domains.

Overall, MuDiS4LS aims at setting up a high-performance multi-site infrastructure that will enable life scientists to address challenges of integrative bioinformatics and accelerate the national transformation towards digital economy. Importantly, the implementation will benefit from strong links between IFB and "wet" biological research labs as well as *in silico* research labs working on methods and algorithms in computational biology and information technologies.

1.1.2. Foundations of the MuDiS4LS project

IFB occupies a strategic position at the crossroads of biological data: given its multi-institutional membership (CNRS, INRAE, Inserm, CEA, INRIA) and its additional partners (Universities, Pasteur, CIRAD, IRD and several medical institutes), its target communities include research teams from basically all the French communities in life sciences and health.

Continuation and deepening of the orientations from IFB 2018-2021 roadmap

During the years 2018-2021 we have developed a framework of shared material and human resources which laid down the founding elements of the present project.

- The **National Network of Computing Resources (NNCR)** currently federates 9 platforms to provide access to high-performance servers available in cluster and/or cloud mode.
- **Mutualized task forces** were established as early as 2017 to ensure the collective management of the development and deployment of the NNCR, by sharing part-time engineers based on different platforms.
- We built **software environments best suited for open science.** To this aim, advanced methodologies (continuous Integration, code versioning, functional testing) and deployment technologies (Ansible, Conda, containers) were used, allowing the installation of isolated, modular and versatile software environments.
- **Open services to life sciences and health communities.** IFB platforms ensure consulting and support at all the steps of research projects: experimental design, primary analysis, custom workflows, interpretation of the results. They also organize **training** for biologists and bioinformaticians in order to disseminate skills and best practices to all communities.
- **Collaborations with the other national infrastructures for life sciences and health (INBS)**. Since its 2018-2021 roadmap, the IFB has established privileged partnerships with most of the data-producing INBSs, in the context of pilot-projects in integrative bioinformatics, and of a specific action "Mutualized resources between research infrastructures".

Data management throughout the life cycle.

Since March 2020, IFB initiated a project for the development of machine-actionable Data Management Plans for Life Sciences (**maDMP4LS**), which help managing the allocation of digital resources on the IFB servers (calculation, storage) through a programmatic access to the data management plans deposited in DMP OPIDoR at INIST[1]. This will lead researchers to design the data path throughout the research lifecycle. It will also change the DMP into a dynamic, adaptive document that will be updated over the course of research projects to request additional resources as soon as the evolution of the project requires it.

[1] https://www.inist.fr

Data FAIRification

Our previous achievements placed IFB at a strategic position to implement all the requirements to make the data FAIR (Garcia *et al.*, 2020; Lamprecht *et al.*, 2020). This includes the automated DMPs mentioned above as well as software best practices (Lamprecht *et al.*, 2020) (open code, version management, application of unit and functional tests, automated deployments, etc.), formal documentation of procedures with reusable workflows, development of resources for data and software interoperability (at national and European level), development of the EDAM ontology[2] (Ison *et al.*, 2013) (biological data types and formats, bioinformatics operations, topics), training and support to users in their implementation of data management plans (training of biologists and bioinformaticians, collaboration with INBSs, automation).

In this framework, MuDiS4LS will benefit from the integration of IFB within the ELIXIR bioinformatics network (ESFRI), which develops and disseminates good practices for open science (FAIR) and provides an entry point for European data-related projects: EOSC, ELIXIR-CONVERGE, Beyond 1 Million Genomes, Machine-Learning for Health, etc. In particular, IFB is now strongly integrated in the ELIXIR-CONVERGE project which aims at developing an European network of experts and resources for data management.

1.1.3. Scientific and technological objectives

The main objective of the MuDiS4LS project is to develop a framework of storage and computing resources anchored in the national and regional data centers, enabling scientists to orchestrate data flows throughout their project: elaboration, data production (by the INBS), "hot" storage during the exploitation phase, mid-term and securing long-term, deposition in national or international repositories. This orchestration, primarily led by the end-users, will serve institutional stakeholders (funding bodies, research organisms, national infrastructures, providers, etc.) to undertake virtuous data governance, by enabling to monitor the use of computing resources and to anticipate the evolution of the needs of life sciences communities. This objective will be achieved via three specific goals described hereafter.

Goal 1. Consolidating the services to researchers in life and health sciences

Transform data management plans (DMPs) from a static document to a living tool, powered by a programmatic access (**machine-actionable DMP**, maDMP), systematically invoked at the main steps of a research project: allocation of numerical resources, definition of access rights, data transfers between the sites of production, analysis, preservation and deposition. This evolution will be led in synergy with the international efforts undertaken by our European ELIXIR[3] partners in the context of the ELIXIR-CONVERGE project[4].

The use of maDMPs to **smoothen the data flows** between the different stages of the data life cycle will **engage life science communities to adopt FAIR principles** (Findable, Accessible, Interoperable and Reusable) at each step of their projects.

[2] http://edamontology.org/

[3] https://elixir-europe.org/

[4] https://elixir-europe.org/about-us/how-funded/eu-projects/converge

Addressing the needs at the national level implies significant evolution of the NNCR equipments, in synergy with the "wet" and numerical equipment managed by the data-producing national infrastructures. This evolution implies to (i) mutualize storage and computing resources between national infrastructures; (ii) share human resources, skills and practices through the implementation of use cases; (iii) extend the current NNCR by involving additional platforms and regions. At the European level, this goal is in full harmony with the active involvement of IFB in the ELIXIR-CONVERGE project but also by establishing deposition procedures in international repositories (data brokering, see Goal 3).

Goal 2. Securing and rationalizing the equipment by anchoring the NNCR in labeled data centers

Rationalizing the compute and storage resources. Figure 1 summarizes the current state of IFB numerical resources (left side) and the 2025 target organisation (right side). The infrastructure already proposes both cluster and cloud services provided by 10 IFB platforms covering 7 regions. Four of these computing facilities are already hosted in national data centers, in particular IFB-core-cluster at IDRIS and IFB-core-cloud at CCIN2P3, and 2 platforms already collaborate with regional mesocenters. The NNCR developed during IFB 2018-2021 roadmap currently federates 8 platforms covering 7 regions (highlighted in blue on the left panel).

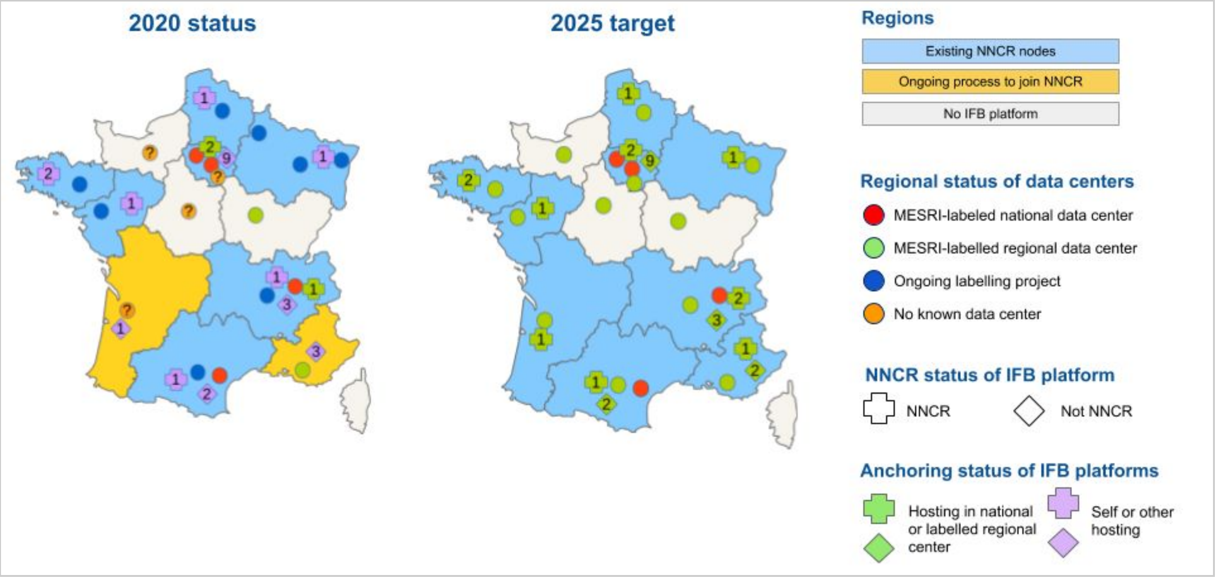

2020 status
2025 target
Regions
Existing NNCR nodes
Ongoing process to join NNCR
No IFB platform
Regional status of data centers
MESRI-labeled national data center
MESRI-labelled regional data center
Ongoing labelling project
No known data center
NNCR status of IFB platform
NNCR
Not NNCR
Anchoring status of IFB platforms
Hosting in national or labelled regional center
Self or other hosting

Figure 1. Current status (left) and 2025 projection (right) of the IFB National Network of Computing Resources (NNCR).

This proposal aims at (i) a progressive migration of all IFB servers towards labeled data centers; (ii) the development of synergies with regional mesocenters and national data centers; (iii) an extension of the NNCR to all the regions covered by IFB platforms; (iv) the use of this extended infrastructure to foster the development of new services addressing the needs of user communities.

At the end of the installation phase (M48 of this project), the NNCR will encompass 17 platforms covering 9 French regions (Figure 1, right side). IFB platforms co-hosted in the same regional (e.g. EskemmData, DROcc) or national site (IDRIS, CCIN2P3, CINES, TGCC) will share the local equipment, which will ensure resource balancing for their respective users. This will be particularly relevant for computing resources, which are characterized by successions of intense and moderate usage.

Strengthening human means and skills by extending the national task force to the new participating platforms, and developing regional task forces which will mutualize the efforts of platforms located in the same region.

Consolidating the position of France in EOSC. Federating numerical resources at the national level is strategic in the perspective of the European Open Science Cloud (**EOSC**) construction. By strengthening the collaborations between IFB platforms and the national data centers, MuDiS4LS provides a powerful lever to enhance the French participation in EOSC. In particular, IFB already collaborates with France Grilles to deploy cloud computing services. The study cases of the MuDiS4LS project will develop new skills in Artificial Intelligence applied to different domains of life sciences, which will be key elements to apply for an increasing number of EU funding calls.

Goal 3. Supporting the management and exploitation of the data produced by the other national infrastructures for life sciences and health

Rationalizing the use of numerical resources by hosting on NNCR servers a part of the data produced by national infrastructures for life and health sciences (INBS). In parallel, IFB will also carry on and strengthen its offer to end-users (shared spaces for team projects) within the regional and national platforms. A case study will also be developed in collaboration with France Génomique to assess the impact of centralized (at the TGCC national data center) versus local hosting solutions (in regional centers, close to the data production centers and research projects). By securing data storage for all researchers, IFB aims to reduce the need of keeping large data storage facilities on data production platforms or research units and thus limit costly and uncontrolled data replication. In addition, specific monitoring of the execution of the Data Management Plans will ensure that the data is stored any time at the planned location.

IFB is currently starting a collaboration with European Nucleotide Archive (ENA), the international sequence repository hosted by the European Bioinformatics Institute (EBI), to develop **data brokering services** with France Génomique, a national sequencing infrastructure already engaged in data brokering with ENA (cf. Appendix A6.2). To face the challenges of integrative bioinformatics, data brokering will be extended to support the deposition of other data types (multi-omics, imaging) in dedicated repositories. IFB will act as intermediate between end-users and various national and international repositories. This will rely on the following elements: (1) specification, from the project conception to the final destination of each dataset via the maDMP; (2) development of domain-specific procedures to validate the data and metadata before attempting to submit them; (3) preliminary deposition of the data as soon as it is produced (FAIR from the first day) with an embargo to ensure their protection during the exploratory phase; (4) training and user support.

1.1.4. Implementation of the technological objectives

The MuDiS4LS work plan has been divided into 4 technical Work Packages (cf. appendix A3).

WP 1. Orchestrating data flows for life sciences

This WP will focus on data management and stewardship for data hosted on geographically distributed IFB platforms. Since IFB acts as a hub for data originating from various sources and institutes, it is of utmost importance to provide data access coping with scientific and institutional constraints. However it is technically challenging and costly to provide rich and complete metadata required for better data access and reuse. As IFB members are in contact with many communities, IFB is in a prime position to bring about a change in the attitude of its users towards the central role of efficient research data management. The aim of this WP is (i) to lower the cost of populating DMPs through FAIRifying the data from the first day and (ii) to provide incentives for researchers to produce and share qualified data for open and reproducible sciences.

Specific tasks:

1. Developing procedures relying on machine-actionable DMPs to enable a swift management of the data fluxes between data producing infrastructures, computing facilities, and repositories.
2. Instrumenting the data and computing infrastructure to capture metadata (e.g. provenance) and feed maDMP thus lowering the human cost of maintaining and allowing the automatic update of data management plans during research projects' lifecycle.
3. Disseminate the maDMP towards data-producing national infrastructures for life and health sciences, in order to ensure a FAIR data management from the first day.

WP 2. A Distributed data infrastructure for project-life-long secured storage and backup

This WP will lay down the physical infrastructure underlying the whole project, by setting up a distributed compute and storage infrastructure for life sciences, anchored in regional and national data centers, which will be managed by a mutualised task force regrouping members of IFB platforms and support from the mesocenters. This WP will focus on providing mid-term secure storage on all NNCR (National Network of Computing Resources) sites.

Specific tasks:

1. Rationalising the equipment of IFB federated platforms by installing all the facilities in labeled regional or national data centers.
2. Support the Core and regional NNCR nodes, by combining HPC and mid-term secured storage.
3. Expanding the services to regions not yet covered by the IFB NNCR.
4. Build a back up network between sites and within the NNCR network.
5. Create shared data spaces (data lake) enabling the integration of different data types and their access by different computing technologies in a transparent way.

WP 3. Data access and outreach

This WP will deal with the final destination of the data, be it an international or local repository. IFB is recognized as a data hub for Life Sciences Data by CNRS-INSB and ELIXIR. CNRS-INSB has mandated IFB for the setup of a Dataverse repository. ELIXIR has solicited IFB to act as data broker for ELIXIR deposition databases. These two missions are intertwined and will help IFB to develop a strong connection between data conservation and data publication in international repositories. IFB will endorse a strategic role in research data management enabling biologists to adopt FAIR principles and take a big step toward Open Science.

Specific tasks:

1. Create a national repository (BioDataVerse) that will strengthen and complement existing institutional repositories.

2. Liaise and interact with ELIXIR Deposition Databases services for the data brokering.
3. Creation of a permanent role of data coordinator as interface for the thematic communities.
4. Host and run thematic communities that will put together their expertises to enable seamless curation and validation of datasets for a wide variety of life-science and health domains.
5. Connect with meta-data portals in order to give visibility across the world to all datasets hosted by IFB and partners.

WP 4. Intensive Computational Biology (Access to national HPC/AI resources)

The work package "Intensive computational biology" aims at enabling the life science communities to use for projects with intensive and Artificial Intelligence (AI)-related computing needs the existing intensive computing resources (HPC, AI, Bigmem) available in the four national centers IDRIS, TGCC, CINES (both affiliated to GENCI) and CC-IN2P3. The challenges are to deploy on these national facilities the usual tools and data in life science, to provide users with common workflow environments and scientific gateways and web portals. The use cases include AI projects, and projects requiring very large computing resources, such as health applications related for instance to COVID-19 or large scale microbial genomes analysis. Running applications for intensive computations requires that tools and pipelines have been adapted and benchmarked at a preliminary step in representative regional HPC/AI environments (e.g. CBP-PSMN).

Specific tasks:

1. Deploy on the national HPC/IA resources the required tools and databases.
2. Deploy common life science workflow engines in HPC facilities (nextflow, snakemake, CWL).
3. Provide application developers with benchmarking environments
4. Share data between HPC/IA and IFB resources (WP2 and project FITS) using e.g. iRODS tools.
5. Evaluate integration of HPC/IA resources with public scientific gateways of IFB.
6. Train developers and users to HPC/IA computational environments.

1.2. Structure and building of the equipment

1.2.1. Element 1: NNCR Core Resources

In its 2018-2021 road map, IFB chose to deploy a national infrastructure proposing two complementary access modes: IFB core cluster, hosted by IDRIS (national data center in Paris region) and IFB core cloud, hosted by CCIN2P3 (national data center in Lyon). This dual infrastructure, funded by the previous round of the PIA, is operated by mutualized task forces contributed by regional platforms. The core resources also provide a shared solution for users from the French regions where IFB has no local implantation (white areas on **Figure 1**). They also provide an answer to occasional overload of some regional nodes and can serve as a buffer to ensure an efficient balance of the infrastructure load at the national level.

The investments requested for this funding call will enable to strengthen and ensure the sustainability of the Core infrastructures, with a specific focus on storage, which is currently insufficient to propose mid-term secured hosting to users. Our current estimation of the needs indicates that the storage and compute capacity of the Core resources should be doubled over the next 8 years. Concomitantly, the infrastructure should acquire GPU nodes for specific applications of life sciences: genomics, structural biology, imaging. These resources will also be dimensioned to enable prototyping before being put in production on the Jean Zay national GPU supercomputer. The deployment of GPU-based software and workflows on Jean Zay will be led in collaboration with the IDRIS partner.

The storage infrastructure will be based on technologies already mastered by the IFB task force. Indeed, the IFB Core Cluster was equipped mid-2020 with a new storage infrastructure based on DDN hardware solutions and Open Source Lustre Filesystem. This investment was followed by the engineers of 5 IFB platforms and allowed them to get acquainted with this storage solution.

1.2.2. Element 2: Regional platforms of the NNCR

The MuDiS4LS project takes ground in a long-lasting collaboration between French bioinformatics platforms started with the ReNABi (Réseau National de Bioinformatique, 2008-2013) and pursued with the PIA2-funded Institut Français de Bioinformatique (2013-2019, extended with extra-funding to 2025). These initiatives led the regional platforms of bioinformatics to federate their compute and storage facilities by forming IFB National Network of Computing Resources (**NNCR**).

In the installation phase of the MuDiS4LS project, following the example of NNCR-core infrastructures, regional platforms will scale their infrastructure by adding computing power (CPU, GPU), high-performance storage for computing-intensive tasks and mass storage to host project data. In order to rationalize and secure the digital equipment financed by MuDiS4LS, the regional platforms will rely on the labelled data centers. In regions where several platforms co-exist, their respective infrastructures will be grouped together in the same data center, which will lead to savings in equipment and human resources.

1.2.3. Element 3: Inter-site data replication

In order to reduce the global data footprint, and to simplify their management by the community, it is mandatory that they are organized within stable and trusted infrastructures. For this purpose, each infrastructure must be able to store its data in a secure way. Thus, IFB will use the NNCR as a support to organize inter-site data replication to ensure in each infrastructure the secure storage for each project. This replication network will use the high-speed connexions between the different labelized datacenters (RENATER), the harmonization of the infrastructure management system, and the mutual competencies from NNCR for its perennial behaviour. Each infrastructure will thus have access to one or more regional sites to host its replications. The site choice could change depending on security prerequisites for each hosted project. Backup systems and mechanisms will use the primary storage capacities from infrastructures. This approach will help users to store data only in the place where they will be requested at each stage of a project (acquisition on a storage facility close to the data providing equipment, computation on high performance storage on HPC, and long-term storage on highly replicated facility). Such mechanisms will help to provide national or international labels to different services or databases (e.g. CoreTrustSeal).

1.2.4. Element 4: Health data hosting and secured research environments

Similarly to the general data, sensitive data should follow FAIR principles and similar good practices. Within the framework of national regulation (RGPD) it should become possible to share metadata in order to identify existing data sets, their owner and the associated research topics. Since storage and processing of such data cannot be operated on the same infrastructure as non-sensitive data, it will be hosted by data centers certified for health data hosting (HDS) such as CINES. Secured spaces, composed of storage space and computing resources, will be implemented relying on virtualization techniques; this will allow research teams to instantiate their tools and apply their processes in state-of-the-art and HDS compliant environments. The storage of sensitive data will rely on two types of technologies: classical storage (e.g. to operate non-sensitive databases without performance degradation), and object storage similar to non-HDS with the same principle of inter-site replication to guarantee data resilience and security while enabling them to be processed by HDS certified centres. To operate at best all the national resources, we will develop data export mechanisms paired to pseudo-anonymization systems, data watermarking, and we will deploy, in full compliance with the regulations, a secured file exchange platform enabling to process some of the data on non-HDS infrastructures. The design and integration of this HDS environment will serve as a pilot that will enable us to overcome technical difficulties related to partitioning and securing environments. It will serve as a basis to formalize and share various deliverables (including process, operating mode, installation scripts and architecture document) on which the research community will be able to capitalize for the design and implementation of HDS compliant technical environments, and for future HDS certification campaigns.

1.2.5. Element 5: BioDataVerse

To initiate a coherent network of data repositories that will support researchers in the referencing and publication of their data sets, a repository will be created to give access to storage resources for data description, sharing, preserving and citing. This repository, called **BioDataVerse**, will be operated by the core node at the IDRIS national centre. It will also serve as the support of the data brokering mechanisms that IFB wishes to implement in order to support researchers in the referencing and publication of their data sets on data repositories of international scope such as ENA. The BioDataVerse repository will be implemented by using the Dataverse open source web application. It will be replicated on at least one separate site. The BioDataverse will join the BRIDGES cross-repository portal (https://www6.inrae.fr/bridge/) currently implemented by IRD, INRAe and CIRAD.

1.2.6. Element 6: Ease the access to the Jean Zay Supercomputing facility

IFB will develop strong collaborations with GENCI-IDRIS related to HPC/AI activities. IFB and GENCI-IDRIS will join their efforts and share their expertise and teams to train and assist life scientists and engineers to deploy large scale analyses on the national HPC/AI facilities. One of the common actions will be to ease the access of the life science community to the Jean Zay Supercomputing facility, where IFB members can have in a few days access to a pool of resources and to benefit from support teams, especially with the existing GENCI's dynamic access to IA resources.

1.2.7. Synthesis of the equipment for Element 1, 2, 3, 4 and 5

Table 1. Total requirements for Element 1, 2, 3, 4 and 5 (core and regional servers). Full version is in **Appendix A5.2**

Element	Nb CPU blades (112 cores)	Nb GPU blades (4 cards)	Nb Fast-access storage arrays (2Po)	Nb Mass storage arrays (2Po)	Nb of backup storage array (2Po)
1&3 - Equipment for NNCR core resources	51.17	12.54	0.26	2.09	1.51
2&3 - Equipment for regional IFB platforms	146.02	21.78	1.18	7.08	1.90
4 - Health data hosting and secured research environments	2.64	0.00	0.00	0.66	0.00
5 - BioDataVerse	0.00	0.00	0.00	0.33	0.33
TOTAL	**199.83**	**34.32**	**1.44**	**10.16**	**3.74**

Table 2. Localisation of the equipment per Element. Bold: new infrastructures; regular: extension of existing ones. A detailed version with computing and storage capacities per region and site is provided in **Appendix A5.2**

Target labelled data centre | Elements			
CCIN2P3	1, 2, 3, **5**	DROCC	2, 3
CINES	2, 3, **4**	EskemmData	2, 3
DACAS	2, 3	IDRIS	1, 2, 3
Data Center Régional Nouvelle Aquitaine	2, 3	TGCC	2, 3
Datacentre Hauts de France	2, 3	Unistra	2, 3

1.3. Originality, innovative feature and sustainability of the equipment project

1.3.1. A comprehensive framework for life-long management of Life Science data

The first innovative feature of the project lies in the scientific data management aspects. By bridging all the steps of data life (production, analysis, valorisation, mid-term securing, long-term preservation, deposition; cf. **Figure 2**), the project will tackle, in an integrated way, recurrent problems encountered in Life Sciences resulting from several factors: ever-increasing volumes, heterogeneity of the data types, challenge of interoperability complexity of the biological concepts.

The 4 WPs described in section 1.1.4 will lay down the foundation of a general framework enabling life scientists to manage their data in a seamless way at each step of their projects.

maDMP will concretely transform data management into a dynamic process, where the DMP will enable to streamline data flows between the different sites hosting the data at its different stages (production, analysis, public opening). The systematic use of maDMP to handle data transfers and resource allocation will also ensure a permanent synchronisation of the DMPs with the evolution of the projects (modification of the data volumes, inclusion of new data, modification of the collaborators, etc.). In addition, provenance metadata will be stored in a FAIR centralized and accessible knowledge graph leveraging the W3C PROV Ontology (https://www.w3.org/TR/prov-o/). This graph will be linked to data production and analysis research context, provided by the maDMP.

It will result in a domain-specific knowledge resource aimed at fostering community data sharing and reuse without disclosing possibly sensitive raw datasets.

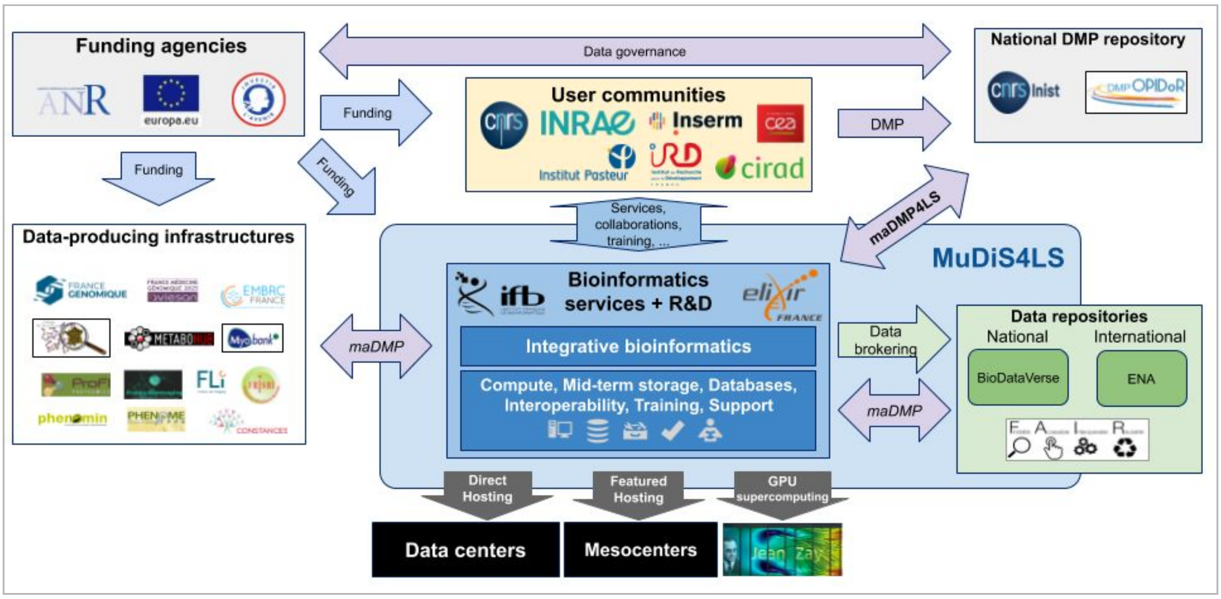


Figure 2. ***Orchestration of the data flow.*** *The project is structured around the management of data flows through machine-actionable DMPs to ensure the coordination of data fluxes at the national and international levels.*

This will transform DMPs from a static document into an operational data governance tool for all the stakeholders: data providers to monitor the usage of their resources and anticipate the future needs; funding agencies and research organisms to ensure a global audit and follow-up of the utilisation and evolution of data resources.

While DMPs allow researchers to set up a data management strategy from the first day of their projects, data warehouses represent an essential step in the data lifecycle for storing and publishing data, thereby enabling to fully embrace the open science era. INRAE, CEA and CIRAD have already started to develop institutional data warehouses, however, these initiatives are restricted to their own personnel and are far from covering all national needs, especially for biology. Given the issues of a researcher's work belonging to his or her home unit or organisation and intellectual property issues, it seems consistent that the management of the repositories should be specific to each organisation.

IFB has been mandated by the CNRS Institute of Life Sciences (INSB) to deploy a data warehouse that meets the needs of CNRS biology and health researchers. This BioDataVerse will be developed within the framework of the NNCR to ensure its implementation within a national data center. Replication on a remote site will ensure the security and durability of the data. The selected software solution will be based on the experience already acquired by INRIA, IRD, CEA and INEE in order to allow technological harmonization and pooling of skills.

The essential steps to ensure sound data management are: data curation, definition of the expected metadata for each dataset and establishment of metadata validation procedures. These steps will be implemented through the work groups of the Implementations Studies, which are representative of diverse user communities, and will be involved in the data management of their themes.

Through the publication of reusable deployment recipes via the national NNCR network, the IFB wishes to promote the creation of a network of data repositories. In compliance with FAIR principles, all the repositories thus created will be grouped together within a single meta-portal supporting cross-disciplinary research. The different nodes of this network will make it possible to adapt management rules to the needs of local research communities, particularly in terms of data conservation policy.

1.3.2. Mutualization of human resources

The founding element of IFB actions led since 2017 was the mutualisation / pooling of human resources. This applied not only to the federation of material compute and storage resources, but also to the sharing of expertise, as well as the development and deployment of software environments. A similar mutualization has been developed by the cloud federation, Biosphere, which interconnects 6 clouds hosted by IFB-core cloud and regional facilities. The targeted organization model creates a balance between data movement, which may be very costly in some cases and even impossible with confidential data, and the automatic deployment of appliances on local clouds, the closest possible to data repositories but requiring a local cloud-compatible infrastructure.

These core shared resources and mutualised task forces have enabled the IFB to set up and consolidate **procedures, good practices and prototypes** that are then adopted by the NNCR, and serve as building blocks for transnational integration. This includes for example the parallel deployment of scientific tools on several national infrastructures[5], the adaptation for IFB Core Cluster of an account manager initially developed for a regional platform, the sharing of Ansible playbooks to harmonise the instances of the NNCR. The task forces also animate the IFB community support site[6] which brings together users, system administrators and scientific experts in order to answer questions and address issues in a collective mode. This sharing of components contributes to the stabilisation of regional infrastructures and ensures sustainability by reducing the maintenance cost and by fostering the sharing of knowledge and the adoption of good practices. Of note, all these developments were born within specific platforms, and later refactorized, generalized and deployed on the IFB Core Cluster to finally be redistributed to other regional platforms.

Furthermore, since 2019, the NNCR platforms have defined a common policy for the use of their resources and have been working on the construction of a **unified economic model** in order to propose a harmonized tariff model to user communities (section 4.2.2).

1.3.3. Bringing life science communities to the digital space

The adoption of the MuDiS4LS computing and storage infrastructure by the target communities will be catalysed by 5 implementation studies that will address thematic fields representative of the French life science research (health, agriculture, environment, marine biology and microbiology). These 5 implementation studies will drive the technological choices of the 4 technological work packages described in 1.1.4, and provide realistic study cases to evaluate the relevance of the infrastructure to address the actual needs of user communities.

[5] https://gitlab.com/ifb-elixirfr/cluster/tools

[6] https://community.france-bioinformatique.fr/

A detailed description of each Implementation Study is provided in **Appendix A3**, with a definition of the goals, tasks and deliverables. The list of participating partners, limited to the ISs co-coordinators, is given in **Appendix A4.2**. We provide hereafter a condensed description of each IS:

IS1. Integration and FAIR sharing of imaging and multi-omics data. This IS brings together the main image-producing infrastructures (FBI, FLI, FRISBI, EMBRC, Phenomin), which will delegate a part of their image data to IFB computing facilities in order to enable their integration with other data types (multi-omics). These national infrastructures will develop synergies to (i) define the roadmap to create the French IDR (Image Data Repository for public archiving), (ii) manage the life cycle of imaging data by elaborating a DMP for image management structure and for scientific projects, including links to public archiving, (iii) ensure the FAIRification of data and its FAIR exposition for heterogeneous data integration and analysis, and (iv) develop specialized workflows and enable a multiscale interpretation (from atomic/protein to cells to small animals or patients) in a phenomics perspective (in collaboration with PHENOMIN).

IS2. Marine biology data integration and dissemination. The aim of this implementation study, in close connexion with the WP6 of the AO-EMBRC project, will be to enhance and extend cross-referencing of environmental descriptors, taxonomy, multi-omics, modelling and imaging data as well as analysis pipelines generated by research on marine organisms and ecosystems promoted notably by the EMBRC infrastructure and the TARA consortium. It will rely on construction of marine specific DMP in collaboration with the WP1 inspired by the work initiated in the framework of the ELIXIR marine metagenomics community. It will contribute, in collaboration with WP2, to the development of a national infrastructure to ensure regular processing and dissemination of the data produced by the marine stations and observatories. It will promote the FAIRfication of marine models and augmented observatories datas and their dissemination in national and international ecological data infrastructures (DataTerra, Emodnet Biology) on the one hand and with genomics and imaging data warehouses on the other hand (ENA, EuroBioImage).

IS3. Bioinformatics solutions to handle health data. Health data require specific storage and computing environments to ensure compliance with regulatory policies. France defined a certification to host health data (HDS, "Hébergement de données de santé"), which is so stringent that not a single academic actor is currently in state to provide end-user services. This Implementation Study will benefit from a physical, technical and human environment located at the CINES data center, which has undertaken the procedure to achieve the level 1 of this certification (security layers), and is at the crossroad of key national infrastructures (France Cohortes, France Médecine Génomique 2025). The goal of IS3 will be to implement all the intermediate layers to set up services to (i) manage, process, benchmark, host and share sensitive health data, (ii) evaluate and adapt new technological approaches from WP1 and WP4 to sensitive data (iii) provide guidelines, templates and tools for writing and implementing biomedical DMP (FAIR principles), through the adaptation to sensitive data of the Researcher Digital Environment. This IS will foster interactions with the N4HCloud digital platform for health data (integration of imaging and omics modalities, data management and processing solutions).

IS4. FAIR integration and sharing of new data deluge in microbiome research. This IS aims at setting-up a shared space for the integration of FAIR (meta)-omics (genomics, transcriptomics, metabolomics, ...) data obtained on (i) a large number of microbiome samples from the human body, animals or various environments, and (ii) libraries of bacterial genomes and their genotypic nomenclatures. The Healthy French Microbiome program (100,000 metagenomics samples) and the sequenced genomes of the Institut Pasteur libraries will be the starting input data to IS4. Its goal will be to (i) provide the guidelines and templates for implementing DMP including FAIR principles covering minimal standards, (ii) establish minimal requirements for metagenome annotation strategies in particular for antibiotic resistance genes and mobile genetic element, and (iii) deploy machine learning and Artificial Intelligence methods on the Jean Zay computer (IDRIS) for the analysis of large cohorts. Specific use cases will address the question of antibiotic resistance prediction and its evolution, and the characterization of human gut microbiomes infected by SARS viruses.

IS5. Integration and FAIR sharing of genetic and multi-omics data for agriculture. The objective of this work package is to develop services that will link biological resources to highly heterogeneous data types including 'omics' data, images, phenotypic and environmental measurements, with a specific focus on holobiont studies for agricultural animals and plants, as well as their commensal, symbiotic and pathogenic microorganisms. It will be led in collaboration with the RARe CRBs, which are at the heart of many research programs intended to explore the living organisms and ecosystems as well as valuing biodiversity for agriculture and industry, food, environment and health. It will identify holobionts of interest to several user communities (plants, animals, food and environment), rely on international standards to FAIRify existing data, rely on WP1-4 of this project to manage data and develop integrative workflows, and engage communities via dedicated training activities.

The Implementation Studies are based on the technological developments proposed in the 4 WP described in section 1.1.4. Their use cases highlight a body of transversal needs that cover data management in labeled national and/or regional data centers based on a dynamic Data Management Plan (maDMP) to ensure the dynamic life cycle of the data generated during the research project, the data FAIRification required for multi omics data integration, and the data sharing and dissemination thru the adaptation of data brokering procedures in international repositories developed in WP3. Specific tasks will also require the elaboration of DMP for the management of various kinds of data (image, NGS, diseases and patients, environmental metadata etc), and the definition of standards and ontologies based on already international initiatives.

Altogether, these implementation studies, conceived in consultation with research infrastructures, cohorts, industrial demonstrators, are representative of the main user communities of our bioinformatics resources. They will serve as starting points to launch several innovative projects in those strategic research domains, which aims to respond to the challenges of integrative bioinformatics by confronting the scientific and technological obstacles reported by users.

1.3.4. Sustainability

The sustainability of the MuDiS4LS equipment relies on the engagement of IFB research organisms, the pricing strategy and the economic model of IFB, and co-financings. The economic model is described in **section 4.2.2**.

Integration of MuDiS4LS in institutional strategies

The MuDis4LS project is strongly supported by the partner organisms, as evidenced by the opening of 7 permanent positions during the course of the project (**Section 4.2.2**). We also received 6 support letters from the research organisms members of IFB direction board (CNRS, Inserm, INRAE, CEA, INRIA) as well as from IRD (**Appendix A6**). These organisms emphasise the importance of MuDiS4LS for their own institutional strategies (through the 5 Implementation Studies). They also highlight the importance of the project's foundation: developing a framework anchored in the national and regional data centers that will enable life scientists to orchestrate their data flows.

Funding model / pricing strategy and co-financing

The full cost exercise launched by the DGRI-MESRI in 2017 provided us with a first global view of IFB costs and incomes: around 21 M€/year (2016 and 2017 exercises) including indirect costs (20%), salaries of permanent and non permanent positions (57%), functioning (13%) and equipment amortization (10%). Considering the new policies of the French ministry and research institutes, research infrastructures are requested to include project contribution as an element in the funding of the operational cost. Indeed, IFB platforms are aware of the need to control the increasing demand for storage and computing resources, and to establish pricing for services to guarantee increasing self-financing contribution. Identified incomes for the renewal and evolution of equipment include revenues from infrastructure access pricing, national and international collaborative research projects involving IFB platforms as partners, support from research institutions and support from french regions and government (this call). The envisaged priority actions are presented below:

- **Increase the rate of self-funding** to 20% of incomes by an in-depth work on the service offering. The objectives of the model are to work towards a responsible use of resources, a project driven demand and a fair distribution of storage and computational resources. The fees will be adapted according to the profile of the user (academic or from a private company, funder (equipment acquisition, donation) or not), and will tend to converge between platforms according to the type of service: (i) computational needs, beyond a basic and free bundle, units of CPU hours and storage capacities, (ii) training sessions and (iii) data analysis. For the last one, collaboration with academic non-bioinformatician scientists will be preferred and payment will be asked only for equipment and functioning needs.
- **Increase the part of IFB own contractual resources** to 5% of incomes. The links established with National and European bioinformatics infrastructures, and the broad coverage of skills that can be mobilized around ambitious projects should contribute to attractiveness, visibility and thus a better integration of IFB platforms as partners in national and European calls. This should increase the part of own resources and/or equipment pooling. The main challenge will be to reach a mean rate of 5% to fund required equipment and associated operating cost.
- **Increase IFB - Industry partnership** to 5% of incomes. As explained in section 2.2, the recent recruitments of the IFB cell "Communication and valorisation" should increase significantly the partnerships with private companies. The pricing of the services (e.g. training, data management, data curation, etc) will contribute to increase the part of own resources of the IFB platforms.

In their report, the international jury for the evaluation of IFB (June 2019, **Appendix A2.3**) pointed out the fact that "*There will be a need for continued funding of platforms - sustainability by user fees is "unrealistic"* ". Indeed, it is not conceivable that only the supervisory bodies cover the remaining expenses of all the IFB's regional platforms. Moreover, it would not be fair that only funded projects could access the IFB/MuDiS4LS services. Contribution of the science operators is necessary to maintain a well-scaled and competitive infrastructure. Thus, we expect strong support from French government (this call), but also research institutions and French Regions via co-funding mechanisms such as "Contrats de plan État-Région" (CPER 2020-2025, under evaluation).

Beyond the time of research projects, it will not be possible for any one of the small/medium/large infrastructure such as ours to meet the needs of long term conservation without a guarantee of long-term funding. With regard to international repositories that assume this function for the life science community, it is expected that projects will be able to still use them in the future.

1.4. Technical environment and sharing

1.4.1. Existing infrastructures

Table 3. Current state of the NNCR computing infrastructures (06/2020)

Platform	Localization	Computing (#CPU HT)	Storage (#TB)	RAM (#GB)
Federation of clusters				
IFB Core	IDRIS	4300	400	20008
ABiMS	Roscoff	2600	2500	10600
GenoToul	Toulouse	6192	4096	34400
GenOuest	Rennes	1866	2300	11616
BiRD	Nantes	560	600	3800
MIGALE	Jouy en Josas	1016	350	7000
TOTAL		**16534**	**10246**	**87424**
Federation of clouds				
IFB Core	CCIN2P3	3936	408	20408
GenOuest	Rennes	600	350	2600
PRABI	Lyon	448	144	1500
BiRD	Nantes	512	50	2048
Bistro	Strasbourg	200	50	1024
Bilille	Lille	192	0	768
TOTAL		**5888**	**1002**	**28348**
GRAND TOTAL		**22422**	**11248**	**115772**

1.4.2. Shared equipments

To address the needs of life scientists and support developments in bioinformatics, IFB will enhance the power of its infrastructures and strengthen the national network of computing resources that became indispensable to manage the data fluxes produced by high-throughput technologies. This will be led in close collaboration with the main data-producing research infrastructures (**INBS**) in order to contribute to a better management of data life cycle, compliant to the FAIR practices. On the other side of the data cycle, its outreach will require a specific handling defined in interaction with international repositories, complemented with the setting up of national institutional repositories.

Core resources already provide the hosting of central services (available to all users and IFB platforms, such as the national Galaxy instance (usegalaxy. fr) or the deployment of online resources developed by French teams and registered in the Service Delivery Plan (SPD) of ELIXIR: Phylogeny.fr (ATGC), LoRDEC (ATGC), RSAT (TAGC), Genomicus (ENS), Workflow4Metabolomics (ABiMS) as well as novel resources developed by Pasteur (NGphylogeny.fr , COVID-19 tools). This resource pooling also enables developers to delegate to the task force the administration of the "lower" layers (operator system, software environment, job load management, ...) and to focus on their vocation (bioinformatics).

From the beginning, MuDiS4LS has adopted a bottom-up approach by asking the INBS about their computing and storage needs. Thus 5 partner INBS will delegate, through this project, all or part of their infrastructure requirements. The equipment that will be purchased on their behalf will be pooled on the IFB's core resources or regional nodes. Beyond that, the ESR/Equipex+ ALADIN (IBISBA) project plans to pool its calculation servers to the IFB Core Cluster. The IFB will thus be able to take advantage of unused resources and vice versa. In addition, ALADIN will take advantage of the IFB Mutualized Task Forces already in place for implementation, maintenance and support.

Finally, in close collaboration with France Génomique, in the French regions, where several modes of access to computing (HPC/HTC/Cloud) co-exist within the same data center, storage spaces will be shared and made accessible for computing in a transparent manner for the user (data lake). The integration into the IFB scope of the Cloud services provided by France Génomique for sequencing will rationalize the storage and computing services provided by the two infrastructures.

1.4.3. Anchoring of the NNCR in national and regional data centers

IFB's development strategy is consistent with the structuring strategy roadmap of the Ministry (**Figure 3**), which will organize the national landscape by labelling regional data centers.

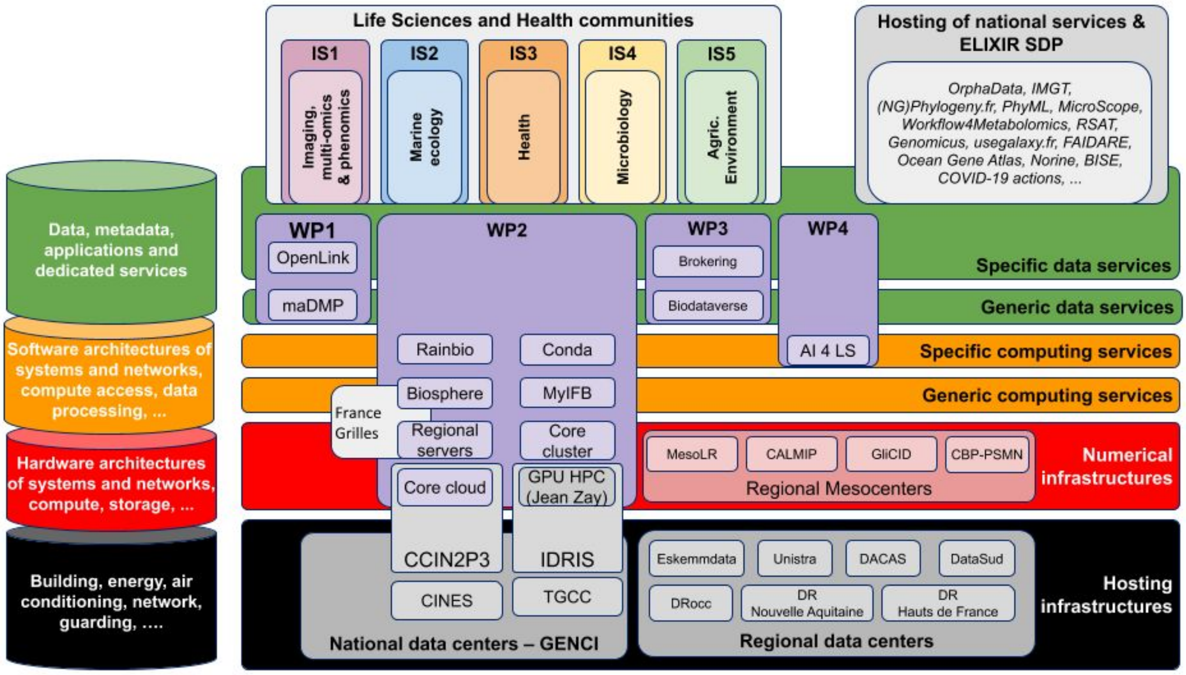

Life Sciences and Health communities
IS1
Imaging, multi-omics & phenomics
IS2
Marine ecology
IS3
Health
IS4
Microbiology
IS5
Agric. Environment
Hosting of national services & ELIXIR SDP
OrphaData, IMGT, (NG)Phylogeny.fr, PhyML, MicroScope, Workflow4Metabolomics, RSAT, Genomicus, usegalaxy.fr, FAIDARE,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

Figure 3. Positioning of the project in the e-infrastructure model.

The IFB regional platforms have already identified the regional data centers with which interactions are or will be set up (cf. section 1.3.1, Goal2). Several models already exist, and are likely to continue: dry hosting, integration into existing meso centers or data centers, featured hosting. In addition, new approaches will emerge such as the creation of new meso centers including one or more platforms. The impact of these integrations of bioinformatics platforms with meso centers or data centers will have the effect of bringing new constituted communities towards these infrastructures, amplifying the structuring role of the project at the national level. The multi-site infrastructure promoted by MuDiS4LS crosses the hardware and software layers in order to provide all the partner units (IFB platforms, INBS, Institutes) and end-users (basically, all the life science communities) with a broad provision of specific services encompassing data, metadata and software applications.

2. Dissemination and exploitation of results

2.1. Impact of MuDiS4LS in the scientific community

The present proposal is in line with the current IFB road map. Impact of the IFB on the scientific community has been measured over the three past years, showing a clear added value of the provided services (**Appendix A1**, section indicators). In 2019, we can highlight (i) 270 publications in which IFB and/or one of its platforms is (co-)authored or acknowledged; (ii) 800 collaborative and service projects, of which 1/3 involve foreign collaborators; (iii) about 8,000 users, more than 18,500 downloads of the tools available in our catalog, and 72 innovative software developed; (iv) 138 training courses and about 2,000 trainees (17% from private companies).

The core of the present proposal is based on several surveys conducted over the last two years, about the bioinformatics needs of IFB partner institutions' research teams. Clearly, a very strong demand has been expressed for data analysis (with the needs in GPU processors), data management and mid-term data storage during the life-cycle of research projects. These surveys also highlighted the request for new services around DMP, data FAIRification, integration of multi-omics data, and data deposition in national and international repositories. Indeed, the increasing number of large-scale projects are currently generating a huge demand from users of research units and private companies for access to such services, training activity or skills to support scientific projects.

Moreover, in the context of the "Shared services with other national Research Infrastructures (RI)" action of the IFB road map, and during the preparation of MuDiS4LS, we have gathered the needs of each national RI in terms of data storage: they were asked to give us a projection of their needs in each geographic region (i.e., where the data are produced), with an estimation of the proportion of data storage they wish to delegate to IFB-MuDiS4LS. Two types of interaction have been defined: (i) three RI became partners of the project with part of their numerical needs delegated to IFB: France Génomique, France Bio Imaging, and EMBRC-FR, (ii) Three other RI are positioned as collaborators and will essentially rely on the results of MuDiS4LS for the automatization of the data management flow (maDMP) and for the data FAIRification: FRISBI, MetaboHUB, ProFI. Moreover, two industrial demonstrators are also involved as partners in MuDiS4LS: MetaGenoPolis, IBISBA (cf. **section 3.2**).

Altogether, our project is structured to provide significant support for a broad range of applications in life science and health (cf. **section 1.3.1**). Goal1 and Goal3 will respond to the needs of these various communities and thus directly impact on the efficiency and quality of their research projects. Our main objective is to propose new tools and processes that will be interoperable within PFs of our RI and between others RI, with the aims to guarantee the widest possible access and sharing of data and metadata for a better scientific exploitation, and to ensure their durability through common approaches and mutualized digital infrastructures.

It is noteworthy that regarding the French personalized medicine (FMG2025) plan[7], links have already been identified between the CREFIX[8] component and IFB in connection with research (I.e., launch of joint research project calls, especially in integrative bioinformatics) and development activities (i.e., benchmarking of tools). Moreover, the positioning of IFB regarding the CAD[9] of the FMG2025 plan (in charge of the collection, organisation and management of the sequenced genomic data and part of the clinical data) and the Health Data Hub[10] (HDH) is pretty clear: with its competences in integrative bioinformatics, IFB is a kind of "proxy" able to retrieve different types of data (which are not in the CAD or in the HDH, *i.e*. proteomic and metabolomic data) and to answer specific biological questions.

Thus, MuDiS4LS contributes to integrate the various local and national initiatives (CPERs, other ESR/Equipex+ projects) and, importantly, European (ELIXIR, EOSC) and International initiatives (RDA). The role of IFB, as the French node of the ELIXIR bioinformatics infrastructure (www.elixir.eu) is summarized in the IFB 2017-2018 report[11]. Through our leading role in several communities that are instrumental to drive priorities for development and capacity building with researchers (Plant, Human CNV, Marine metagenomics, proteomics, metabolomics, Microbial biotechnology communities and Galaxy), scientific communities associated to MuDiS4LS will benefit from the impact of ELIXIR to the French node: (i) a powerful environment supporting the European contribution to the development of standards and generic tools for the community in life sciences, (ii) the support of co-development activities through internal projects between ELIXIR nodes, staff exchanges, but also through hackathons, (iii) the interactions with other ESFRIs (EMPHASIS, EMBRC, ECRIN, MIRRI, IBISBA) and the participation to EU-funded projects (4 ongoing H2020 projects led by ELIXIR hub and 4 others submitted or in preparation led by ELIXIR node).

2.2. Impact of MuDiS4LS on public policies and in the socio-economic world

Beyond its contribution to academic research, the objectives of the MuDiS4LS project are totally consistent with those of national public policies. First of all, the proposed pooling of digital and human resources in national and regional data centers that are currently being labelled by the French ministry of Higher Education, Science, Research and Innovation (MESRI), is in line with the ongoing rationalization of computing infrastructures in France and aims to contribute to the preservation of our environment. Second, the MESRI has engaged a National plan for Open Science to ensure that "the results of scientific research are open to all, researchers, companies and citizens, without hindrance, without delay, without payment", and in strong coherence with this plan, research organisms are implementing a roadmap to accelerate this process. The MuDiS4LS project, in

[7] https://www.aviesan.fr/en/aviesan/accueil/toute-l-actualite/plan-france-medecine-genomique-2025

[8] Centre de Référence, Innovation et Transfert

[9] "Collecteur Analyseur de Données"

[10] https://solidarites-sante.gouv.fr/IMG/pdf/181012_-_rapport_health_data_hub.pdf

[11] https://doi.org/10.5281/zenodo.3520131

collaboration with one main actor of the CNRS roadmap (INIST), is totally in line with this plan, especially on the main objective "Develop a culture of data management/sharing among all stakeholders in the data life cycle: researchers, engineers, computer scientists, librarians, etc. based on the implementation of the FAIR principles (findability, accessibility, interoperability and reusability)"[12]. Third, the five Implementation Studies will drive the technological developments on the required equipments; they will be the starting point for research projects addressing societal challenges, especially the "One Health" concept (combining human health to environmental and animal health, and to food quality and the control of microbiota to fight the pathogens), and the adaptation to climate changes and the development of green economy.

The economic valorisation of IFB/MuDiS4LS services will mostly rely on industry - platform collaborative contracts and on the pricing of the services. We are setting-up a "Valorisation and communication" cell composed of IFB representatives (of which 2 CDDs respectively dedicated to the IFB communication and to the IFB links with industries), and representatives of the knowledge transfer departments of the IFB supporting research institutions (CNRS-innovation, INRA-Transfert, CEA-Valo, INSERM-Transfert). Following the IFB Industrial Advisory Board (IAC) recommendations on current contracting modalities, a one-stop shop procedure allowing to handle globally all the requests addressed to IFB should be discussed in priority. Composition of this IAC will also be revised according to existing industrial collaborations of some MuDiS4LS partners (*i.e.*, MetaGenoPolis, IBISBA). Partnership with private companies should indeed be largely increased as MuDiS4LS can insure them with scaled and secured infrastructures, support (expertise and tools) to make FAIR data and tools, but also cutting edge expertise for training and data analysis, in support of their scientific programs.

3. Management framework

3.1. Management

3.1.1. Relevant experience of the project manager

The project is managed by a team composed of Jacques van Helden (scientific management), Gildas Le Corguillé and Julien Seiler (technical management).

Jacques van Helden (M), 55 y/o, is Professor of Bioinformatics and Biostatistics at Aix-Marseille Université. Since June 2017, he is co-director, with Claudine Médigue, of the Institut Français de Bioinformatique (IFB). Since 1997, his research activities are dedicated to the conception, implementation, evaluation and application of bioinformatics approaches to analyse genome regulation and biomolecular networks. He developed Regulatory Sequence Analysis Tools (RSAT, http://rsat.eu/), as well as network-based approaches to infer metabolic pathways from sets of functionally related genes (operons, co-expression clusters, phylogenetic profiles, ...). His current research focuses on integrative approaches to genomic regulation based on multi-omics data.

Gildas Le Corguillé (M), 37 y/o, is a bioanalyst/bioinformatic engineer (Sorbonne University), at the Station Biologique de Roscoff (FR2424) within the ABiMS bioinformatic platform since 2009 where he participated in different local and national projects: genomic, phylogeny, transcriptomic, genetic, etc.

[12] https://www.science-ouverte.cnrs.fr/wp-content/uploads/2019/11/CNRS_Roadmap_Open_Science_18nov2019.pdf

In 2013, he shifted to explore the Galaxy framework. In constant interaction with the Galaxy community, he became in 2017 co-leader of the ELIXIR Galaxy Community. Currently, technical leader of the ABiMS platform, he leads the third open HPC infrastructure for life science in France. Since 2018, he is dedicated at 20% for the IFB where he co-leads the IFB Core Cluster and the IFB NNCR Cluster. He also leads the implementation of usegalaxy.fr, gathering effort in France for a main national Galaxy instance. Finally, he is part of the project maDMP4LS.

Julien Seiler (M), 37 y/o, is an engineer in computer sciences working at CNRS for 10 years. Since 2011, he has been head of IT at the Institut de Génétique et de Biologie Moléculaire et Cellulaire in Strasbourg. He is leading a team of 15 people managing all IT infrastructures of the IGBMC, including a dedicated HPC cluster (800 cores) and distributed storage (more than 2Pb) for 800 users. Since 2018, he has devoted 20% of his working time to IFB where he co-coordinates the core cluster taskforce. He led the implementation of the SLURM infrastructure and storage and initiated a collaborative tool deployment solution based on Git. From 2020, he leads the OpenLink project, a gateway between scientific data management tools to enforce FAIR principles, closely linked to the maDMP4LS IFB project. He is involved in bioinformatics training (Degree in integrative bioinformatics (Paris University) and the School of Bioinformatics (Aviesan and IFB)).

3.1.2. Coordination modalities

The general governance of the project will rely on IFB general governance scheme (**Appendix A2.1**). In particular, the coordination with the partner organisms will be ensured by the existing IFB board (which emanates from IBiSA, the coordination of all the national infrastructures for life sciences and health), and the coordination with the leaders of IFB bioinformatics platforms will be ensured by the CIPI cell (Cell of Interactions between regional Platforms and IFB-core).

Three additional bodies will be defined to ensure the coordination of the MuDiS4LS project.

MuDiS4LS steering committee ensures the general coordination of the project and takes the strategic orientation choices. It is composed of

- The IFB director, Claudine Médigue
- The scientific and technical coordinators of the MuDiS4LS project: Jacques van Helden, Gildas Le Corguillé and Julien Seiler
- The coordinators of the NNCR cloud federation: Christophe Blanchet and Olivier Collin
- Economic model: Christine Gaspin
- Three delegates of the partner INBS: Edouard Bertrand (FBI), Gemma Jiménez Papiol (EMBRC), Pierre Le Ber (France Génomique)
- One delegate of Inserm DSI
- One delegate of GENCI, who will also establish the link with the national computing centers

MuDiS4LS operational cell is in charge of ensuring the technological implementation of the 4 work packages and the coordination between them, as well as to address the requirements of user communities. It is composed of (i) the scientific and technical coordinators of MuDiS4LS, (ii) the coordinators of each work package, and (iii) one representative of each national computing center partner unit (IDRIS, CINES, TGCC).

MuDiS4LS user committee is in charge of following the development of the five Implementation Studies, ensuring their coordination both with the 4 work packages and between others ISs. It is composed of (i) the coordinators of the 5 implementation studies and (ii) one representative of each collaborating INBS.

3.2. Organisation du partenariat / Collaboration organization

3.2.1. Relevance and complementarity of partners

MuDiS4LS regroups 36 partner units, including 20 IFB platforms, 5 data-producing INBS (FBI, France Génomique, EMBRC, FRISBI, FLI), 2 industrial demonstrators (IBISBA, MetaGenopolis), an additional cloud infrastructure (CBP-PSMN), the infrastructure that federates biological resource centers (RARe), the Inserm IT department (DSI), 2 national computing centers (IDRIS, CINES), 3 mesocenters (MESO@LR, CALMIP, GliCID) and the INIST. A summary description of each partner and their involvement in this project are provided in **Section 6**.

The complementarity of these partners is evidenced on **Figure 3**, where they occupy all the layers of the e-infra schema, from the physical infrastructure to the end-user communities. The project takes ground on an excellent distribution of the disciplinary skills between these different types of actors: the national and regional data centers (black layer on Figure 3) provide a highly secured and optimised physical environment where IFB platforms will deploy their numerical infrastructure (red layer), with different hosting modalities (dry, features) depending on the regional opportunities. Of note, the participation of Inserm SI and CINES pave the way to initiate academic solutions to address the crucial need of a numerical environment compliant with the new regulation about health data hosting (HDS).

IFB platforms will also act as facilitators to encourage life sciences to integrate the complex yet crucial novel digital technologies (e.g. AI, deep learning) in ambitious and innovative projects. Importantly, the capacity of the actors to operationally manage the interconnections between all the layers is demonstrated by the fact that they already form the basis of the current deployments of IFB core services (cloud at CCIN2P3, cluster at IRIS) which will be generalized to the extended NNCR targeted by the MuDiS4LS project.

In addition, this project established collaborations with several other INBS who do not explicitly appear as partners but provide support letters (**Appendix A6.3**). These collaborations correspond to a support from IFB to engage the user communities of these infrastructures for the FAIRification of their data and integration with other data types, the adoption of the maDMPs and the development of data brokering with the repositories of reference in their domains.

Altogether, this project pursues and extends the collaborations established with 17 INBS (France Genomique, France BioImaging, EMBRC-FR, ProFI, MetaboHUB, FRISBI, etc.) and other PIA2-funded structures (cohorts, Plan France Médecine Génomique 2025) since 2018, for the pilot projects in bioinformatics. These pilot-projects established long-term relationships, which led us to carry out a systematic estimation of the requirement of storage resources for each INBS in each region, and to host a defined proportion of their data on mutualised spaces of the IFB infrastructure. The complementarity between partners is also highlighted by the involvement of the partner organisms in the different implementation studies, which reflects the diversity of their domains of application (agriculture, microbiology, marine environment, health).

Key Performance Indicators for this project will be derived from those collected since 2017 by IFB in relation with its objectives of being a national or world scientific leading RI and an enabling facility to support science (OECD, 2019, DSTI/STP/GSF(2019)1/FINAL): number of active users, CPU + GPU time, data storage, publications acknowledging the infrastructure, number of projects with external grants. Additional KPIs will be collected in relation with the new services developed through this project on its structuring effects: number of projects adopting the maDMP, HPC GPU computing, number of projects involving the data brokering service.

We have identified a few risks. The first one concerns maDMP, which is central to the project and requires a long term involvement of INIST. INIST is not a platform of IFB and might change its priorities or lose its support. Alternative maDMP solutions emerging in the ELIXIR network will be considered as replacement. The project relies on key human skills and competences that might not be sufficient to scale up; recruiting skilled informaticians is very competitive. We will actively search for complementary funds, train new people and help them to develop their career to make IFB an attractive place to be.

3.2.2. Qualification, role and involvement of the partner units

Prénom Nom	Position	Discipline / Domain	Unit	Organization/ Establishment	Contribution in the project
Anne Françoise Adam Blondom	DR	Genetics and genomics	PlantBioinfoPF	INRAE	Scientific coordinator of the Plant pilar of RARe and deputy Head of nodes for ELIXIR-France. Co-lead of IS5
Abdelkader Amzert	IR	Computer sciences	Inserm DSI	Inserm	Co-lead of IS3
Edouard Bertrand	DR	Computer sciences	FBI	CNRS	Scientific coordinator of the FBI PIA3 project. Member of MuDiS4LS steering committee
Christophe Blanchet	IR	bioinformatics	IFB-core	CNRS	NNCR-cloud co-head. Biosphere coordinator. Co-lead of WP4
Yves Bourne	DR	Structural biology	FRISBI		Delegate of the FRISBI INBS for MuDiS4LS
Olivier Collin	IR	Bioinformatics	GenOuest	CNRS	NNCR-cloud co-head. Co-lead of WP3
Erwan Corre	IR	Bioinformatics	ABiMS	CNRS	Economic model. Head of ABiMS plateform Co-lead of IS 2
Fayza Daboussi	DR		IBISBA		Delegate of IBISBA for the MuDiS4LS project
Frédéric de Lamotte	CR	Data Science	Agap	INRAE	Co-lead of WP1
Boris Dintrans	DR	Computer sciences	CINES	CNRS	Head of the CINES. Co-lead of IS3
Michel Dojat	DR	Medical imaging	FLI	INSERM	FLI-IAM contribution to FAIRification of hterogenous data (IS1)
Jean-François Dufayard	CDI CIRAD	Computer sciences	South Green	CIRAD	Mutualisation between national infrastructures for life sciences and health. Co-lead of IS1
Emmanuel Faure	CR	Computational Biology	France Bio Imaging	CNRS	Delegate of FBI. Co-lead of IS1
Philippe Hupé	IR	Bioinformatics, Statistics	Curie	Curie	Co-head of Curie platform. Co-lead of WP4
Pierre-François Lavallée	DR	Compuer sciences	IDRIS		Delegate of IDRIS for MuDiS4LDS
Gildas Le Corguillé	IE	Bioinformatics	ABIMS	Sorbonne Université	MuDiS4LS technical coordinator. Co-head of NNCR-cluster. Co-lead WP1
Lucas Leclère	CR	Genomics	LBDV		Delegate EMBRC. Co-lead of IS2
Valentin Loux	IR	Bioinformatics	MIGALE	INRAE	Head of MIGALE plateform. Co-lead of IS5
Claudine Médigue	DR	Bioinformatics	MICROSCOPE	CNRS	Co-head of IFB. Co-lead of WP3. Co-lead of IS3

Denis Milan	DR	Genomics	France Génomique	INRAE	Articulation with France Génomique (WP2)
Ivan Moszer	IR	Bioinformatics	iCONICS	ICM	Co-lead of IS3
Jérôme Pansanel	IR	Computer sciences	BigEst	CNRS	Head of the SCIGNE platform, Biosphère partner
Perrine Paul-Gilloteaux	IR	Bio Image Informatics	France Bio Imaging	CNRS	Delegate of FBI. Co-lead of IS1
Eric Pellettier	DR	Environmental Genomics	France Génomique	CEA	Co-lead of IS2
Nicolas Pons	IR	Bioinformatics	MetaGenoPolis	INRAE	Co-lead of IS4
David Salgado	IR	Bioinformatics	MMG-GBIT	Inserm	Co-lead of IS3
Olivier Sallou	IR	Bioinformatics	GenOuest	CNRS	Co-lead of WP2
Julien Seiler	IR	Computer sciences	BigEst	CNRS	MuDiS4LS technical coordinator. Co-head of NNCR-cluster. Co-lead WP3
Guillaume Seith	IR	Computer sciences	BigEst	Inserm	IFB Core Cluster. Co-lead WP3
Michèle Tixier-Boichard	DR	Animal genetics	RARE	INRAE	Director of CRB RARe IS5 co-lead

4. Justification des moyens demandés / Funding justification

4.1. Justification des moyens demandés par élément / Funding justification by element

4.1.1. Element 1 et 2: Compute and Storage equipments for the core and the regional IFB platforms

Equipment bought for Elements 1 and 2 will be standardized as much as possible in order to reduce costs, including human resources dedicated to their implementation and management. However, we will adapt to local constraints of hosting data centers. It is also important to keep in mind that configurations and prices will keep evolving until their acquisition.

CPU nodes

The compute node base will be Dell C6420 type servers with 4 nodes per 2U enclosure. Each node will benefit from 112 hyperthreaded cores, 384 GB RAM, a dual network connection at 25 Gb/s and a 7 years expandable warranty. The average cost of one node for 8 years is estimated at 26,880 €, including associated network gear. Across the whole NNCR network, we plan to acquire 200 servers of that kind, summing up to 22400 cores and 76 PB RAM.

To ease the budget estimate for the project, we have identified an average generic configuration of the compute nodes. However, keeping constant costs, we are considering acquiring Bigmem type compute nodes with 1 to 3 TB RAM, depending on needs. Such machines are recommended for genome assembly needs, among others.

GPU nodes

The hardware solution relies on Dell R740-type GPU servers composed of 2 to 4 GPU cards, 80 CPUs and 64 GB RAM and a dual network connection at 25 Gb/s. The average cost of a GPU server is estimated at 20,000€ for an 8-year duration. The cost of that kind of machine may vary significantly depending on the integrated GPU model.

Across the whole NNCR network, we plan to acquire 34 servers of that kind, adapting the choice of GPU cards to the needs.

To date, we have identified 3 professional GPU models able to answer the bioinformatics uses :

Model	Performance (simple precision)	RAM	Unit cost
Quadro RTX 6000	13.69 TFlops	24 GB	3500 €
Tesla T4	7.757 TFlops	16 GB	1900 €
Tesla V100	16,67 TFlops	16 GB	7500 €

Storage equipments

Based on the experience of the Core Cluster which has evaluated several storage solution, two important criteria must be taken into account for the storage equipment :

- The average cost of a professional storage facility is around €400,000 for 2 PB.
- This cost is roughly the same whether one is aiming for a capacitive infrastructure, with a data replication system to ensure its durability, or whether one is aiming for a high-performance infrastructure with very short data access times (but less reliability in terms of data preservation in the event of a component failure).

The DDN solution and Lustre Open Source File System on which we based the budget estimate for this project offers a good level of reliability with a declustered RAID system as well as a high availability rate with full case redundancy and fractional reconstruction.

Given the reliability and performance qualities of the Lustre system, this type of solution can be considered to host both mid-term storage or to meet specific performance needs in scratch mode. The DDN quote on which our budget estimate is based is valid for a 2PB capacity solution with a 5-year warranty.

4.1.2. Element 3: inter-site data securing

Replica sites will be accessible through standard protocols to allow a simple but secured access from any infrastructure (http access). The storage software will rely on an object technology (openIO) in order to give that modern and low cost access. That solution will allow to easily increase the available capacity on each replica site according to the evolution of infrastructure backup needs and to keep a history of changes on a limited timeframe (versioning on X days).

Backups will be made using tools integrated in primary storage (storage to replicate) or the open source tool Rclone (https://rclone.org). This will minimize the network impact, for the infrastructure to backup as well as for the replica site. This tool based on the famous and robust rsync program will only copy the latest modifications (daily or weekly).

The OpenIO licence costs 645,000€ for 4 PB during 8 years. On the hardware side, the solution will rely on Dell R740XD2 type storage nodes, 22 x 14 TB HDD, 20 cores, 64 GB RAM, dual 10/25GB connector (Mellanox), linked to 48 ports 10/25GB/s Dell S5248F-ON type PoweSwitches.

Hardware costs for 4PB usable storage (servers: 478,584 € for 7 years; power switches: 13,600 € for 7 years). The storage systems will be connected to Renater with 10GB/s minimum speed.

4.1.3. Element 4: Health data hosting and secured research environments

The health data computing infrastructure will be based on a set of Dell R740 hypervisors (40 cores, 768 GB RAM, 23 TB of SSD and a 25Gb dual attachment). The infrastructure will be complemented by a firewall, network switches and a network monitoring solution based on SEC technology.

The hardware cost is estimated for 5 years as: hypervisor : 47k€ each, Palo Alto firewall : 141k€, SEC : 60k€, network switches and cables : 106 k€.

The health data storage infrastructure is built on NetApp storage arrays for file storage and an object storage solution based on the Scality solution.

Storage cost estimation (5 years): NetApp: 405€/Tb w/backup, Scality Apollo 4510 : 2847Tb, 66 k€.

4.1.4. Element 5: BioDataverse

The BioDataVerse will be developed on a base unit of 1PB replicated storage initially. This base will be designed with a goal of scalability in order to answer progressively INSB needs. That infrastructure will rely on three main technological blocks:

An object storage service implementing S3 specifications

That service will be composed of 6 Dell R740XD2 type storage nodes of 300GB each with dual 25GB connection. The object storage system will be implemented with an SDS solution as OpenIO or Scality allowing data securitisation for all the storage nodes through an Erasure Coding algorithm.

The DataVerse software and its components

The data warehouse will be managed using the DataVerse (https://dataverse.org/) open software. This software requires additional components as the Solr indexing engine. All these tools will be deployed on two Dell R640 type hypervisor with 48 cores and 512GB RAM

Distant replica site

Finally, an object storage replica will be implemented on a distant site. That solution will rely on an infrastructure similar to the primary storage one (6 nodes Dell R740XD2 and OpenIO).

4.2. Summary of funding justification

4.1.1. MuDiS4LS contribution to the global needs of the IFB infrastructure

In the context of this project, IFB platforms and the INBS partners evaluated their needs in computing and storage infrastructure for the 8 next years. The ESR/Equipex+ MuDiS4LS project aims at covering the following proportion of these total needs.

- 33% of the equipment acquisitions
- 31% of the functioning costs including dry hosting and maintenance of the equipment
- 33% to 40% of the sub-contracting for the featured hosting
- 100% of the Jean Zay supercomputer converged platform fast-track access

The complement should come from other funding sources: CPER, self-funding (tarification, collaborative projects), long-term funding by the supporting research organisms and stakeholders, future national calls for equipment.

4.2.2. Economic model

Involvement of permanent and requested personnel

The change of scale and the scientific and technological ambitions developed in this project require a significant contribution in staff both during phase 1 of the project (installation) in relation with equipment scaling and implementation of new services and during phase 2 (exploitation) on the support for the new proposed services. These two phases will benefit from a continuous investment of permanent and non-permanent staff already present in partner units of IFB and whose salaries are covered by the partner organisms (see **Appendix A4** for summaries of the personnel involvement).

In addition to this established task force, it will be necessary to recruit people to support equipment scaling and build new services around the data (this project, 6 FTEs are requested for WP1-4 during the 4 years of the installation phase) but also to support the evaluation of the usability of these new services in the context of several case studies representative of the needs of communities in Life Sciences (this project, 7 FTEs for 2 years, will start on middle of phase 1 to support IS1-5). In phase 2 of the project (exploitation), the partner organisms made a commitment to open 7 permanent positions to consolidate the staff of partner units around the new proposed services.

Opening of permanent positions by the partner organisms

This project benefits from a strong support from the partner organisms, which are opening 7 permanent positions to ensure the sustainability of the infrastructure:

1. **CNRS** opens a permanent position of **System Administrator for IFB-core facilities** in 2020, to ensure the system administration of the mutualized NNCR cluster and cloud facilities.
2. **CNRS** will open a second permanent **Sys Admin position for IFB-core** in the course of the project.
3. The Computing sciences institute of **CNRS** (INS2I) opens another position to develop **user interfaces** to give access to bioinformatics workflows for life scientists
4. The MATHNUM division of **INRAE** will open a **DevOps position** (Toulouse, Occitanie region) to contribute to the IFB task force. He will contribute to install equipment in datacenters and to ensure that software will work across a diverse set of operating systems and platforms.
5. The MATHNUM division of **INRAE** will open a position in statistics (Jouy-en-Josas, Ile de France) to contribute to data analysis and integration in relation to implementation studies IS4 and IS5.
6. **Inserm** recruits an Ingénieur de Recherche (with PhD in bioinformatics) to develop **bioinformatics for health**.
7. **CEA** recruits an Ingénieur de Recherche to take care of the **applications of Artificial Intelligence** to life sciences and health

In addition to the above-listed MuDiS4LS-specific positions, CNRS signed support letters for two additional permanent positions to be open at IFB-core before 2025, including a secretary who will contribute to the management of this project (acquisition of the equipment, maintenance, hosting, …) and a help desk to support the coordination of the services ensured by the 20 IFB platforms. The support letter signed by INRAE for IFB 2021-2025 road map in June 2019 includes the commitment to open 2 additional positions on INRAE platforms in line with the development of the national missions of IFB. This includes a DevOps position for the NNCR (in total 4 positions for 2021-2025).

Co-financing of the project

This project is co-funded by a 2.8M€ budget allocated to IFB by ANR for the 2021-2025 extension of the PIA2 funding. This budget secures the functioning and hiring of temporary personnel to ensure the coordinating role of IFB-core, as well as the actions defined in the 2021-2025 road map. Some of these actions are complementary to the present project, in particular interoperability (for the

FAIRification of the data), mutualisation between INBSs, training (with a special focus on DMP), call to challenges in integrative bioinformatics. IFB also benefits from a grant from ANR to hire a software developer for 12 months to work on the interoperability between OPIDoR and IFB infrastructure management software. Five IFB platforms applied to the CPER call for equipment complementary to and non-overlapping with the present project (these equipments are not part of the MuDiS4LS application): GenoToul (4,200k€), BiRD (1000k€), CBIB (855k€), ABIMS (700k€), Bilille (80k€).

Pricing system

So far, all the NNCR platforms didn't reach the same level of maturity in their self-funding based on the pricing of their computing infrastructure. Some have years of feedback, some never managed to put a price on it. A first pricing strategy was established, which relies on shared principles including two user categories (academic and private) and the definition of a free package of storage space and CPU hours. CPU hour and storage space per year were chosen as value units. Based on a full cost analysis including operating costs, a depreciation rate of 5 years for equipment and indirect costs, we established a price for the CPU hour per year (5c€ for academic users, 10c€ for private users) and the To of storage per year (250€ for academic users, 500€ for private users). This founding pricing will be applied as soon as a user will need more storage space and/or CPU hour allocation than defined in the free package. It will serve as a basis for evaluating resource costs in project calls either for a real use of resources or to propose bundles including discounts more suited to specific uses. Although the self-financing rate is still very uneven between platforms, we plan that 20% of our incomes per year should come from this pricing strategy in the next five years for most of the platforms. This income will cover hosting fees and part of equipment evolution. In line with this pricing strategy, the Core Cluster, in production since late 2018, recently shaped 9 bundles (Table 4). The two cursors should be the disk space and the quota of CPU.

Table 4. Projection of the IFB Core Cluster Pricing (academic / private company)

		Disk quota		
		< 250 GB	**250 GB < < 1,5 TB**	**1,5 TB**
CPU Quota	**< 10 k hours**	€ 0	€ 450 / € 900	€ 750 / € 1500
	10 k hours < < 65 k hours	€ 1700 / € 3400	€ 2000 / € 4000	€ 2250 / € 4500
	65 k hours <	€ 3000 / € 6000	€ 3375 / € 6750	€ 3750 / € 7500

4.2.3. Synthesis of the cost per element

Table 5. Synthesis of the cost for Element 1, 2, 3, 4 and 5 (core and regional servers). Complete version Appendix A5.3

Element	Total equipment	Total functioning equipment	Total Sub-contracting cost	TOTAL
Phase 1				
1&3 - Equipment for NNCR core resources	k€ 3,770	k€ 322	k€ 0	**k€ 4,092**
2&3 - Equipment for regional IFB platforms	k€ 9,573	k€ 565	k€ 513	**k€ 10,650**
4 - Health data hosting and secured research environments	k€ 573	k€ 29	k€ 190	**k€ 792**
5 - BioDataVerse	k€ 162	k€ 111	k€ 0	**k€ 273**
6 - Fast track access to the Jean Zay Supercomputing facility	k€ 0	k€ 0	k€ 346	**k€ 346**
TOTAL phase 1	**k€ 14,078**	**k€ 1,027**	**k€ 1,049**	**k€ 16,153**
Phase 2				
2&3 - Equipment for regional IFB platforms	€0.00	€0.00	k€ 559	**k€ 559**
4 - Health data hosting and secured research environments	€0.00	€0.00	k€ 230	**k€ 230**
6 - Fast track access to the Jean Zay Supercomputing facility	€0.00	€0.00	k€ 212	**k€ 212**
TOTAL phase 2	**k€ 0**	**k€ 0**	**k€ 1,001**	**k€ 1,001**
TOTAL	**k€ 14,078**	**k€ 1,027**	**k€ 2,050**	**k€ 17,154**

5. Key figures

Indicators	Current values	Estimation at +10 years
Publications and licenses in the relevant domains resulting from research led on the equipment (indicate the co-publications between partners and licenses co-owned with enterprises)	In 2019 the equipment contributed to 430 publications (platforms and/or IFB as co-authors and/or explicitly acknowledged). 69 of these were co-signed by multiple partners of this project, including 33 with France Génomique, 11 with EMBR-C, 9 with the GLiCID Mesocenter.	The number of citations per year is expected to increase: (1) past years indicators show a steady increase of the citations; (2) usage of high-throughput technologies ; (3) demand for integrative bioinformatics; (4) new citation guidelines IFB platform chart. A rough (but reasonable) estimate of 20% increase per year, would reach 2,660 publications in 2029 and a total of 13,400 for 2020-2029.
	The bioinformatics domain is not relevant to patents, especially with the Open Science orientation taken by IFB and ELIXIR. We prefer to avoid providing misleading indicators	
Number of researchers and teachers in the	IFB computing and storage facilities hosted 7811 active user accounts in 2019,	MuDiS4LS redirects users from small local facilities to NNRC mutualized digital space. We

domain (estimation of potential users)	with an approx rate of 80% researchers and teachers (6248).	will thus handle a higher number of users with higher space requirements. We estimate that the infrastructure should cover the needs of 15 000 to 20 000 researchers/teachers in 2029
Number of national partners	In 2019, the 20 IFB member platforms were engaged in 57 national projects.	This will primarily depend on the national funding policy, and on the orientation of the calls towards domains involving integrative biology and bioinformatics.
Number of international partners	In 2019, the 20 IFB member platforms were engaged in 36 international projects. A part of these projects were led in collaboration with the 3 other ELIXIR partners.	For 3 years France increased its participation in ELIXIR activities and ELIXIR-driven EU projects. The number of partnerships and partners is thus likely to further increase, but it would be speculative to predict a number at >=10 years.
Number of non-academic partners (indicate their nature)	For 2019, the 20 IFB member platforms had 18 partnerships with private companies.	The industry partnerships will significantly increase with the recruitment of a full-time person to promote IFB-industry partnership.
Yearly amount of the partner's contributions	For 2019 : collaborative projects raised **1,988 k€** (1473K€ academic + 515K€ with enterprises). The pricing of IFB services raised 2,144 k€ (318k€ for project hosting, **1,091 k€** for consulting and support to projects, and 242 k€ for training, 493 k€ for other services).	Self-funding is expected to raise with the establishment of a pricing on the core facilities and the adoption of the pricing model by other nodes of the NNCR. We target a steady rate of at least 20% self-funding.
Number of Master or PhD students using this type of equipment in the partner establishments	Master + PhD students represent ~16% users on IFB facilities (1040 users in 2019)	Based on the same reasoning as for researchers / teachers, we expect ~10,000 students to be hosted on IFB facilities in 2029. This is probably a lower bond IFB facilities are more and more solicited for university teaching.

References

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* **533**: 452–454.

Garcia, L., *et al.* 2020. Ten simple rules for making training materials FAIR. *PLoS Comput Biol* **16**: e1007854.

Ison, J., *et al.* 2013. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* **29**: 1325–1332.

Lamprecht, A.-L., *et al.* 2020. Towards FAIR principles for research software. *DS* **3**: 37–59.

Mehra, M.R., Desai, S.S., Ruschitzka, F. & Patel, A.N. 2020. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *The Lancet* S0140673620311806.

Millet, E.J. *et al.* 2019. A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios. Portail Data Inra.

Wilkinson, M.D. *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**: 160018.

6. Description of Partners

Institut Français de Bioinformatique (IFB)

Coordinators: Claudine Médigue and Jacques van Helden

Description. The **Institut Français de Bioinformatique** (**IFB**) is the national infrastructure of bioinformatics, which federates 30 bioinformatics facilities (20 member platforms and 10 associated platforms) ensuring services (compute and storage, software development and deployment, databases, consulting and support to projects), training and innovation to address the needs of all the life science and health communities. IFB is also the French node of the European bioinformatics infrastructure ELIXIR. IFB is funded since 2013 by the Programme d'Investissement d'Avenir (PIA) and supported by the main public research organisation (CNRS, INRAE, Inserm, CEA, INRIA). **IFB-core** is the coordinating unit that manages the mutualized resources and represents the infrastructure at the national and international level.

Involvement in the project. IFB is the scientific and technical leader of the MuDiS4LS application. It will contribute to the project with its two compute and storage facilities (IFB-core-cluster and IFB-core-cloud).

ABiMS (http://abims.sb-roscoff.fr/)

Corresponding people: Corre Erwan, Gildas Le Corguillé

The mission of the ABiMS platform is to assist researchers of the marine community and, more broadly, of the life sciences, in the bioinformatic analysis of their data as well as in the development of software and databases. It is one of the national platforms of the French Institute of Bioinformatics (IFB). It is one of the Data and Service Centres of the Ocean Data Cluster - Odatis, it is also associated to the EMBRC infrastructure, is part of the IBiSA network via the regional BioGenOuest project and is ISO 9001:2015 certified. Through its numerous interactions with research units, ABiMS is involved in several projects, with national and European impacts involving bioanalysis activities, software, and e-Infrastructures development. ABiMS provides a computing and storage infrastructure (2600 cores and 2.5 PB) associated with more than 500 softwares dedicated to data analysis. It provides expertise in VREs deployment for data analysis (Galaxy) and genome annotation (Gmod / Apollo), software engineering, bioanalysis and training.

ARTbio (https://www.artbio.fr/)

Corresponding people: Naïra Naouar, Christophe Antoniewski

The Accessible Reproducible and Transparent bioinformatics platform of the Institut de Biologie Paris Seine provides support to biologists and medical doctors for functional genomics and precision medicine. It is implanted on the Jussieu campus of the faculty of sciences of Sorbonne Université and is also strongly involved in clinical research as a partner of the SIRIC Curamus that gathers efforts in cancerology from the 5 university hospitals of SU. ARTbio is running 2 public Galaxy servers, https://mississippi.fr and https://usegalaxy.sorbonne-universite.fr which provide environments for small RNA profiling and variant analyses, respectively. It also provides public servers for R analyses as well as for data storage. A third Galaxy server is dedicated to ARTbio user projects. ARTbio develops open source bioinformatics softwares, most of them available as Galaxy plugin tools (more than 48 tools available in https://github.com/ARTbio/tools-artbio). Beside accompaniments of user projects under contracts, ARTbio is maintaining its own research lines in order to develop top level expertise in three specific domains: small RNA biology and viral small RNAs, statistical and machine learning methods for single cell RNAseq analysis, and predisposition mutations in children and young adult acute myeloid leukemia (partner of the CONECT-AML INCA project). ARTbio has defined training in bioinformatic as a top priority in the incoming years. Thus, in addition to the organisation of regular training sessions, ARTbio is nowadays known for its companionship offer and is also leading the project STARTbio for a University Diploma in Bioinformatics at Sorbonne Université, based on a practical approach of analyses as well as heavy use of e-learning tools (videoconferences and executable tutorials)

ATGC (http://www.atgc-montpellier.fr/)

Corresponding people: Lefort Vincent, Stéphane Guindon, Eric Rivals

The ATGC bioinformatics platform provides services for evolutionary, comparative and functional genomics. It highlights the methodological developments carried out by the Methods and Algorithms for Bioinformatics (MAB) team at LIRMM. The team has a long standing experience in software development (e.g. PhyML, FastME, LoRDEC, RAPPAS). These developments are disseminated to the scientific community as freely accessible services from the platform's website: http://www.atgc-montpellier.fr. Most of the tools can be run directly online, using a dedicated web interface from which each user can load his own data. These interfaces have been designed to facilitate the presentation of tools, analysis and interpretation of results. ATGC developed and maintained phylogeny.fr and developed a new version (NGphylogeny.fr) with Pasteur Hub. More than 20 original bioinformatic tools are available on the ATGC website. Regular training sessions are organized each year in *molecular phylogeny*, *bioinformatics for the processing of high throughput sequencing data* and *Linux and script for bioinformatics.*

Involvement in the project. ATGC will mainly contribute to WP1, WP2 and WP4.

AuBi (https://mesocentre.uca.fr/projets-associes/plateforme-aubi/)

Corresponding people: Peyret Pierre, Mahul Antoine

AuBi is the Clermont bioinformatics platform for life sciences (fundamental biology, microbiology, agronomy, environment, health and epidemiology). AuBi leans on UCA Mesocentre to promote access to computing facilities for data analysis, storage, training in bioinformatics and hosting web services for research. The platform is committed to serving UCA teams and Associates (public sector, companies ...). Important skills are developed to support scientific projects in the field of large scale sequence analysis in genomics (assembling, annotation, variant analysis, DNAseq, ...), transcriptomics (RNAseq, ChiPseq, ...) and epigenomic variations (BSseq), metagenomics, metatranscriptomics, metabolomics, molecular dynamics but also statistics and imaging. Services include data analysis, data storage, bioinformatic training on UCA Mesocentre cluster and Galaxy Network and access to web services (Galaxy and Omero).

Involvement in the project :

AuBi will contribute to WP2, WP3 and Implementations Studies IS1, IS4 and IS5.

bilille (https://wikis.univ-lille.fr/bilille/)

Corresponding people: Marot Guillemette, Touzet Hélène

Bilile is part of the UMS 2014/US 41 Plateformes en Biologie et Santé de Lille (PLBS, University of Lille, CNRS, INSERM, Institut Pasteur de Lille, CHU Lille). It has expertise in bioinformatics, biostatistics and bioanalysis and offers a complete service to Lille research labs in biology and health. This includes : support for scientific projects, HPC resources, training for biologists and bioinformaticians, organization of bioinformatics colloquium.. The application fields cover analyses of omics data (genomics, transcriptomics, proteomics,...), genome annotation and evolution, integrative biology, cytometry, phenotypic screening, glycobiology… Bilille is involved in this project for the participation in the Biosphere cloud federation and will provide human resources in bioinformatics and biostatistics for three implementation studies.

BiRD (http://pf-bird.univ-nantes.fr)

Corresponding people: Bihouée Audrey , Redon Richard, Bourdon Jérémie

BiRD is co-managed by the biomedical research unit "l'institut du thorax" and the digital sciences lab "LS2N". BiRD advises, proposes and develops bioinformatic services based on high-throughput sequencing data. BiRD has expertise in large-scale data analysis and developed dedicated bioinformatics workflows that standardize processing from raw data to biological significance. These services are supported by a dedicated computing and storage infrastructure which is remotely accessible through several services and open to all scientists, regardless of their host institution. BiRD provides computing and storage infrastructure for bioinformatics representing 832 threads and 650TB (Cluster, OpenStack Cloud integrated into the IFB Biosphere project). This infrastructure provides appropriate bioinformatics resources for the processing and analysis of omics data.

Partner involvement

The BiRD will mainly contribute to WP2 and IS1, through the funding of an engineer, member of the NNCR and the Cloud task force. This position, co-hosted with the FBI infrastructure project. The platform will also contribute to WP1 on the maDMP and FAIRification of data and workflows in connection with IS3 on the data integration in biomedical context. BiRD is also a partner of IS2 and IS4 through the expertise on ecosystem modeling of LS2N collaborators.

BigEst (http://bigest.unistra.fr)

Corresponding people: Cognat Valérie, Thompson Julie

The BigEst platform brings together bioinformatics teams and services from the Strasbourg site: GMGM, IBMC, IBMP, ICube, IGBMC, IPHC and LGM. The platform offers expertise, bioinformatics tools and resources, as well as data mining algorithms, focused on evolutionary and functional analyses in various application fields, including biomedical, plant, yeast and bacterial studies. BigEst maintains databases and analytical software in integrative plant biology, genetic diseases, comparative genomics, proteomics, non-coding RNA, and microorganisms. BigEst is involved in the development of the IFB Cloud, training on the Galaxy framework, and the pilot project for integrative bioinformatics. It is also highly contributing to the IFB Core Cluster as co-coordinator of the infrastructure.

Involvement in the project

BiGEst is mainly involved in WP2, WP3 and project coordination tasks. Indeed as an active member of the IFB Core Cluster and Cloud taskforce, BiGEst will work on the distributed data infrastructure of the NNCR and on the deployment of the BioDataverse.

CBiB (http://www.cbib.u-bordeaux.fr/)

Corresponding people: Groppi Alexis, Nikolski Macha

The CBiB is a bioinformatics core facility that provides access to high-performance computing resources, biological data analysis and programming expertise. The resources serve scientists and private labs to fulfill the bioinformatics needs of their research in an efficient and cost-effective manner. We offer state-of-the-art technologies for working with clinical, translational, and basic science data – from acquisition and storage to analysis and sharing. Our resources are secure and standards-compliant. From a few samples to several tens of thousands, the Bioinformatics Centre provides complete DNA, RNA, metabolomics, proteomics as well as image data analysis and integration services. The CBiB provides access to its computing and secure data storage infrastructure, standard "omics" data analysis pipelines, as well as tailored services such as database development and deployment, bioinformatics methods development in particular for "big data" and data integration approaches.

***Partner involvement*.**

In this project, the CBiB will mainly contribute to WP1, WP2, IS1, IS3, IS4 and IS5.

Genotoul Bioinfo (http://bioinfo.genotoul.fr/)

Corresponding people: Gaspin Christine, Hoede Claire, Klopp Christophe

The platform is located in Toulouse, south of France. Its goal is to bring together equipment resources and human skills that offer to life science research programmes an access to state of the art know-how and top level technologies. During past years the team has built up an expertise in diverse applications of sequence analysis. This knowledge has been used in software and pipeline developments as well as in data analysis projects including genome assembling (short/long reads), annotation (coding/non coding), (s)RNA-seq, methyl-seq data and variant analyses, metagenomics (metabarcoding/whole genome) and, more recently, data integration. Supported communities are agriculture, alimentation, human health, ecology and bioinformatics. Services include access to high-performance computing (3000 cores), storage resources (4Po), updated international databanks and software, support for data analysis and integration. Other activities include development of novel bioinformatics methods/tools, organization of training sessions and regional knowledge sharing.

***Partner involvement*.**

In this project, Genotoul Bioinfo will mainly contribute to WP1, WP2, IS2, IS4 and IS5.

GenOuest (http://www.genouest.org/)

Corresponding people: Collin Olivier, Sallou Olivier, Nicolas Jacques

Hosted at INRIA-IRISA, the GenOuest core facility offers a complete bioinformatics environment with hardware and software infrastructure, public databases, software and workflows, all with associated support. The technological portfolio relies on several computing resources (cluster, cloud, docker, galaxy portal), data management solutions (BioMAJ). During the years GenOuest has invested in the development of data-centered tools. Thanks to the CeSGO project, GenOuest offers a set of collaborative tools to manage projects and scientific data in the best possible way. GenOuest offers development of bioinformatics applications as well as training and technological transfer of new tools developed by research teams of the Institute. Services include access to the computational infrastructure (cluster, cloud, docker environment), IFB-core cloud support system, monitoring of the Biosphere cloud federation, hosting more than 30 scientific services and databases freely accessible to the community.

***Partner involvement*.**

In this project, GenOuest will contribute to WP1, WP2, WP3.

iCONICS (https://iconics.icm-institute.org/, https://neuroinformatics.icm-institute.org/)

Corresponding people: Moszer Ivan, Durrleman Stanley

The iCONICS platform (Paris Brain Institute – ICM) develops and makes available software solutions and methodological expertise to meet three important needs for biomedical research: data curation, standardization, annotation, structuration, integration and visualization (data managers, software engineers); high-throughput omics data processing, including NGS data such as whole-exome, RNA-seq (bioinformaticians); basic and advanced statistical analysis, especially integration of multimodal and high-dimensional data (biostatisticians). The platform thus supports scientific and clinical teams at every step of their research projects: from study design to data management, processing, analysis and interpretation, considering a wide variety of data acquisition approaches (clinical, imaging, genomics, etc.). iCONICS is a core component of the Center for Neuroinformatics of the ICM, whose main objective is to promote the harmonization and sharing of best practices in data management and analytics across the Institute.

***Partner involvement*.**

In this project, iCONICS will mainly contribute to WP1, WP4 and IS3.

InstitutCurie (http://u900.curie.fr/)

Corresponding people: Hupé Philippe, Servant Nicolas, Barillot Emmanuel

The expertise of the platform is versatile on many aspects of high-throughput data management, processing, integration and statistical and functional analysis in biology and in clinics. We cover fundamental, translational and clinical research, including clinical trials. In all these domains, the platform also develops appropriate methods, implement tools and automatic pipelines, and package and release them publicly or grant on-line access to the community. Software optimisation for high-performance computing is also part of our know-how. Finally the platform has developed an experience in training biologists, clinicians and bioinformaticians in all above-mentioned fields. Services include collaboration for high-throughput data analysis and training in bioinformatics.

MICROSCOPE (https://mage.genoscope.cns.fr/microscope)

Corresponding people: Medigue Claudine, Vallenet David, Dubois Mathieu

The LABGeM team from the CEA/Genoscope has developed MicroScope, a web-based platform for prokaryotic genome analysis and expert functional annotation. Services include microbial genome annotation (complete, WGS or metagenome-assembled genomes), comparative genomics and pangenomics, function and biological process predictions, metabolic network reconstruction, analysis of transcriptomics data, trainings on prokaryotic genome annotation and on the curation of metabolic networks. Expert annotations are continuously gathered in the MicroScope database contributing to the improvement of the quality of microbial genome annotations. MicroScope combines tools and graphical interfaces to analyze genomes in a comparative and metabolic context.. It provides data from thousands of genome projects and is used as a community resource for comparative analysis and annotation of publicly available genomes but also as a private resource with restricted access rights on genomic data.

Partner involvement. In this project, MicroScope will use the MuDiS4LS infrastructure to provide services to the community of microbiologists and will also contribute to several objectives of IS4.

MIGALE (https://migale.inrae.fr/)

Corresponding people: Loux Valentin, Schbath Sophie, Chiapello Hélène

The Migale bioinformatics platform is a team of the INRAE MaIAGE research unit (Applied Mathematics and Computer Science, from Genome to the Environment). Since 2003, it provides four types of services to the life sciences community: an open infrastructure dedicated to life sciences data analysis (500 Tb, 1000 equivalent CPUs), dissemination of expertise in bioinformatics (annual "Bioinformatics through practice" training session), design and development of bioinformatics applications (genome browser, databases), data analysis in genomics, metagenomics and metatranscriptomics. Services include access to the infrastructure and Galaxy environment, training, analysis services in genomics, comparative genomics, metagenomics and metatranscriptomics. Other activity is to develop tools for Galaxy (proteomics, metagenomics) and databases.

***Partner involvement*.**

In this project, Migale will mainly contribute to WP1, WP2, WP3, IS4 and IS5.

MMG-GBIT (http://geneticsandbioinformatics.eu)

Corresponding people: Salgado David, Béroud Christophe

The MMG-GBIT IFB platform is linked to the Bioinformatics & Genetics research team of Inserm U1251. It benefits from this strong interaction and provide users with the team expertise in human genetics, rare diseases and oncogenetics as well as NGS data analysis. It provides international reference systems to collect, annotate, filter and interpret human genetics data in relation to diseases. These tools, databases, registries and observatories have already received millions of worldwide queries. In addition, we provide trainings and can assist researchers for any bioinformatics project or tools development related to human genetics from design to analysis. Within IFB, David Salgado is co-leading the Work Group on bioinformatics for Health Data. MMG-GBIT plays an important role in the ELIXIR infrastructure as it is leading the hCNV community. The platform is involved in other ELIXIR human communities. MMG-GBIT is part of the EU projects such as "Beyond 1 Million Genomes project" (B1MG) and the "European Joint Program on Rare Diseases" (EJP-RD).

***Partner involvement*.**

In this project, MMG-GBIT will participate in WP4 and co-leading the implementation of the IS3.

Pasteur

Corresponding people: Ménager Hervé, Malabat Christophe, Dillies Marie-Agnès, Gascuel Olivier

The Bioinformatics and Biostatistics Hub is the service part of the Computational Biology Department. The Hub team involves 50 biostatistics and bioinformatics experts. The mission of the platform is to develop methodological research in bioinformatics and biostatistics, give visibility in this field to the Institut Pasteur on an international level, offer support to experimental research units, and develop the computational and analysis skills of the campus. The platform activities comprise participation in research and analysis projects, on-site assignments within campus units and platforms, training and teaching sessions open to our partners, and the provision of a number of resources to the national and international community. The Hub team and all the Department have a long standing experience in the development of web sites (e.g. NGPhylogeny.fr), intensive and parallel computations using both GPU and Big Mems, health data management. All these skills will be shared with the other project partners through common work packages, especially with ATGC Montpellier.

PlantBioinfoPF (http://www.urgi.versailles.inra.fr/)

Corresponding people: Quesneville Hadi, Amselem Joelle

The Plant bioinformatics facility (https://doi.org/10.15454/1.5572414581735654E12), hosted by the URGI INRAE research unit belongs to the BioinfOmics INRAE infrastructure (including the Migale and Genotoul bionfo platforms). The main missions of the platform are to contribute to an open science compatible management of patrimonial data produced by INRA and its partners and to propose tools and suitable environments (including training) for data analysis to the international community of researchers. The platform has also developed a data discovery portal dedicated to the research federation of information systems such as WheatIS .

Services include access to data in GnpIS (https://urgi.versailles.inra.fr/gnpis/), data integration of genetic, genomic and phenotypic data for INRA and international partners included the official repository of the International Wheat Genome Sequence, access to computing resources and and analysis tools with associated storage resources (associated with fees depending of volumes and partners), analysis support, organization of training sessions (Mainly GnpIS navigation and Transposable Elements annotation). The expertise of platform members (particularly in integration and publication of FAIR data, development and maintenance of data discovery portals dedicated to the research federations) will be shared with the other partners of the project. We will also be involved in rationalization of the digital infrastructure to build, articulate, structure links with national centers and regional data centers in secure locations (labeled datacenter), in the co-building of the security of data and software to manage data access, data flows and the deployment of software toward data.

PRABI-AMSB (http://www.prabi.fr/spip.php?article51)

Corresponding people: Perrière Guy, Navratil Vincent, Guyot Dominique

Partner description. The PRABI-AMSB (for Analysis and Modelling of Biological Systems) platform proposes bioinformatics services for biologists who need assistance with specific tools or in-depth expertise for more important projects. The expertise available at PRABI-AMSB covers the areas of expression data (RNA-seq), interaction data (ChIP-seq, Tap-tag/MS, Yeast Two Hybrid screens), metagenomics and metatranscriptomics, comparative genomics and phylogeny, genome assembly,annotation as well as systems biology (metabolic, protein interaction and regulatory networks) The PRABI-AMSB has also a strong experience in virus/host systems biology (http://virhostnet.prabi.fr) and is an active FAIR player in the race to better understand SARS-Cov-2 and COVID-19 disease in collaboration with the IFB task force against COVID-19 and the European Virus Bioinformatics Center (http://evbc.uni-jena.de/). The PRABI-AMSB has also an expertise in high-performance computing with the development of load balancing software (paraload). Services include bioinformatics training and hands-on (Linux, R language, NGS analysis with Galaxy, systems biology, phylogeny).

Partner involvement. The PRABI-AMSB will contribute to WP2, WP3, WP4 and IS4, IS5. The PRABI-AMSB will bring together biologists and the state-of-the-art in bioinformatics through user-oriented cases studies organised along an Environment-One Health axis. This will be done in partnership with the Université de Lyon IdEx including wet lab biologists involved in the Ecofect LabEx. The PRABI-AMSB involvement in this project will be to link users from the FR BioEnviS and the IFB Core cloud resources by creating local thematic task forces and through dedicated bioinformatics pipeline, web service, training and hands-on.

RPBS (www.bioserv.rpbs.univ-paris-diderot.fr)

Corresponding people: Tufféry Pierre, Rey Julien

RPBS is a platform dedicated to structural bioinformatics. It federates contributions from different units in Paris and Ile de France. It proposes the development of structural bioinformatics methods/protocols, service hosting and on-line deployment of services of the field, training and consulting in the field, calculation hosting through PAAS. Services include on-line RPBS services at bioserv.rpbs.univ-paris-diderot.fr and mobyle.rpbs.univ-paris-diderot.fr. Other services cover protein structure and function analysis and modeling.

South Green (http://www.southgreen.fr/) (SG)

Corresponding people: Rouard Mathieu, Ruiz Manuel, Tranchant Christine, Pitollat Bertrand, Tando Ndomassi

Partner description. South Green is a bioinformatics platform dedicated to the genomics of tropical and Mediterranean plants and related pathogens. It federates bioinformaticians from different units and institutes of Montpellier (CIRAD, IRD, INRAE and Bioversity) with a multidisciplinary expertise in data integration, software development, sequencing data analyses and high-performance computing. South Green provides free access to a wide range of original tools and information systems such as GreenPhyl, SNiPlay, Gigwa or AgroLD, and offers bioinformatic pipelines through Galaxy (ELIXIR-FR SDP), SnakeMake and TOGGLe workflow managers. The platform has also built a strong expertise in genomics data analysis and in the development of Genome Hubs (ELIXIR-FR SDP), integrated information systems, data FAIRification. A significant part of activities comprises hands-on training that are regularly offered in the local community as well as with partners in Africa and Asia on the following topics: Galaxy , NGS analyses, R, Perl, Python, Linux, HPC administration. Besides, South Green provides access to computing facilities for both users and developers engaged in this scientific area.

***Partner involvement*.** SG will contribute to WP1 (DMP), WP2 (HPC management), WP3 (Dataverse) and WP4 (workflows). Moreover, SG will co-coordinate IS1 and contribute with its expertise in multi-omic data integration for plant datasets in IS5.

France BioImaging (FBI)

Delegates: Perrine Paul Gilloteaux (Expert Research Engineer-IRHC, CNRS. MicroPiCell, Nantes); Emmanuel Faure (CRCN, LIRMM, Montpellier)

***Partner description*.** France Bioimaging (FBI) was created in 2011, in the first series of PIA-INBS. The FBI is identified as the French node of ERIC-EuroBioimaging. The FBI delegates are also the FBI's mission data officers. The FBI is mainly interested in imaging technologies, whatever the biological model of interest (from plants to human brain tumors). FBI brings together 18 large biological imaging facilities associated with specialized R&D imaging laboratories in 6 local nodes and one transversal node (https://france-bioimaging.org/locations/). The FBI provides access and expertise to state-of-the-art technologies to more than 6000 users/year nationally and internationally, through the opening of advanced microscopy systems on its imaging facilities. The transversal node, the "Image Processing and Data Management Node-IPDM '', brings together the " infrastructure strengths on "BioImage Informatics". France BioImaging's digital perspective includes three challenges based on image data: data management, Artificial Intelligence for image analysis, and image data visualization.

Partner Involvement. These challenges are addressed through close collaborations. A shared staff and a common roadmap are defined, with well-defined and complementary roles. The IFB will bring to the project its expertise in the work on national meso-centres and software deployment, in the definition of data management plans and data Fairification FBI will support the paradigm shift for Imaging facilities and users. FBI will co-lead IS1 and ensure the complementarity of the FBI. DATA and Multi4LS projects, providing knowledge of the open scientific imaging landscape. It will provide specialized staff in the implementation and operations phases. The FBI will contribute to the interoperability between the IFB, FBI-IPDM and other NRI platforms. Specific joint pilot projects (in Montpellier and Nantes) and use cases (My.EMBRC.image, with EMBRC) are defined to address the accessibility of biological imaging data and how they can be integrated with other biological data to improve scientific knowledge at the "Phenomics" level.

France Life Imaging (FLI)

***Delegates:* Michel Dojat, Régine Trébossen**

***Partner description*.** Since 2012, France Life Imaging (FLI) has been bringing together, in 6 regional and 1 thematic (Information Analysis Management: IAM) nodes, the main French centers of *in vivo* imaging research: methodological, instrumental or data analysis research, for clinical or preclinical applications. FLI helped initiate a national policy for *in vivo* imaging equipment making them available to all researchers. The current project (2020-24) integrates 3 additional poles: an East pole (Nancy-Strasbourg), a Western pole (Nantes-Rennes) and an Occitan pole (Montpellier-Toulouse). IAM provides a database management and information processing support for imaging populations involved in clinical and preclinical studies. The infrastructure (hardware and software) accessible to users of national *in vivo* platforms to store, manage and process large sets of clinical and preclinical *in vivo* imaging data and associated metadata. The project's implementation followed two stages. The first one, 2013 to 2018, was devoted to the building of image analysis and data management solutions allowing the interoperability between heterogeneous and distributed storage solutions implementing raw and meta-data indexing through the use of semantic models or ontologies. In a second operational stage, the platform will be transferred to an external partner via a contract.

Partner involvement. FLI will contribute to the interoperability between the IFB, FBI and FLI-IAM platforms. First the specific data models will study in order to allow for their alignment. If required specific extensions will be introduced for alignment facilitation. Specific use-cases will be defined to demonstrate how biological and in vivo imaging data can be accessed and fused to improve scientific knowledge.

France Génomique

Delegates: Denis Milan, Pierre Le Ber, Claude Scarpelli

Partner description. France Génomique (FG) is the national infrastructure for genomics gathering 228 engineers (178 FTE) from the two national core facilities of Evry (Genoscope & CNRGH) and from 8 additional facilities in Toulouse, Montpellier, Marseille, Nice, Strasbourg & Paris (ENS, Institut Pasteur, Institut Curie). With 25 % of time involved in R&D, FG is continually developing its expertise. In the past years, the decrease of sequencing cost induces a very large increase of dataset size (up to 6-7 10^{12} nucleotides / run). More recently, numerous new applications have emerged, with amazing developments in long read and unique molecule sequencing and with the possibility to perform genomics studies at the single cell level. FG and IFB teams have developed a very strong relationship, since FG project originally also gathered 12 bioinformatics teams, whose activity is now coordinated by IFB.

Partner involvement. France Génomique will go on to share with IFB its expertise on sequencing technologies and cutting edge approaches. They will contribute to *in silico* developments to make available the most relevant tools for genomic and metagenomic applications (IS2 to IS5). As a complementary project in the same EQUIPEX call, France Génomique will set up a new cloud node to join the IFB-core cloud federation and make its resources and tools accessible to a larger user community for sequencing data management and analysis. It will bring to IFB its experience as a broker node for sequencing data. Finally the France Genomique GeT platform will contribute to carry out as a pilot to evaluate storage and computing equipment closer to sequencers versus this same equipment hosted remotely in a labeled regional center or at TGCC national center.

National Marine Biological Resource Centre (EMBRC France)

Delegates: Erwan Corre, Bernard Kloareg

Partner description. The National Marine Biological Resource Centre (EMBRC France) is an Infrastructure Nationale en Biologie Santé operated by Sorbonne University (SU) and the Centre National de la Recherche Scientifique (CNRS). It is one of the Research Infrastructures (RI) identified in the French IR roadmap, and it is the French node of the European Marine Biological Resources Center (EMBRC-ERIC). Its mission is to provide access to marine biological resources and ecosystems to the scientific and industrial communities, both French and international. In the final evaluation of the Programme d'Investissement d'Avenir "Infrastructures Nationales en Biologie Santé" (PIA1, 2019), the Infrastructure was extended for 5 years, with the highest mark (A+).

Its services are distributed among the marine stations Institut de la Mer de Villefranche-sur-Mer (IMEV), Station Biologique de Roscoff (SBR) and Observatoire Océanologique de Banyuls-sur-Mer (OOB). The science strategy of EMBRC-France is focused on: exploration and monitoring of marine biodiversity; functioning of marine ecosystems ; domestication of marine species ; and development of tools for genotyping and phenotyping. These activities rely on the cross-referencing of genome, transcriptome, proteome, metabolome and bioimage data with reference external data, ranging from curated taxonomic databases to physical and chemical descriptors of marine environments.

Partner i*nvolvement*. In this Expression of Interest, EMBRC-France is teaming with IFB and FBI to implement the MyEMBRC-Image project (IS1). The RI will also be instrumental in the integration and dissemination of marine biology data (IS2). Noteworthly here, EMBRC-France is leading an other ESR-Equipex+ project, referred to as AO-EMBRC, on the development of Augmented Observatories, which is not overlapping but complementary to the current proposition..

FRISBI

Delegates: Yves Bourne (Yves.Bourne@afmb.univ-mrs.fr)

Partner description. The French Infrastructure for Integrated Structural Biology (FRISBI, http://frisbi.eu) offers user access to a broad range of state-of-the-art infrastructures in integrated structural biology, including sample preparation, biophysical characterization, 3D structural analysis by cryo-electron microscopy (cryo-EM), NMR and crystallography. Since 2012, the five nodes (Strasbourg, Grenoble, Montpellier, Marseille and South-Paris) pioneered a major effort to federate the national community through transversal working groups and inter-infrastructure networks, including the consultation and coordination of future strategic orientations. FRISBI reinforces the technology transfer between academic research groups and industry for technological developments. FRISBI contributes to coordinate training in structural biology through co-financing and co-organization of courses and workshops, and has played a pivotal role in the launch of the first French National training network in structural biology (ReNaFoBis). FRISBI manages more than 800 access user projects per year and more than 200 scientists are trained, and the scientific activities resulted in more than 750 FRISBI-acknowledged publications and 37 patents. As part of its essential role in defining future orientations, FRISBI coordinated in 2016 a national roadmap about the urgent need of a major national investment in cryo-EM to stay competitive in the field. This roadmap included new 300 kV microscopes at threebio national centers (Strasbourg, SOLEIL, Grenoble) and new generation 200 kV microscopes at regional sites, including upgrade with direct electron detectors.

Partner involvement. Our partnership with the IFB initiative will permit mid- and long-term secure storage and archiving at the NNCR IFB national and regional nodes to manage the massive amount of data (up to 1.5 To / day, i.e. 1.5-2 Pb / year) generated by each of the three new 300 kV microscopes of the "France-Cryo-EM" project. High-performance pipelines will be implemented to offer state-of-the-art IFB computing resources for remote data processing, and this distributed compute and storage infrastructure will be extended in the future to the other regional cryo-EM sites.

PHENOMIN

Delegates: Yann Herault (herault@igbmc.fr)

Partner description. Rodent Models are essential in research to gain knowledge on fundamental biological processes and veterinary and biomedical progress. They are essential when it comes to understanding integrative physiology or the expression of a genetic character in its environment, for example to study infection disease with a new virus. This statement is supported by many scientific academia, including the French Academy of Science. As science moves forward with ethical aspects being increasingly important, there is a growing need to reinforce data reproducibility and robustness. Whether some irreproducibility is inevitable, some factors could be worked on as the use of quality certified biological resources (scientific collection), of a controlled environment, the monitoring of variations, the best study designs, the use of cross-validated protocols with state-of-the-art equipment, with data analysis and reporting. To consider this issue, PHENOMIN has undertaken several strategies. Indeed, the creation of models, their analysis, their preservation and their distribution are processes requiring the use of certified resources (scientific collection), of validated, robust and standardized methods, based on forefront technologies, with a study design guaranteeing the quality of the modifications carried out, the statistical power and reproducibility of functional analyses, ensuring the quality of resources and the sustainability of data and models preserved for future research. This animal research must be done with respect to ethics, expressed in the 3'R rules, by reducing the number of animals linked to the refinement of tests (non-invasive equipment) and developing alternative integrative studies using isolated cells or organs.

Partner involvement. PHENOMIN-ICS will require to upgrade its current informatics infrastructure to increase the storage and compute capacity to capture raw data and to provide primary analysis. The national infrastructure "Institut Français de Bioinformatique" (IFB) will guarantee the long-term storage, archiving and availability of the data.

MetaGenoPolis

Delegates: Nicolas Pons (nicolas.pons@inrae.fr)

Partner description. MetaGenoPolis (MGP) is an unit of INRAE with over 40 people (http://www.mgps.eu/) funded by the French Investissement d'Avenir program, to establish an infrastructure able to accompany exploration of the role of the human and animal microbiome in health and disease. MGP is based on a unique set of metagenomics platforms (ISO 9001:2015 certified) focused on research, service and delivery, open to academic and private communities. MGP characterizes the gut microbiome in health and chronic diseases by quantitative metagenomics, starting with reception, storage and processing of fecal samples, sequencing of DNA, and analysis of data. MGP includes computing and storage facilities (800 cores and 3.3 PB respectively) and develops the bioinformatics and statistical softwares to integrate big data in order to extract signatures and clinically relevant algorithms from metagenomic profiles and related data. Range of activities covers varied needs, from fully automated analysis to expert data analysis for specific needs. It also develops tools and updates its databases for the analysis, enabling, among other examples, reconstruction of pangenome of bacterial species, specialized gene catalogues and specific databases (antibiotic resistance determinants of the Human microbiome)). More recently, MGP is involved in the Million Microbiomes of Humans Project (MMHP), initiating the French Gut citizen science project aiming to recruit 100,000 individuals for microbiome profiling.

Partner involvement. MGP will co-coordinate IS4 and contribute with its expertise in the characterization of the microbiome components and in multi-omic data integration for the human and animal microbiome analysis in various cohorts. MGP will play a role in the articulation between the French Gut project and MuDiS4LS as the produced data will be regularly delivered in Open Data. This microbiome dataset will be used in the different tasks of IS4. As WP4 and IS4 are linked for the AI aspects but also for the computing of large dataset, MGP will also contribute to WP4.

RARE (CRB)

Delegates: Michèle Tixier-Boichard (michele.boichard@inrae.fr)

Partner description. RARe (www.agrobrc-rare.org) is a distributed research infrastructure which gathers Biological Resource Centers (BRC) of French research institutions active in life sciences for agriculture, forestry and food (INRAE, CIRAD, IRD, CNRS) and their partners (technical institutes, higher education institutions). It has been registered on the national infrastructure roadmap by the Higher Education and Research ministry since 2016. The leading objective of RARe is to improve the national and European visibility of biological resources maintained by its constitutive BRC. Maintaining a large diversity of well documented resources, collecting new ones, contributing to their characterization, distributing them, and managing the related data, give a central role to RARe in numerous research programs aimed at exploring the living world and at making value of biodiversity for agriculture and biotechnologies. RARe maintains germplasm, to regenerate individuals or populations, genomic and biological resources for research in agriculture, food safety, dynamics of genetic diversity, plant and animal health. Producing data and giving access to data on these resources is a main lever for RARe to enhance their visibility and attractivity. RARe involves five components, also called pillars, which differ by the type of biological resources handled and by the research community using them: animal, plant, forest, microorganism, environment. This is a total of 35 BRCs with 160 full-time equivalent (permanent) and 20 temporary equivalent positions. RARe is managed by a steering committee representing INRAE, CIRAD and IRD and the executive team is in charge of coordination of all cross-cutting activities (quality, Access & Benefit Sharing, communication, joint projects).

Partner involvement. RARe is contributing to 2 projects by providing access to microbial resources and their associated data. This includes collections of defined strains and collections of microbial consortia associated either to a human or an animal host, or to a complex matrix (food, soil).

IBISBA

Delegates: Fayza Daboussi (daboussi@insa-toulouse.fr) Partners involved: Micalis (INRAE), TWB (INRAE), TBI(INRAE-INSA Toulouse), Genoscope (CEA)

Partner description. In the field of industrial biotechnology, the DBTL (Design-Build-Test-Learn) engineering cycle is widely used to describe the iterative cycle of developing catalysts with improved performance or capable of producing new functionalities: design, build, test and learn from the results generated to launch a new cycle. It is possible today to construct and screen literally thousands of genetic variants per week covering a wider range of possible solutions than was possible even a few years ago. The equipment requested by IBISBA (ALADIN project) will be mutualized with IFB equipement. It i) will facilitate the automation of feedback loops implemented in the DBTL cycle for the engineering of strains and acellular bioproduction systems, ii) will significantly increase the combinations of biocatalysts tested in each cycle, in order to significantly reduce iterations (in number and duration, and therefore in cost) and iii) will generate sufficient data to implement active learning, an artificial intelligence method that opens up the possibility of efficiently exploring large combinatorial spaces. We propose here a national infrastructure built around leading scientific teams with the necessary expertise, who will actively organize and facilitate access to such a platform.

Partner involvement. While the partners of this EQUIPEX will develop all the necessary data computational tools within the Galaxy-SynBioCAD platform, the platform will need to be deployed on a cluster offering sufficient resources for the national industrial biotechnology community. IFB will host that cluster and provide advices on state-of-the art methods regarding compartmentalization and the Galaxy system.

Inserm DSI

Delegates: Abdelkader Amzert

Partner description. Inserm brings together 15,000 researchers, engineers, technicians, and administrative staff around one common goal: to improve the health of all by advancing knowledge of life and disease, innovation in treatment, and public health research. Since its foundation in 1964, Inserm has played a part in many key medical advances, including the first prenatal diagnostic tests, understanding of the HLA system, the first in vitro fertilization, identification of the AIDS virus, radiotherapy for cancer, the first skin graft, deep brain stimulation, and gene therapy. Its mission is supported by the work of 9 theme-based institutes, whose role is to monitor progress and take a lead on research in their respective fields. The Institute is distinguished by both the scientific excellence of its staff, and by its ability to provide benchtop to bedside translational research. Inserm is the leading European academic biomedical research institution, and with nearly 12,000 publications a year; is second in the world only to the National Institutes of Health (NIH). According to the 2016 ranking by Thomson-Reuters, Inserm is also the world's 9th most innovative public research organization.

Partner involvement. Inserm's IT department designs IT services for scientific research teams based on the researcher's digital space, respecting a secured environnement for sensitive patient data storage, compliant to GDPR and national requirements (e.g. HDS certified IT environment). Its main goal is to help research teams throughout operational reform of their daily data management in order to use FAIR datasets without increasing the workload. Inserm's team will be involved in WP1 with the design of the researcher's digital space that aims to set up a data management layer (API) to collect, update and share metadata across the different steps of a research project. This data management layer will automatically interact with the maDMP. We will also be part of IS3 to implement an HDS-IT environnement related and consistent with the works of WP1.

France Grilles

Delegates: Jérôme Pansanel (jerome.pansanel@iphc.cnrs.fr)

Partner description. France Grilles (http://www.france-grilles.fr/accueil/) is a Research Infrastructure and Scientific Interest Grouping created in 2010. Its main objective is to provide human and digital services to meet the needs of processing, storing and sharing massive scientific data. This infrastructure is registered on the French national roadmap of the Higher Education and Research ministry since 2008. The services offered by France Grilles are based on a distributed computing infrastructure accessible through several technologies (computing grid, Cloud Computing, iRODS data management) and are open to all scientific communities. This technical infrastructure is provided by a network of 17 sites in France. Through the France Grilles service offer, researchers have access to 100,000 computing cores and 50 petabytes of storage. France Grilles also represents France in EGI (https://www.egi.eu) and contributes with the other member states to its functioning. Over a number of years, France Grilles and IFB have developed a very strong relationship, as illustrated by the sharing of expertise and know-how in Cloud Computing and iRODS-based storage, or by the pooling of Cloud infrastructures.

Partner involvement. France Grilles will bring its experience on two main themes. First, France Grilles will share its expertise in the field of distributed storage, in particular with the use of iRODS, its integration with an authentication and authorization framework and the dynamic placement of the computation according to the location of the data. Secondly, France Grilles will also get involved in the field of artificial intelligence, and will share its experience such as the one acquired within the European project EOSC-Pillar. This will result in the involvement of France Grilles in WP2 and WP4.

GENCI

Delegates: Stéphane Requena

Partner description. GENCI, in charge of providing high-performance computing and processing data, has the mission, at national and european level, to promote the use of intensive computing associated with Artificial Intelligence for the benefit of french academic and industrial research communities

INIST

Delegates: Claire François, Paolo Lai, Jean-Michel Parret

Partner description. Inist-CNRS provide a range of services aiming at incentivizing, training and accompanying researchers as well as data librarian, IT, data-related service providers in the construction of a FAIR ecosystem: DorANum for training, DMP OPIDoR for drafting DMPs, CatOPIDoR for identifying national services related to research data, Loterre for sharing controlled vocabularies, DataCite agency for attributing persistent identifiers (DOI). The INIST DMP OPIDoR team (9 software engineers and information specialists) ensures technical evolutions of the software and offers some training with the tool as well as advice regarding models and DMPs. In addition, DMP OPIDoR team has been exchanging with a large number and variety of stakeholders at the national (research organisms, universities, Genci, ccIN2P3, ANR, Cines) but also international level (DCC, UC3, EPFL), is fully engaged in RDA active DMP related working groups and endeavor to disseminate knowledge and information through communities so as contribute to the harmonization of best practices and application of FAIR principles.

Partner involvement. Inist will develop machine actionable DMP.

IDRIS

Delegates: Pierre-François Lavallée

Partner description. IDRIS is the French national centre for intensive numerical calculations of high performance computing (HPC) and artificial intelligence (AI) serving the research branches of extreme computing for the CNRS (National Centre for Scientific Research). In 2019, IDRIS opened a new chapter in its history by making available to the national scientific community a new generation hybrid accelerated supercomputer, Jean Zay, one of the most powerful in Europe (28 Pflop/s as of September 2020) both for the usages of numerical simulation and data processing but also, henceforth, for artificial intelligence within the framework of the plan "AI for Humanity" launched by French President Emmanuel Macron in March 2018. This extension in usage parameters and missions of IDRIS makes Jean Zay the first national platform for the AI research community. IDRIS serves and accompanies a community of users consisting of more than 1300 researchers and engineers working on approximately 450 projects from all scientific disciplines by offering a very high quality applied support service (accompanying, advice and expertise). In addition to its national missions, IDRIS is hosting resources (IFB Core Cluster, Ruche and Fusion machines of the Moulon CentraleSupelec Mesocentre, ENS [Ecole Normale Supérieur] Paris-Saclay), and is thereby strongly rooted in the great Paris-Saclay campus project currently being constituted. Lastly, IDRIS is highly implicated in the current construction of the European ecosystem of intensive computing as a partner in different European projects, including PRACE (www.prace-ri.eu), EOSC (https://www.eosc-portal.eu/) and AQMO (http://aqmo.irisa.fr/fr/le-projet-aqmo/).

Partner involvement. IDRIS has a long-lasting collaboration with IFB, as the hosting site for IFB core cloud (2013-2016) and IFB-core-cluster (since 2017). In the framework of the MuDiS4LS project, 2 additional IFB platforms located in the région Ile de France will migrate their computing facilities to IDRIS. Moreover, IDRIS will play a key role by providing access to the Jean Zay supercomputing facility.

CINES

Delegates: Boris Dintrans

Partner description. CINES (National Computing Center for Higher Education) is a French public institution, located in Montpellier (south of France) and supervised by the French ministry for Higher Education and Research. CINES offers services to the scientific community through national strategic missions :(1) High performance computing; (2) Long term digital preservation; (3) National hosting computer platforms. To achieve these missions, CINES is equipped with a high-level secured infrastructure (in particular 2 general electricity supplies : 2.6 MW and 10 MW), complying with the new requirements for environment footprint, power offer and cooling capacity. Besides these two main missions, CINES offers a secure hosting for national servers operating strategic applications for the higher education and research community. This activity takes advantage of the high level infrastructure. About 60 people work at CINES. These include technical teams of engineers and experts responsible for the operation and optimal use of the computing resources as well as for training and user support.

Partner involvement. The CINES is involved in IS3 "Bioinformatics for health", where it will play the role of secure infrastructure to host health data.

DROcc - CALMIP (University of Toulouse)

Delegates: Hervé Luga

***Partner description*.** The University of Toulouse, created in 1229, is a comprehensive and research oriented university clustering all higher education bodies and research organization in the south west of France. The University of Toulouse is a world-class academic institution, renowned in many areas such as economics (Nobel Prize 2014), aeronautics & space, health, engineering, cognitive sciences, A.I., archeology, tourism, Agriculture & hospitality… The University of Toulouse gathers many research and Higher Education Institutions throughout the Midi-Pyrenees region. Over 100.000 students are enrolled in its 23 higher education member institutions, 7 research organizations and 1 University Hospital. Its research forces rely on its 143 laboratories, its excellence recognized by PIA instruments (LABEX, EUR) and its artificial intelligence institute 3IA ANITI.

***Partner involvement*.** UFTMip operates the datacenter where the equipment of Genotoul Bioinfo will be hosted in the next five years. UFTMIP will provide INRAE and Genotoul the infrastructure necessary to this hosting. Altogether, the CALMIP mesocentre, UFTMIP and Genotoul Bioinfo will set up computation and storage resources on which they will host/operate service offers including those of Genotoul Bioinfo for Life sciences. This partner will bring its experience in HPC, system architecture, and storage to build a service offer in a secured place with the objective to mutualize a regional data lake for data access and sharing. They will also work together on intersite replication. This contribution takes place in WP2.

DROcc - MESO@LR (University of Montpellier)

Delegates: Anne Laurent

Partner description. In Occitanie, digital infrastructures and services are currently being put under the umbrella of the common Data Center Régional Occitanie (DROcc). 2 mesocenters are working together for the Mesonet project at the level of Occitanie: CALMIP and Meso@LR. Operated by the University of Montpellier, the Meso@LR platform (formerly HPC@LR) offers shared resources and advanced architectures for HPC and HTC. Open to companies and academic, public and private stakeholders, it acts as a proactive player of the French Tech in Montpellier. It also contributes to the strong dynamic of the Montpellier site around ISITE MUSE, especially in the fields of Agriculture, Environment and Health, for addressing the MUSE three major intertwined challenges: Feed, Protect, Care.

Meso@LR currently offers 308 nodes (8624 cores + 2 SMP Nodes (3To RAM)). In 2020, a massive storage infrastructure is being implemented. Computing resources are accessible in a flexible way, either on the fly, either by the hour or by providing dedicated and secure environments. The technical resources of the mesocentre are hosted at CINES. The majority of Meso@LR users are located in Eastern Occitanie, with a national opening. Meso@LR works closely on links to data science and artificial intelligence, especially in the framework of the Institut de Science des Données de Montpellier (ISDM). Moreover, particular interest is paid to teaching activities and capacity building, in the framework of practical work, workshops and a training cycle operated by the mesocentre. Meso@LR has received support from the OCCITANIE / Pyrénées-Méditerranée Region and from Montpellier Méditerranée Métropole as part of the 2015/2020 CPER project.

Partner involvement. UM operates the Meso@LR mesocentre, which is currently linked with ATGC for part of the needs and aims at strengthening links with Southgreen. Meso@LR will provide infrastructures for storage and computing. It will closely work with DROcc-CALMIP in order to build a federated data lake working on both horizontal and vertical axes for optimizing resource consumption. Meso@LR will bring resources and experience in data and computing management, together with experience in multi-partner data/HPC resource governance.

GLiCID (Nantes Pays de la Loire)

Delegates: Yann Capdeville, Yann Dupont

Partner description. With a goal of pooling material and human resources, GLiCID merges the current efforts of 5 structures in the Pays de la Loire region into a single entity with equipment that will ultimately be physically located in the same data center. The BiRD core facility and the CCIPL are two of these 5 structures. The Centre de Calcul Intensif des Pays de la Loire (CCIPL) is a regional parallel computing centre of intermediate size of Tier 2 or mesocentre type. Its missions are to meet the high-performance computing needs of scientists from all higher education and/or research organisations in the region, within the limits of its resources; to facilitate the pooling of laboratory equipment; to provide training and development assistance in scientific computing for its users. The CCIPL has a long experience and expertise in terms of administration of intensive computing machines (CPU and GPU for artificial intelligence) and storage of scientific data. For more than a year, the CCIPL and the BiRD core facility and CCIPL have become closer and their respective infrastructures are in the process of merging. Altogether, GLiCID actors currently cumulate 11,000 compute cores and 1.2 Petabytes of fast storage and 1 Petabyte of long-term storage. The computing resources offered by the various project member entities are currently promoted by more than 150 publications per year, some sixty labelled projects (including 24 ANR), public-private contracts/partnerships, patents and/or software.

***Partner involvement*.** The GLiCID structure aims to increase the scope of laboratories with access to computation and to provide the resources necessary to offer services based on computation, such as, for example, services developed by BiRD for life sciences or platform projects facilitating developments around AI, particularly for Bioimaging. GliCID will particularly contribute to WP2 with its experience in the implementation of secure storage solutions for scientific data. It will participate, through the expertise of the BiRD core facility and CCIPL engineers on distributed storage, to the implementation of mid-term storage in a secured place and inter-site replication. The engineer, system and network administrator of the BiRD core facility and member of NNCR, is already involved in structuring projects with GlyCID engineers. GlyCID will also be part of IS1, in connection with the FBI infrastructure, around the management of Bioimaging data, and artificial intelligence for image analysis. More specifically, it will be in charge of making these data available and deploying these analysis methodologies.

CBP-PSMN (HTTP://WWW.CBP.ENS-LYON.FR, HTTP://WWW.ENS-LYON.FR/PSMN)

Delegates: Ralf Everaers (CBP), Hervé Gilquin (PSMN) and Olivier Gandrillon (LBMC)

***Partner description*.** Together, the PSMN and the CBP play the role of a resource and competence center in computational science at the ENS Lyon and complementary services to the specific capacities of the LBMC, RDP, IGFL, CIRI, SFR BioScience, PLATIM, BioAster and other ENSL associated laboratories and platforms across all disciplines. The PSMN is a high-performance computing center providing access to resources acquired by the school and ENSL research laboratories or through European, national or regional grants and which are, for the most part, mutualized between all users. The CBP was set up as one of the first "modeling houses" in France with a support mission for research, training and scientific animation in the field. It disposes of offices, seminar rooms and a digital infrastructure (lab room, servers, software forge etc.) for collaborative work. In particular, the engineers of the CBP develop and test innovative software and hardware solutions before they are deployed in the PSMN. The two structures currently manage more than 700 servers (more than 15,000 cores and more than 100 graphics accelerators) installed in the state-of-the-art Datacenter of the ENS de Lyon, which was inaugurated in December 2017 and is part of the CINAuRA regional data center. Equipped with a 1.2MW power supply, it houses all the scientific IT resources of the ENS de Lyon in 66 bays. The engineers of the CBP and the PSMN ensure the maintenance of the clusters, the software installation, as well as training and support for 500 researchers and an equal number of master and bachelor students.

***Partner involvement*.** The CBP-PSMN are involved in the IFB cloud federation and provide IFB members with access to their local cloud resources. CBP-PSMN will participate in the planned distributed storage network and inter-site replication (WP2), and associate data treatments through pre-defined virtual environments as virtual machines or containers available on their cloud site. The CPB-PSMN will also contribute to the HPC/IA activities (WP4) by providing application developers with a benchmark and validation environment. Developers will be able to evaluate their applications on a large variety of different hardware configurations before deploying them throughout the IFB cloud and on national resources. CBP-PSMN will deploy the required tools, data and workflows, and bring their expertise in HPC computing to support and train the community, especially with respect to GP-GPU intensive computing.

PIA3 / Appel à manifestations d'intérêt ESR-EquipEx+ / *Call for proposals ESR-EquipEx+*

2020

MuDiS4LS

Document de soumission / *Submission form*

Mutualized Digital Spaces for Life Sciences (MuDiS4LS)

Appendices

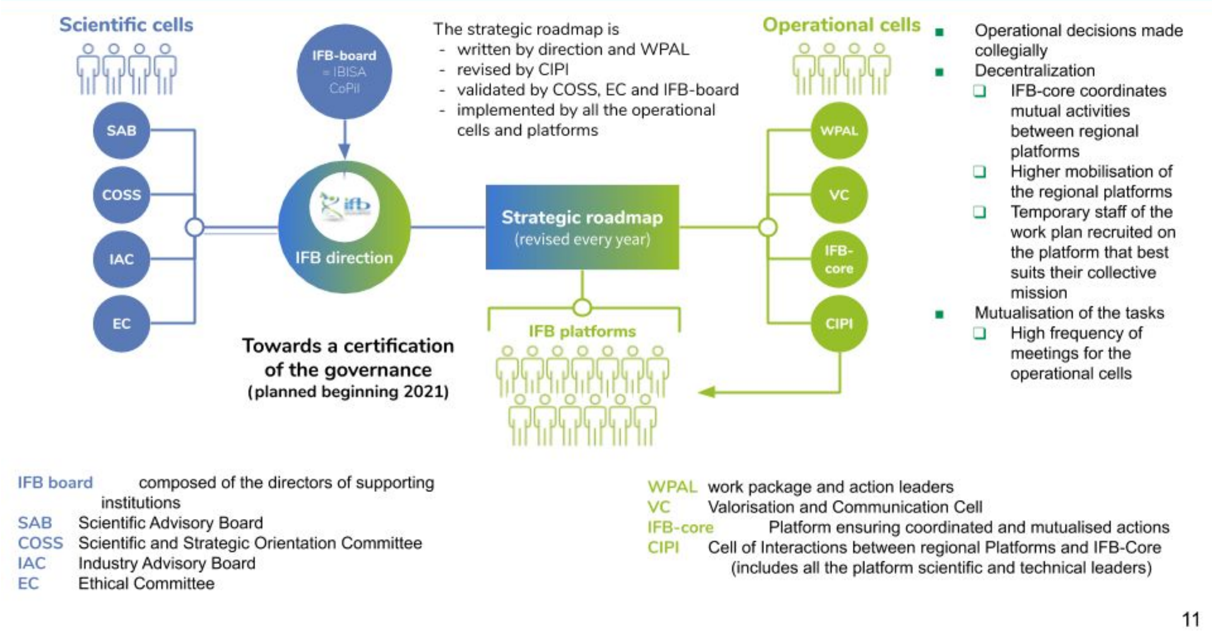
A1. IFB 2017-2018 Activity report

The 2017-2018 Activity Report (https://doi.org/10.5281/zenodo.3520131) contains the following elements.

1. A panorama of the activities undertaken since the restructuring of the national infrastructure
2. The active collaborations between IFB and the data producing infrastructures during the pilot projects in integrative bioinformatics, which contributed to ground the MuDiS4LS project.
3. Our activities in the European bioinformatics infrastructure ELIXIR
4. A detailed descriptions of each IFB platform
5. Indicators for 2017 and 2018

A2. Structuration and evaluations of IFB

A2.1. IFB governance scheme



The strategic roadmap is
- written by direction and WPAL
- revised by CIPI
- validated by COSS, EC and IFB-board
- implemented by all the operational cells and platforms

- Operational decisions made collegially
- Decentralization
  - ❑ IFB-core coordinates mutual activities between regional platforms
  - ❑ Higher mobilisation of the regional platforms
  - ❑ Temporary staff of the work plan recruited on the platform that best suits their collective mission
- Mutualisation of the tasks
  - ❑ High frequency of meetings for the operational cells

IFB board composed of the directors of supporting institutions
SAB Scientific Advisory Board
COSS Scientific and Strategic Orientation Committee
IAC Industry Advisory Board
EC Ethical Committee

WPAL work package and action leaders
VC Valorisation and Communication Cell
IFB-core Platform ensuring coordinated and mutualised actions
CIPI Cell of Interactions between regional Platforms and IFB-Core (includes all the platform scientific and technical leaders)

11

A2.2. Report from IFB SAB (March 2019)

Report of the Scientific Advisory Board for the Institut Français de Bioinformatique (IFB) 2019

Scientific Advisory Board (SAB) members:

Søren Brunak (chair), University of Copenhagen
Amos Bairoch, University of Geneva and SIB - Swiss Institute of Bioinformatics
Christine Orengo, University College London
Anton Nekrutenko, Penn State University
Lodewyk Wessels, Netherlands Cancer Institute (absent at the 2019 meeting)
Ana Conesa, University of Florida

Date of the SAB meeting: 26–27 March 2019

Background and summary of conclusions

The French Institute of Bioinformatics (IFB) is a national service infrastructure in bioinformatics. It was established following the award of a proposal in the 'National Infrastructures in Biology and Health' call. During 2017 and 2018, following a review by the Agence Nationale de la Recherche (ANR), IFB underwent a major restructuring, for example installing in May 2018 a new leadership whereby Claudine Médigue and Jacques van Helden now act as co-chairs. The SAB meeting thus evaluated the achievements of a new, restructured IFB project that have been delivered in less than a one-year period. The SAB found that the new leadership clearly had restarted the project in a very powerful way and that it now, much more than before, involves numerous highly competent stakeholders across France. Many of these contributors gave excellent presentations at the SAB meeting and engaged in discussions. The goal of the IFB is to enable the development of research infrastructure in biology and health at the national level and to facilitate the implementation of the associated national and European roadmaps. The SAB found that significant improvements have been made towards this goal by creating a federated infrastructure of clouds and clusters from local resources. IFB has also contributed a tremendous national effort in bioinformatics training in this short period of time. IFB clearly plays a very active role in the European ELIXIR organization as well. Some branding of the new IFB project is still in process. Overall the SAB was impressed by the developments and commends the new IFB leadership in getting the new organization into a coherent and productive mode in such a short period of time. This momentum should be used to expand further the vision and strategic goals for IFB.

The SAB made a number of observations across the areas presented to the board. In the report the SAB decided to focus on constructive comments that can be used to further improve the value of the infrastructure.

Services to the communities

IFB attempts to bring together a collection of nationally distributed hardware resources. These resources consist of five academic clouds as well as six conventional HPC servers, the IFB clusters. Such decentralization is necessary for ensuring effective use of existing resources and for addressing the full spectrum of distinct types of computational analyses. IFB is to be commended on bringing together so many resource providers across France. Not only has it allowed them to establish a comprehensive facility that can probably meet a wider range of local needs than a single facility, but it is a mechanism for collaboration within IFB of multiple platforms across the regions, which is likely to lead to a much wider knowledge of the IFB in both bioinformatics, medical and biology communities and a much greater buy-in of IFB activities across France.

The cloud system, BioSphere, has a unified dashboard allowing members of French research institutions to launch a variety of virtual machines instances. The current allocation has hard limits of 100,000 SU of CPU time and 1 PB of "hot" storage (~400 TB on the IFB-core-cloud) and it is unclear how these can be extended. While it is highly important to have dedicated bioinformatics resources like the ones IFB organizes, it is also important to relate in the forward going strategy to other national supercomputer resources, and to emerging international initiatives, such as EuroHPC. On the shorter term it could be relevant to invest further in GPU resources to support the increasing use of machine learning in the bioinformatics domain.

Training

The SAB was impressed by the training efforts undertaken by the IFB. Training activities both at the level of platforms and in the curriculum of major French academic institutions are key to the future of bioinformatics activities in France. In this aspect the IFB has been highly active in federating and organizing training activities of various kinds. These activities are fully integrated with the various ELIXIR training programs thus taking advantage of training efforts organized across the European countries. We believe that these impressive training activities need not only to be continued but even expanded as there is a high demand in life science research groups by both biologists and embedded bioinformaticians in acquiring the knowledge necessary to understand the scope and challenges of modern bioinformatics data analysis. The recent organization by the IFB of a very successful European-wide BioHackathon is also something that exemplifies the value of the IFB training activities. The SAB recommends evaluation of the long-term impact of these training activities (i.e. surveys one year after completion of the course) and specific actions for training of health professionals in relevant emerging areas of precision medicine.

Innovation - towards integrative bioinformatics

One of the missions of the IFB is to promote innovation in bioinformatics. IFB has chosen to establish an innovation axis in integrative bioinformatics, which has a goal to promote the integration of different omics platforms such as metabolomics, genomics, proteomics, etc. IFB launched a call for projects in integrative bioinformatics that was well received and selected a number of pilot projects that are currently running. The SAB evaluate positively the initiative to choose a specific innovation axis where the infrastructure organization can concentrate efforts and make a significant contribution to the state of the art. The IFB is committed to maintain this effort. The SAB felt that this was a good model that could potentially increase uptake of IFB tools and resources and publicize the activity of the IFB to the wider community. However, despite some progress on this difficult task, the SAB felt that a clearer vision was needed as to what should be achieved in this area of integrative bioinformatics, who to involve, and what the roadmap for this goal will be. Although some timing of actions has been proposed, we do not see that they follow a clear strategic plan. Recommendations are made as to clearly define this innovation goal, the strategic plan and the roadmap for achievements. This exercise will require of an evaluation of the state of the art, the position of France within integrative bioinformatics and the logical course of innovative projects. Ideally, this should be started by addressing interoperability needs, platform communication and databases, and then move into new integrative methods and tools to finally reach applications to the health system and the bioindustry. Also, this plan should include actions to put in contact different stakeholders in genomics to facilitate the creation of integrative projects.

EU integration

The SAB found that IFB had taken part in the European ELIXIR efforts in a strong way, and that the involvement in eight implementation sub-projects provided evidence that France has a lot to offer in the European context. Similarly, as already remarked above in relation to the training effort IFB has played a major role at the European level. This was also clear in the context of tools infrastructure and in other interoperability efforts. In particular the Hackathon organised in Paris last year had clearly been successful and the fact they had been requested by the ELIXIR platforms to hold this in Paris again was a clear recognition of IFB's major role in successfully leading this initiative.

Interaction with industry

The interaction with industry is relatively widespread and as many as 80% of the IFB partners are involved in industry collaborations. However, the SAB felt that IFB could put this aspect more positively by landscaping these collaborations and the contacts they provide. By collecting information on tools/services/training provided to companies and perhaps building a catalogue they can ensure that all industry contacts are aware of resources available from the platforms etc. The SAB noted that a stronger involvement of pharmaceutical companies is a possibility and that it was unclear to what extent that sector has a strong involvement with IFB.

Interaction with healthcare providers

The SAB noted that there seems to be a large unexploited potential for more interaction with hospitals and other healthcare stakeholders. While there may be hindrances that relate to person-sensitive data, such barriers do not apply to workflows that are used in genomic medicine or in clinical proteomics, or to efforts on reference genomes, common and rare variation of relevance for the diverse French population. In many countries there are close interactions between academic genomics environments, healthcare and industry. In 2016, INSERM and its Aviesan partners drafted an ambitious plan (Plan France Médecine Génomique 2025) to ensure that France should play a major role in the use of whole genome sequencing in clinical practice. We believe that this plan should capitalize on the competencies of the IFB. In particular, we are happy to learn that there are ongoing discussions on how the IFB can play an important role in this context. We also believe there should be a connection between the IFB innovation axis in integrative bioinformatics and the efforts to involve IFB in the national genomic medicine plan.

IFB communities

The IFB has developed an ambitious programme to establish a network of bioinformatics resources which includes a federation of computer clusters and clouds. Whilst challenging, this has several benefits including agile response of regional platforms to strong local user needs and development of specific infrastructures and solutions that could then be shared across other platforms where appropriate.

The SAB felt that the creation of working groups collecting information on the needs for specific user communities eg. in health, agriculture, fundamental biology, environment and microbial biotechnology, was a positive move that will enable IFB to respond to particular requests from these communities by identifying the appropriate partners and assessing necessary allocation of resources. For example, the survey of INSERM laboratories is a good example of how IFB plans to landscape the requirements and then plan for tools development, services, training etc. This will be rolled out to other organizations. The activity in itself will also have the benefit of alerting the user communities to IFB capacities and will allow them to make a stronger case for the need for IFB and the growth of its facilities.

Furthermore, their plans for surveying the bioinformatics needs of the other national infrastructures (INBS) and devising mechanism to share their resources with these infrastructures will promote IFB value to an even broader community and ensure wider uptake of their tools/facilities. It will also bring significant benefits by ensuring good practices in other institutes and may facilitate the identification of potential collaborators for their challenges in integrative bioinformatics.

Funding, IFB economic model

Due to the complexity of the French research funding mechanisms the full cost models that were presented to the SAB are quite opaque in terms of what will be the strategy of the IFB to ensure that the needs of the life science research community will be met. However, what is clear to the SAB is that based on our experience of what other countries have achieved or are currently setting up there is a huge financial gap between what is planned for the next 5-10 years and the real needs. This should not be seen as a lack of vision by the direction of the IFB but rather the almost impossible task that they have been asked to carry out: estimate the full costs of activities that are under the responsibility of 8 different “tutelles” whose budget allocation for bioinformatics activities are either not known or at least not available. The funding of French bioinformatics resources (both data resources and software tools) that are part of the French ELIXIR service offer needs to be planned. One of them (Orphanet) is an ELIXIR core resource and is well funded, the others are not but at least some of them are widely used and internationally used. There is therefore a need to insure the sustainability of these precious resources.

It should be emphasized that France not only needs to find a way to fund the exponential needs of the French life science research community in bioinformatics analysis but also provide a mechanism to continue to fund bioinformatics research activities. This mechanism should consider the three fundamental pillars of bioinformatics: service to users, transference of bioinformatics developments into sustained tools that can be used by the community, and bioinformatics innovative research. As the IFB manages a federation of resources, each with different funding rules, it is not clear how all this can be reconciled into one funding model (for example, some institutes can invoice third parties, some others not). Some resources are paid by regional funds and might want to treat differently users from different parts of the country. This requires a serious and complex financial study that is completely missing. This detailed funding study that results in a clearly sustainable funding model is necessary to guarantee the future of IFB. Obviously, different funding mechanisms are possible. One is that the IFB receives significant funding and provide free-of-cost services to the community. Another is that IFB does not get strong funding but creates fees for their services and these feeds should be built into the research projects of the different researchers/institutions that will require the service. In any case, the funds would probably come from the same source, but the numbers must be made, and the users of the French bioinformatics infrastructure need to be informed of the model. Hence, the financial plan must disclosure all these aspects of the IFB mission and activities. Finally, there is no consideration of the growth that data-based medicine and economy will certainly have in the future and this is awkward. This needs to be considered and the economic needs calculated accordingly.

Branding of IFB

For an organization like IFB branding is extremely important. The community needs to build trust in IFB as a resource provider, and in particular in the stability and robustness of infrastructure elements that their work depends on. It is therefore important to run a web-site that clearly describes the organization, the opportunities and services. Even if the new IFB management has gone far already in building this trust, the SAB found it problematic that the IFB web-site was not updated as to reflect the new structure post May 2018. At the SAB meeting the web-responsible was present but provided no explanation for the missing update of the web-site, that, for example, did not describe the governing bodies, such as the IFB Board, with the correct membership.

A2.3. Report from the international jury evaluating French National Research Infrastructures, June 2019

Evaluation: ReNaBi-IFB

1. Scientific excellence of the project

1.1. **Achievement of the proposed objectives, potential evolution and compliance with the recommendations made in 2016**:

The plan of action for the infrastructure was totally restructured after the negative mid-term evaluation. A new governance system was established with lead of two co-coordinators. The new road map has new work packages directed to 1) distributed national environment of bioinformatics basic services for biologists from all fields of life sciences, and 2) development of innovative bioinformatics service (now targeted for integrating data of heterogeneous nature) 3) links with industry and international networks, 4) training and outreach, and 5) management and outreach.

The funding was frozen until the new work plan was accepted in May 2018 - so the current evaluation covers a period of one year. The changes in the work plan provide adequate responses to all of the concerns from the mid-term evaluation, and substantial progress has been made. The IFB IT environment relies on a mutualized task force from 12 IFB platforms. The cluster component has been implemented and available to the user community with a high effectiveness.

1.2. **Quality of the project presented for the forthcoming period, including expected scientific added value and its international visibility**:

The funding is secured until 2021 due to temporal freezing of funds after mid-term review. The quality of the plan for the forthcoming period is good, aiming at meeting emerging user needs, promotion of the infrastructure nationally and internationally and continuing developing the IFB as the French node of ELIXIR, as well as interactions with the other INBS and ESFRIs. New planned developments include: moving software to the data (cloud computing), providing expertise for FAIRification of data (support DMP), applications of AI to integrative bioinformatics.

2. Scientific structuration promoted by the project

2.1. **Added value of the infrastructure for the scientific community**:

Software development and deployment have been an important activity and the number of developed databases and tools as well as shared ones have been increasing and are currently available for the community. Training events have also been successful both for users and trainers. Total number of users increased threefold in 2018 compared to 2017.

2.2. **Alignment/integration in the strategy of academic research institutions**:

Alignment seems to be strong. The permanent staff in the platform related institutions take actively part in work package tasks. The local platforms are directly integrated into the respective institutions. In addition, IFB is collaborating with a large number of academic institutions, and the scope and quality of collaborations is visible in the large variety of high quality publications.

2.3. **Socio-economic impact (impact for public decision-makers, patents and licences, expertise performed, industrial partnerships and contractual collaborations established)**:
Impact in terms of publications and training is high. Patents are not typical in this field, but valorization of IBF workflows and software could be improved. The regional platforms are aiming at developing partnerships with industries. 32 projects in 2018 included industrial partners, and a significant number of users trained were from industries. Clear governance with central contact point is a very good idea to enhance visibility (coordinated actions, lobbying), and access by externals incl. industry.

3. Consolidation of the infrastructure for the 2020 --2024 period and beyond

3.1. **Organisation and governance with the perspective of sustainability (without additional funding by PIA beyond 2024): self-assessment capability, funding mechanism of the infrastructure, valorization policy**:
The new governance concept leaves a competent impression. There is a clear vision on the requirements towards sustainability. During the last years the platforms have tried to develop their economic model. They have identified potential resources for the renewal and evolution of equipment (tarification of storage and computing, national and international collaborative projects, support from the supporting research institutions and support from the French Regions). Priority actions to 2024 are envisaged, accompanied by indicators as: increasing the rate of self-funding of platforms (minimum 5%), targeting national and European calls (target 10%), developing industry partnerships, identifying non-sustainable resources (for which support must be coming from elsewhere)

3.2. **Budget request and involvement of partner institutions**:
The budget request 2.8 M€ for the extension period (beyond 2021) is modest, calculated based on current level of budgeting. They however expect to have an increase in costs during the coming years (suggested by the SAB).

3.3. **Sustainability after 2024**:
Seems to be possible, IF there is no substantial increase in costs.

4. Global assessment

4.1. **Main strengths:**
Organization is structured to optimize sharing of knowledge. The program is well positioned across life science fields and they have good plans to respond to users' needs. It is a coherent national infrastructure integrated into international community (ELIXIR). Broad utilization of resources and expertise, high number of high quality publications; development of algorithms, open-source software; training of demanded tools (e.g., galaxy)

4.2. **Main weaknesses:**
Funding of bioinformatics is commonly kept at a low level in biology projects-applications. There will be a need for continued funding of platforms – sustainability by user fees is "unrealistic".

4.3. **Overall comments and recommendations**:

This is an important infrastructure, very much needed also in the future by the scientific community to keep up with the rapid evolution of the field and the growing need of user support and training. IFB is well connected with other infrastructures within PIA. Connectivity will likely further increase as those other infrastructures realize the benefits achievable for their respective own impact.

A2.4. Report from ELIXIR SAB, Feb 2020

ELIXIR NODE PERIODIC REVIEW SUMMARY FORM

The ELIXIR Review Summary form is for use by the ELIXIR Scientific Advisory Board (SAB). The purpose of the form is to capture the SAB's views and comments on each ELIXIR Node presented for periodic review.
This document is confidential and is not intended for public circulation.

Name of Node	*ELIXIR France*
Date	*11 February 2020*

For information:

Characteristics of ELIXIR Nodes
The information below was given to the Nodes to help them complete the periodic review form. It details the services that an ELIXIR Node may provide.

Nodes providing data resources
Where the Node institute(s) provide generally accessible data resources - databases, knowledgebases, repositories - these can be included in the Service Delivery Plan. This applies to mature, established data resources such as the ELIXIR Core Data Resources and Deposition Databases, as well as to more specialised data resources that are only needed by a small part of the community, but are nonetheless of significant scientific importance.

Nodes offering compute provision
Nodes including major computing centres such as High Performance Computing, virtual machine hosting and cloud storage provision may provide large-scale resources for data and computationally intensive tasks needed by ELIXIR scientists and their collaborators. ELIXIR also works on supporting services for user access, e.g. AAI, or deployment of bioinformatics tools and workflows. Such resources support rapidly evolving software environments and reproducible workflows for biological and biomedical data analysis, and provide computational access to large reference data protected by the ELIXIR infrastructure.

Nodes offering training provision
Nodes that have the appropriate training expertise and facilities will provide training to developers, researchers and trainers throughout Europe on bioinformatics resources provided by the ELIXIR Node or on use cases based on the Node's specialty. Training can be provided through regular training events and/or an eLearning infrastructure. Training providers will implement the ELIXIR training Toolkit, which includes the registration of the training event and publishing reusable (FAIR) training materials in the ELIXIR training portal TeSS, and the deployment of the short-term feedback and long-term feedback surveys after the event. It is likely that many Nodes will be involved in providing training, specialising in their particular area of expertise and combining forces to develop curricula and deliver a comprehensive suite of training tools and courses.

Nodes offering tools infrastructure
Nodes may wish to deliver software tools that they have developed, whether standalone packages, workflows or web services. We encourage Nodes to adopt software best practices when developing their tools and include a plan that address tools sustainability. Nodes will need to provide the capability for users to discover their tools in the bio.tools

registry and enable potential benchmarking of their tools within OpenEbench. We encourage the Node to facilitate ease-of-use and interoperability when developing software and also deposit relevant training material associated with the software within the TeSS training registry.

Nodes providing interoperability and standards infrastructure
A Node offering interoperability infrastructure should provide capability to address FAIRification activities in one or more of the following: interoperable programmatic access to reuse databases and tools; biological and medical nomenclature; services that enable tools and data to be interoperable; controlled vocabularies and ontologies; and/ or reporting requirements for data deposition and exchange.

For completion:

Organisation(s) involved in the ELIXIR Node

Note: Provide a short summary of the appropriateness of the organisation(s) involved to continue carrying out the services provided by the Node in the context of ELIXIR. This should include the relevant experience and track-record in providing the services to ELIXIR. Consider Node's status with regards to gender balance and diversity. Has the Node taken steps to ensure that compliance with the ELIXIR ELSI Policy is monitored?

The French Node is a large and highly distributed group that seems to be operating well, especially given the very varying institutes and components of this distributed Node. The SAB is pleased that the Node has set up objective criteria for selection of Node service that opens selection to any French group. This has also proven valuable as it allows groups that are not successful in their application to determine what is still required for them to be in a position to provide their resources as an ELIXIR Node service, and to get support to get to achieve this.

The SAB also especially commends the Node for its efforts in surveying its national users to aid in identifying their needs.

The SAB commends the Node in its efforts to learn from a failure in Node-industry partnerships through the formation of an Industry Advisory Committee.to make future recommendations and to oversee the process of collaboration.

The SAB further commends efforts towards gender balance and diversity, recognizing that a vigilant effort has to be maintained.

Impact of the Node Nationally and Internationally

Note: Provide a short summary of the impact of the ELIXIR Node. This could include:

- *The scientific and technical impact*
- *If relevant, the uniqueness of the services and the extent to which the services have been adopted by the community (i.e., pervasiveness)*
- *Any positive impact of the Node on the national bioinformatics community; and*
- *Any economic and societal impact*

The SAB commends the French Node on its development and evolution, especially given the early challenges in building a national bioinformatics infrastructure. It is noted that this development was encouraged and supported by ELIXIR, thereby aiding in the assembly of the national bioinformatics community and government funding for the Node.

The Node has now established excellent core infrastructure to support the French community. This includes compute and (limited) storage capacity, consulting and support, training, and software development. In future long-term data management and storage will be an area that the Node needs to address.

ELIXIR Impact

Note: This could include the added-value of the Node to Europe and ELIXIR and the impact of ELIXIR on the Node.

The SAB is pleased to see the French Node's collaboration with other ELIXIR Nodes, especially within the tools platform. The Node also has a leading role in the ELIXIR tools platform and in particular its role in the GALAXY and EU Hackathon are to be commended.

Keeping in mind its maturity, the SAB feels that the French Node is in a good position to help mentor smaller Nodes.

The French Node will benefit from ELIXIR-Norway in its establishment of a local EGA mirror.

Finances and resources committed by the Node

Note: Provide a short summary whether the resources committed (staff and budget) will be necessary and sufficient to enable the ELIXIR Node to continue to carry out its proposed services satisfactorily.

The French Node seems to be well supported by government funding until 2025.

Final summary and recommendations

Note: Provide a final summary of the ELIXIR Node as a whole. If necessary, please provide any suggestions for improvement or modification, including recommendations regarding significant gaps and opportunities, which could be explored.

The ELIXIR SAB congratulates the French Node on their success in establishing the Node and concludes that the periodic review of the French Node has been successfully conducted and has no further recommendations to make.

A3. Detailed description of the Work Packages and Implementation Studies

WP 1. Orchestrating data flows for life sciences

WP coordinators

Frédéric de Lamotte, Gildas Le Corguillé

Short description

This WP will focus on data management and stewardship for data hosted on geographically distributed IFB platforms. Since IFB acts as a hub for data originating from various sources and institutes it is of utmost importance to provide data access coping with scientific and institutional constraints. However it is technically challenging and costly to provide rich and complete enough metadata required for better data access and reuse. As IFB members are in contact with a large number of scientist communities, IFB is in a prime position to bring about a change in the attitude of its users towards the central role of efficient research data management. The aim of this WP is (i) to lower the cost of populating DMPs through FAIRifying the data from the first day and (ii) to provide incentives for researchers to produce and share qualified data for open and reproducible sciences.

Specific tasks

1. Developing procedures relying on machine-actionable DMPs to enable a swift management of the data fluxes between data producing infrastructures, computing facilities, and repositories.

2. Instrumenting the data and computing infrastructure to capture metadata (e.g. provenance) and feed maDMP thus lowering the human cost of maintaining and allowing the automatic update of data management plans during research projects lifecycle.

3. Disseminate the maDMP towards data-producing national infrastructures for life and health sciences, in order to ensure a FAIR data management from the first day.

Use cases

U1.1. Leveraging maDMP4LS (Machine-actionable DMPs for Life Sciences) to ensure compliance with GDPR for personal data or formalized organizational constraints (in close collaboration with WP2). (organization)

U1.2. Annotation of raw Life Science data, based on provenance and research context metadata captured from INBS and IFB computing facilities and populating a global life-science data provenance hub. (in close collaboration with IS3). (machine)

U1.3. Interconnection between electronic laboratory notebook and maDMPs, so that the many research partners are dynamically aware of used and produced data artifacts (datasets, softwares, training material, etc.) and eased in their data curation and publishing activities (in close collaboration with WP3). (humans)

Deliverables

D1.1: Software tools ensuring the seamless interoperability between the NNCR computing infrastructure and maDMP platforms.

D1.2: Life science provenance metadata knowledge hub.

D1.3: Inter-connected electronic laboratory notebooks with maDMPs.

D1.4: Guidelines for users to promote the usage of DMP platforms.

Partner INBS

ProFi, MetaboHub, France Génomique

Other partners / collaborators

INIST; Inserm DSI

WP 2. A Distributed data infrastructure for project-life-long secured storage and backup

WP coordinators

Guillaume Seith, Olivier Sallou

Short description

This WP will lay down the physical infrastructure underlying the whole project, by setting up a distributed compute and storage infrastructure, anchored in regional and national data centers, for life sciences, which will be managed by a mutualised task force regrouping members of IFB platforms and support from the mesocenters. This WP will focus on providing mid-term secure storage on all NNCR (National Network of Computing Resources) sites.

Specific tasks

1. Rationalising the equipment of IFB federated platforms by installing all the facilities in labeled regional or national data centers.

2. Support the Core and regional NNCR nodes, by combining HPC and mid-term secured storage.

3. Expanding the services to regions not yet covered by the IFB NNCR.

4. Build a back up network between sites and within the NNCR network

5. Create shared data spaces (data lake) enabling the integration of different data types and their access by different computing technologies in a transparent way.

Use cases

U1. Mid-term securing for all NNCR nodes.

U2. Occitanie data lake setup in collaboration with 'mesocentres'

U3. FAIRifying the data from the second day.

Deliverables

D2.1. Infrastruture convergence within the NNCR: equipments and practices

D2.2. Strengthen mid-storage of the NNCR nodes

D2.3. Build an NNCR backup network

D2.4. A demonstrator of Data Lake will be developed based on needs to integrate sequencing data with other data types. Data space on Clément Ader reserved to France Génomique for the integration of NGS data with other data types (challenging a regional organization vs a centralized national one for scientific projects).

D2.5. Demonstrator for the combination of HPC on big data based on sequencing + other data

Partner INBS

France Génomique; IBISBA

Other partners / collaborators

INRAE DSI

WP 3. Data access and outreach

WP coordinators

Julien Seiler, Olivier Collin, Claudine Médigue

Short description

This WP will deal with the final destination of the data, be it an international or local repository. IFB is recognized as a data hub for Life Sciences Data by CNRS-INSB and Elixir. CNRS-INSB has mandated IFB for the setup of a Dataverse repository. Elixir has solicited IFB to act as data broker for Elixir deposition databases. These two missions are intertwined and will help IFB to develop a strong connection between data conservation and data publication in international repositories. IFB will endorse a strategic role in research data management enabling biologists to adopt FAIR principle and take a big step toward Open Science.

Specific tasks

1. Create a national repository (BioDataVerse) that will strengthen and complement existing institutional repositories

2. Liaise and interact with Elixir Deposition Databases services for the data brokering

3. Creation of a permanent role of data coordinator as interface for the thematic communities

4. Host and run thematic communities that will put together their expertises to enable seamless curation and validation of datasets for a wide variety of life-science and health domains

5. Connect with meta-data portals in order to give visibilities across the world to all datasets hosted by IFB and partners

6. Connect with the FLI-IAM portal to give access to both biological and in vivo imaging repositories.

Use cases

U1. Data brokering pilot-project : facilitating the submission of French sequencing data to EBI-ENA
U2. FAIRifying the data for the day after

Deliverables

D3.1. Procedures for the validation of metadata quality

D3.2. Procedure for the automated submission to repositories

D3.3. Training material (e-learning and live courses)

Partner INBS

France Génomique

Other partners / collaborators

ENA-EBI

WP 4. Intensive Computational Biology (Access to national HPC/AI resources)

WP coordinators

Christophe Blanchet, Philippe Hupé

Short description

The work package "Intensive computational biology" aims at enabling the life science communities to use for projects with intensive and IA-related computing needs the existing intensive computing resources (HPC, AI, Bigmem) available in the four national centers IDRIS, TGCC, CINES (both affiliated to GENCI) and CC-IN2P3. The challenges are to deploy on these national facilities the usual tools and data in life science, to provide users with common workflow environments and scientific gateways and web portals. The use cases include AI projects, and projects requiring very large computing resources, such as health applications related for instance to COVID-19 or large scale microbial genomes analysis. Running applications for intensive computations required that tools and pipelines have been adapted and benchmarked at a preliminary step in representative regional HPC/AI environments (e.g. CBP-PSMN).

Specific tasks

1. Deploy on the national HPC/IA resources the required tools and databases.
2. Deploy common life science workflow engines in HPC facilities (nextflow, snakemake, CWL).
3. Provide application developers with benchmarking environments
4. Share data between HPC/IA and IFB resources (WP2 and project FITS)
5. Evaluate integration of HPC/IA resources with public scientific gateways of IFB
6. Train developers and users to HPC/IA computational environments

Use cases

U1. Running health applications in a HPC/AI environment (coll. with IS3)

U2. Running microbial genomes analysis in a HPC/AI environment (coll. with IS4)

U3. Benchmarking applications with different computing resources

U4. Moving data between HPC/AI resources (Jean Zay at IDRIS) and external resources (CC-IN2P3) to make them accessible and interoperable to the life-science community. (coll. with project FITS)

References

* Jarlier F, Joly N, Fedy N et al. QUARTIC: QUick pArallel algoRithms for high-Throughput sequencIng data proCessing [version 1; peer review: 1 approved with reservations]. F1000Research 2020, 9:240 (https://doi.org/10.12688/f1000research.22954.1)

* Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O, NGPhylogeny.fr: new generation phylogenetic services for non-specialists, Nucleic Acids Res. 2019 Jul;47(W1):W260-W265.

* COVID-Align: Accurate online alignment of hCoV-19 genomes using a profile HMM

F Lemoine, L Blassel, J Voznica, O Gascuel

bioRxiv

* Sarah Cohen Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, Christophe Blanchet. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Generation Comp. Syst. 75: 284-298 (2017) https://doi.org/10.1016/j.future.2017.01.012

Deliverables

D1. Integration of life-science applications in HPC/AI, including tools, data and workflow engines

D2. Recommendations on life-science tools adaptation and benchmarking for HPC/AI

D3. Connection of HPC/AI resources and public gateways for life-science data accessibility and interoperability

D4. Training contents for HPC/AI usage

Other partners / collaborators

- Pasteur
- Curie
- ICM
- IDRIS
- GENCI
- CCIN2P3
- CBP-PSMN

IS 1. Imaging data integration and FAIR sharing

WP coordinators

Jean-François Dufayard (IFB), Perrine Paul-Gilloteaux (FBI), Emmanuel Faure (FBI)

Short description

Imaging data is a key resource among the French life science community. This IS aims to manage the life cycle of imaging data among pilot projects and shared human and material resources between IFB and other INBS producing imaging data. FBI , FRISBI, EMBRC-Fr and PHENOMIN plan to delegate IFB infrastructures some refined image data, and use IFB computing facilities to run some specialised workflows. NIRs FBI and EMBRC-Fr data and process flows for management and fairification will serve as the implementation study (WP1-WP2-WP3-WP4), but in close collaboration with the other NIR using imaging. FLI, FRISBI, EMBRC-Fr, PHENOMIN and FBI will contribute to the FAIRification of imaging data toward their integrated analysis, by ensuring access to multi-scale imaging data from different databases in an integrable and interpretable way. This will also contribute to the other IS (2-4-5) for the representation of data extracted from images for the integration of heterogeneous data.

Complement of description

To define and implement the imaging data flow, two pilots have been defined (EMBRC-Fr and FBI). The reasons to choose these two pilots are : 1) the EQUIPEX/ESR+ FBI. Data from FBI is defined on a complementary aspect with MuDiS4LS and has planned dedicated human resources and road map toward the homogenisation of their data management plan and its implementation. The complementarity and close collaboration with IFB will be ensured by shared staff planned in both Equipex (Bird IFB-Bretagne Loire FBI) and data mission officers from FBI co-leading this IS . 2) EMBRC-Fr, FBI and IFB have already started to work on imaging DMP (myEMBRC-image). These two pilots also represent two main data organization (centralization of data for a distributed infrastructure for EMBRC-Fr, distributed data for a distributed infrastructure but with a common entry point for FBI), they will be constructed with the other mentioned NIRs, and the data flows developed or preconised software layers for deployment and use cases identified will be ready to be translated to other infrastructures (Phenomin and Frisbi). It has to be noted that part of the data produced are delegated to IFB: image acquisition system produce tremendous amount of data (up to several Tb a day per system) and that this data needs to be preprocessed and pre analyzed locally before a selection is sent to mesocenters for high throughput processes. - Regarding the FAIRification of data for heterogeneous data integration, the purpose will be to create coherent description of the extracted markers or phenotypes from imaging data, which covers in these NIRS the protein to the whole patient or animal scale (through cells , tissue and organs scale), and multi temporal aspect (from nanoseconds to days or years). FLI has already gained experience in collaboration with IFB for integration of imaging data extracted from imaging data (ICAN) , and the other NIRs have experience in multiscale and multi temporal imaging data integration. All NIRs are closely related to european infrastructures , such ensuring the coherence of any national choice with the european strategies.

Specific tasks

1. Software deployment and interoperability procedures between implied INBS.

2. Elaboration of data management plans.

3. Training (train the trainers, and train final users), for both DMPs and information systems.

4. Define the roadmap to create the French IDR (Image Data Repository for public archiving).

5. Define the FAIR exposition of imaging data for integration and analysis of heterogenous data.

Tasks details

Concerning task 1, the software deployment and interoperability procedures will be developed between many infrastructures: IFB, FRISBI, Phenomin (part of CELPHEDIA), EMBRC, FLI and FBI. Concerning task 2, there are two types of data management plan : those dedicated to image management structures, and those for scientific projects, including links to public archiving.

Use cases

U1. myEMBRC-image (collaboration with a third infrastructure EMBRC with a scheme of centralized data server for 1 distributed infrastructure: one source of data for WP5.

U2. FBI-biological imaging Equipex: geographically distributed and diverses interoperable management systems with a unified entry point

Deliverables

D4.1 List of standards ontologies for imaging data

D4.2. structure DMP for image data management systems

D4.3. deployment on image facilities.

D4.4. deployment of shared workflows for image analysis.

D4.5 Training material for image data management

D4.6 recruitment of mutualised SI staff between IFB and FBI to manage shared resources.

Partner INBS

FBI, EMBRC-FR, Frisbi, Phenomin, FLI

IS 2. Marine biology data integration and dissemination

WP coordinators

Erwan Corre (IFB + EMBRC), Lucas Leclère (EMBRC), Eric Pelletier (France Génomique + CEA + FR-2022 GO-SEE)

Short description

Research on marine organisms and ecosystems is experiencing a major revolution with the omics methods now available. The aim of this implementation study, in close connexion with the WP6 of the AO-EMBRC project, will be to enhance and extend cross-referencing of environmental descriptors, taxonomy, multi-omics, modelling and imaging data as well as analysis pipelines generated by research on marine organisms and ecosystems promoted notably by the EMBRC infrastructure and the TARA consortium. It will rely on construction of marine specific DMP in collaboration with the WP 1 inspired by the work initiated in the framework of the ELIXIR marine metagenomics community. It will contribute, in collaboration with WP2, to the development of a national infrastructure to ensure regular processing and dissemination of the data produced by the marine stations and observatories. It will promote the FAIRfication of marine models and augmented observatories datas and their dissemination in national and international ecological data infrastructures (DataTerra, Emodnet Biology) on the one hand and with genomics and imaging data warehouses on the other hand (ENA, EuroBioImage).

Specific tasks

1. Set up specifically defined maDMPs for marine data and augmented observatory data.

2. Implementation and dissemination of pipelines and reference data on the various regional infrastructures to ensure optimal load balancing when processing data. Establishment of an IT infrastructure for baseline analysis of European Genomic Observatory data.

3. Use the NNCR infrastructure to provide a single interface for making imaging and genomic data of marine model organisms available.

4. Promote the interoperability of data generated from marine model organisms or augmented observatories with national and international ecological data infrastructures (DataTerra, Emodnet Biology) on the one hand and with genomics and imaging data warehouses on the other hand, in agreement and application of the FAIR principles.

5. Propose a first framework for integrative modelings, to link systems scales, using genes to ecosystem models. Establishment of an infrastructure to link genome-scale (i.e., https://www.ebi.ac.uk/biomodels/), community-scale, and ecosystem-scale models. Ultimately these models will be linked to large-scale biogeochemical and climate models. 6.Publication and dissemination of raw and elaborated marine data by the use of data brokering services provided by the infrastructure.

Use cases

U1 Extension and enhancement of integration tools for data integration and visualization The Tara-Oceans project data (which includes multi-omics, imaging, and environmental descriptors) and existing query and visualization services (OGA/OBA) will serve as a sandbox for the development, interoperation, and implementation of reusable services.

U2 FAIRification of marine data analysis workflows and results Barcodes data analysis process is yet to be standardized. The expertise of the IFB framework will help in such walks. Other omics data will then be taken in charge, as well as imaging. https://www.openaire.eu/how-to-make-your-data-fair

U3 Marine model organism data structuration initiates the organisation and development of a common framework to describe and host data (omics, images) related to marine model organisms.

U4 Data brokering pilot We will adapt the procedures developed in WP3, deliverable D3.2, to set-up a workflow allowing to facilitate the submission of marine data to EBI-ENA. This use-case will benefit from the ELIXIR community "Marine metagenomics" collaborations which included teams from the EBI directly connected with the ENA (https://www.ebi.ac.uk/ena).

Deliverables

D2.1: Generic data biogeography query and visualisation tools for genes / functions / genomes / diversity / environmental parameters / images, including time series

D2.2: National instance of the EcoTaxa imaging data service through the IFB infrastructure, and interconnected with omics data.

D2.3: FAIRification of metabarcode data analysis pipelines.

D2.4: List of recommendations for marine model organism data integrated representation (omics / imaging / modelling) and visualisation - Ecotaxa / Galaxy Genome Annotation

D2.5: Standardised pipeline for marine data submission to ENA (and other data warehouses)

Partner INBS

EMBRC, France Genomique, DataTerra-PNDB

Other partners / collaborators

OSU Stations Marines de Sorbonne Université, FR-2022 GO-SEE Univ. Nantes OSU Pythéas Aix-Marseille Université

IS 3. Bioinformatics solutions to handle health data

WP coordinators

Abdelkader AMZERT (DSI Inserm), Boris DINTRANS (CINES), Ivan MOSZER (ICM/iCONICS), David SALGADO (MMG-GBIT)

Short description

Health data is sensitive: it requires specific storage and computing environments to ensure compliance with regulatory policies. Routine care, diagnosis and prognosis data should live in "HDS" environments, while clinical research protocols should be subject to pseudonymisation and appropriate agreements (CPP, CNIL), in particular. This Implementation Study (IS) will benefit from a physical, technical and human environment located at the CINES (National Computing Center for Higher Education), which is at the crossroad of key national infrastructures such as France Cohortes and the CAD (Collecteur Analyseur de Données of the France Médecine Génomique 2025 plan). The objective of this IS is to extend this environment to users at the national scale. This infrastructure will support large-scale workflows for the management, processing and sharing of sensitive health data. To initiate this infrastructure, four use cases dealing with biomedical data will be developed.

Specific tasks

1. Based on the frameworks provided by the 4 technological WPs, set up services to (i) manage, process, benchmark, host and share sensitive health data, (ii) evaluate and adapt new technological approaches from WP1 and WP4 to sensitive data.

2. Provide guidelines, templates and tools for writing and implementing biomedical DMP to enforce FAIR principles, through the adaptation to sensitive data of the Researcher Digital Environment (WP1).

Tasks details

1. Based on a new HDS-certified IT environment and the frameworks provided by the 4 technological WP of this project, set up 3 types of services to (i) manage, process, benchmark and share anonymized health data, (ii) make available analysis results through specialized databases, (iii) evaluate and adapt new technological approaches to sensitive data from WP3 an WP4, both from a hardware (GPU, FPGA) and software (parallelized file systems, Hadoop, federated EGA, GA4GH encrypted file) viewpoint.

2. Provide guidelines, templates and tools for writing and implementing biomedical DMP, with specific recommendations and resources to enforce FAIR principles, including the connection with the maDMP (machine actionable Data Management Plan) strategy and the referencing of publicly available data in long-term storage repositories, through the adaptation of the Researcher Digital Environment to sensitive data (based on D2 from WP1). The researcher's digital space aims to set up a data management layer to collect, update and share metadata across the different steps of a research project. This data management layer will automatically interact with the maDMP. This approach will allow research teams to use metadata (describing sensitive health data) from HDS-IT environments without having access to these sensitive datasets.

Use cases

U6.1. Data sharing with international standards (GA4GH, ELIXIR, B1MG) and national constraints

The main objective of this use case is to evaluate the feasibility of setting up a genomic data sharing infrastructure, which will be secured and compliant with "national rules", based on the recommendations provided by the "Global Alliance for Genomic Health" (GA4GH), ELIXIR and the EU project in which IFB is participating: "Beyond 1 Million Genomes" (B1MG) and EUCANCan. This will focus on the usage of technological frameworks, toolkits and solutions, and their adaptation to French regulatory constraints, such as those from the ELIXIR-Federated Human Data community (local-EGA, Beacon, ELIXIR AAI), the GA4GH Genomic Data (use, access ontologies, phenopackets), and data security toolkits (Crypt4GH, Data access policies, Passports).

U6.2. Resource exchanges with N4HCloud for integrating imaging and omics health data

Strong interactions will be fostered with the N4HCloud platform, designed concomitantly in the framework of this ESR/EQUIPEX initiative. This digital platform for health data is primarily focused on neuroimaging data, with the aim to integrate other modalities. Specific pipelines, software environments and specialized data resources for genomics and multimodal data integration will be transferred to N4HCloud, validating both its architecture and the portability of our resources, following FAIR principles. This will pave the way for innovative studies dealing with heterogeneous data integration. We will rely on the IntegrParkinson pilot project, already funded by IFB, which involves the NUCLEIPARK and ICEBERG clinical studies with several hundred patients followed longitudinally on the early stages of Parkinson's disease (clinical data, MR imaging, PET imaging, genotyping, transcriptomics and metabolomics data).

U6.3. Database hosting: Example of a Covid-19 action

This use case will focus on the portability of a resource specifically developed during the IFB response to the Covid-19 outbreak: the COVIDScan database has been developed to gather results of thoracic CT-Scans from suspected patients with Covid-19. This demonstrator will help to showcase the adequation between the developed infrastructure and the national regulatory constraints for health data. The portability process will rely on the development of standardized, isolated and secured Virtual Machines to host any databases with sensitive data. This use case will focus on Covid-19 as an example: it could however be extended to other health crisis situations.

U6.4 Data provenance for multi-site sensitive data

In this use case, we will focus on multi-site sensitive data with non-relocatability constraints and will benefit from provenance metadata for better findability and reusability. More precisely, we will leverage the data provenance service developed in WP1 in the context of whole genomes stored and analyzed in U6.1 at the CINES data center, and imaging data stored and processed in U6.2. Life science researchers will be able to query the data provenance service to retrieve data lineage (production and transformation process) and their associated research context. We will rely on data from the INEXMED project, funded by IFB as a pilot project to setup a FAIR data infrastructure facilitating the development of IA methodologies, based on large scale and diverse datasets (medical imaging, omics data, clinical observations) in the context of intracranial aneurysms and congenital myopathies.

Deliverables

D6.1. HDS-compliant infrastructure for the analysis and sharing of sensitive health data

D6.2. Specialized DMP guidelines for biomedical research

D6.3. Adaptation of the Digital Environment for Researchers to sensitive data

D6.4. Linking with the N4HCloud digital platform for health data

D6.5. Proofs of concept to use an integrative environment for hosting and sharing health research data and metadata, based on provenance

Other partners / collaborators

CINES, INSERM DSI

IS 4. FAIR Integration and sharing of new data deluge in microbiome research

WP coordinators

Claudine Médigue, Nicolas Pons (MetaGenoPolis)

Short description

This IS aims at setting-up a shared space for the integration of FAIR (meta)-omics (genomics, transcriptomics, metabolomics,...) data obtained on (i) a large number of microbiome samples from the human body, animals or various environments, and (ii) libraries of bacterial genomes and their genotypic nomenclatures (https://bigsdb.pasteur.fr). The Healthy French Microbiome program (100,000 metagenomics samples) and the sequenced genomes of the strains of Institut Pasteur libraries will be the starting input data to this IS. Specific use cases will address the question of antibiotic resistance prediction and its evolution, and the characterization of human gut microbiomes infected by SARS viruses (signatures based on microbiome profiles). This Implementation Study will be based on several well-known databases and tools in microbial bioinformatics developed at the national level and supported by IFB, on the tight collaborations of IFB in ELIXIR actions, and our involvement/collaborations in ongoing international projects.

Specific tasks

Task1. Developing, in collaboration with WP1-4, a shared data space for the storage and integration of massive sequencing data of genomic and metagenomic data and other data types. This task will also provide the guidelines and templates for implementing DMPs including FAIR principles and covering minimal standards for microbiome data acquisition, processing, deposition and interoperability with specific recommendations for multi-omic data integration. Definition of standards and ontologies for FAIRifying data will be based on existing European initiatives (ELIXIR and MIRRI ESFRI) and discussed in collaboration with the MuDiS4LS IS5 and others partners of projects submitted to the same call, mainly MWD (MicroWorld Discovery project), and ALADIN (Active Learning to Accelerate biocatalyst Development for INdustrial biotechnology).

Task2. Establish minimal requirements for metagenome annotation strategies in particular for antibiotic resistance genes and mobile genetic elements involved in their transfer to bacterial pathogens and dissemination across interconnected ecosystems. Task2 will benefit from the strain library of human pathogens hosted by Institut Pasteur and the effort of database mutualisation supported by the antibioresistance Priority Research Program ("Platform of integrated microbiologics and multi-omics data" call), as well as from our involvement in the MAGITICS project (MAchine learning for diGItal diagnosTICS of antimicrobial resistance; https://www.jpiamr.eu/supportedprojects/9th-call-results/).

Task3 : Define recommendations and considerations in structuration of large dataset in regards of microbiome application for AI and HPC implementation on Jean Zay facilities. To address biological questions raised in U3 of this IS, specific analysis workflows will be developed, in collaboration with WP4 (use-case 2). Open data produced in the French Gut project will be used.

Use cases

U1. Dissemination of antibiotic resistance genes

In a one-health scale, this use case aims to document potential dissemination risks of antibiotic resistance genes through interfaced ecosystems. It is in line with the "Research Priority Program" in antibiotic resistances launched by the French research ministry a few months ago. Antibiotic resistance gene (ARG) reservoirs will be described in (meta)-genome data from various ecosystems (human, animal and environment). Genetic context of these genes will be characterized using specialized databases and tools for annotating mobile genetic elements (plasmids, integrons, IS, conjugative elements…) in order to determine potential transfer to bacterial pathogens. Comparison of metagenomic (and genomic) data will allow to track potential fluxes and dissemination of strains (and species) harboring antibiotic resistance genes and pathogen strains as well. Access to large cohorts (French Gut and MMHP e.g.) will permit to explore the diversity and evolution of ARG reservoir according to age, geography, or clinical status.

U1 will rely on national bioinformatic and biological resources addressing these questions (Institut Pasteur, MicroScope, Genotoul, MetaGenoPolis, Migale, BiRD, Bilille…). U1 will benefit from future efforts in platform mutualisation for antibioresistance research (antibioresistance Priority Research Program in the call "platform of integrated microbiology and multi-omics data"), and in capture of massive volume of metagenomic data in French Gut project and MMHP. Links between IS4, IS2 and IS5 will be done in order to address the one-health question.

U2: Understanding the link between microbiome signatures and susceptibility to virus infection

Recent studies begin to show a potential association between SARS-CoV-2 infection and microbiome composition. Although the more general effects of the current pandemic are mitigate, no cure against Covid-19 is available and the risk of emergence of new viruses and of future pandemics, in particular SARS viruses, will remain. U2 will address the question of the microbiome susceptibility or adaptation to a virus infection at the taxonomic and functional levels by aggregating with FAIR principles relevant microbiome dataset for example Covid-19 cohorts progressively deposited in public repositories. U2 will link with efforts in IS3 (management of health data) and WP4 (processing of large volume of data).

U3. Deployment of machine learning and AI methods on the Jean Zay computer (IDRIS) for the analysis of large cohorts

Using machine learning and deep learning approaches, U3 will address the question of the microbiome stratification of individuals in large cohorts, the prediction of phenotype according to microbiome composition as well as the antibiotic microbial prediction up to MIC (Minimal Inhibitory Concentration) identification using NGS and metabolomics data. U3 aims to provide guidelines for the data structuration in the context of Artificial Intelligence (AI) applied on microbiome data, choice of AI methodology and implementation on dedicated computing resources (GPU cluster on Jean Zay computer e.g.). Open data produced in French Gut and MMHP projects will be used. Microbial signatures characterized in U2 could be used as prediction models tested on large cohorts. U3 will be closely linked to WP4.

U4. Data brokering pilot

We will adapt the procedures developed in WP3, deliverable D3.2, to set-up a workflow allowing to facilitate the submission of microbiome data to EBI-ENA. This use-case will also benefit from the ELIXIR community "Marine metagenomics" collaborations which included teams from the EBI directly connected with the ENA (https://www.ebi.ac.uk/ena)

References

Garcia-Garcera, M. & Rocha, E.P.C. 2020. Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nat Commun* **11**: 758.

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., *et al.* 2020. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput Biol* **16**: e1007732.

MetaCardis Consortium, Vieira-Silva, S., Falony, G., Belda, E., Nielsen, T., Aron-Wisnewsky, J., *et al.* 2020. Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**: 310–315.

Moura, A., Criscuolo, A., Pouseele, H., Maury, M.M., Leclercq, A., Tarr, C., *et al.* 2017. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nat Microbiol* **2**: 16185.

Oliveira, P.H., Touchon, M., Cury, J. & Rocha, E.P.C. 2017. The chromosomal organization of horizontal gene transfer in bacteria. *Nat Commun* **8**: 841.

Perrin, A., Larsonneur, E., Nicholson, A.C., Edwards, D.J., Gundlach, K.M., Whitney, A.M., *et al.* 2017. Evolutionary dynamics and genomic features of the Elizabethkingia anophelis 2015 to 2016 Wisconsin outbreak strain. *Nat Commun* **8**: 15483.

Vallenet, D., Calteau, A., Dubois, M., Amours, P., Bazin, A., Beuvin, M., *et al.* 2019. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research* gkz926.

Ruppé, E., Ghozlane, A., Tap, J., Pons, N., Alvarez, A.-S., Maziers, N., *et al.* 2019. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat Microbiol* **4**: 112–123.

Deliverables

D1.1 enrich annotation of metagenomic data at ARG and mobile genetic element levels, in the respect of FAIR principles.

D1.2 provides reproducible workflows for analysing dissemination potential of genes through various ecosystems with a goal of application in other biological contexts (virulence genes e.g.)

D2.1 shared and FAIR repository of microbiome dataset dedicated to virus infection association analysis

D2.2 reproducible workflows for the detection of microbiome signatures in the context of virus infection in human or animal.

D3.1 guidelines for microbiome data structuration in the context of AI approaches, and guidelines for AI approach choices.

D3.2 workflow for data structuration and processing with orchestration in dedicated computing facilities (HPC, massive GPU cluster).

D4.1 workflow for (meta) genomic data submission to the EBI Nucleotide Archive repository.

Other partners / collaborators

MetaGenoPolis

IS 5. Integration and FAIR sharing of genetic and multi-omics data for agriculture

WP coordinators

Valentin Loux (IFB), Anne-Françoise Adam-Blondon (RARe&IFB), Michèle Tixier-Boichard (RARe)

Short description

The ability to maintain a wide variety of well-documented resources, to collect new ones, to contribute to their characterization, to distribute them and to manage associated data places the Biological Resource Centers (BRCs) of the RARe national infrastructure, which lies at the heart of many research programs intended to explore the living organisms and ecosystems as well as valuing biodiversity for agriculture and industry, food, environment and health. The objective of this work package is to initiate the development of services contributing to link biological resources to highly heterogeneous types of data including various types of 'omics' data, phenotypic measurements including images and environmental data in particular for holobiont studies on animals and plants together with their commensal, symbiotic and pathogenic microorganisms.

Task title list

1. Identification of data of interest for holobiont characterization in relation with the RARe infrastructure and its interested users communities (Plants, animals, food and environment).

2. Identification and development of standards for FAIRifying existing data in relation with the chosen use cases and building on international initiatives

3. Developing in collaboration with WP1-4 an infrastructure supporting machine actionable data management plans and data analysis of heterogeneous data about holobionts

4. Pilots will be carried out: FAIRification of data sets, maDMP, reproducible analysis workflow when possible in relation with funded projects.

5. Dedicated training activities will be developed.

Tasks details

1. Identification of data of interest for holobiont characterization in relation with the RARe infrastructure and its interested users communities (Plants, animals, food and environment) and of what is crucial for interoperability. The use cases will be based in particular on the MicroWorld Discovery project (MWD; submitted to the same call) and extended to projects on algae holobionts in collaboration with EMBRC.

2. Based on the experience gained in the Elixir, EMBRC and MIRRI infrastructures, and in collaboration with IS 4, existing standards for fairifying existing data in relation with the chosen use cases will be identified and developed when necessary with an international community. This will benefit from tight links with ELIXIR platforms and communities and with the Phytobiome Alliance.

3. Building on the resources of IFB (already existing or developed in the present project in WP1-4 and IS 4), EMBRC, ELIXIR and MIRRI, an infrastructure supporting machine actionable data management plans and data analysis will be specified to ensure that the IFB infrastructure is in capacity of taking in charge heterogeneous data about holobionts.

4. Pilots will be carried out: FAIRification of data sets, maDMP, reproducible analysis workflow when possible in relation with funded projects.

5. Dedicated training activities will be developed.

Use cases

U1- The scientific objective of the Micro World Discovery (MWD) project is to develop cultural and analytical techniques to isolate, identify and characterize at high throughput level a repertoire of microorganisms (archaea, bacteria, fungi / yeasts, oomycetes and protists)

isolated from different biotopes (soil-environment,, plant, food, human-animal) that constitute a unique continuum relevant to Agriculture. The project will identify, isolate and characterize by various methods in collaboration with the France Genomic and Metabohub infrastructures a large set of microorganisms and conserve them in collections enabling their access for further research in collaboration with the RARe infrastructure. IS5 will help to set up a MaDMP, building on building on the WP1-3 outputs and on standards and guidelines developed in the EMBRC, MIRRI and ELIXIR infrastructures for data FAIRification (H2020 projects Elixir-Excelerate, EOSC-Life, ELIXIR-CONVERGE). IS5 will collaborate with the data brokering service of IFB to prepare the publication of its OMIC data in the EMBL-EBI archives. IS5 will ensure that the different types of data collected : description of the environments of origin of the consortia, of the isolated microbes, OMICs characterization data remains interoperable.

U2 – MWD will also characterize the influence of synthetic microbiote consortia in the expression of plant phenotypes in collaboration with the ANAEE and EMPHASIS infrastructures. IS5 will build on the suite of tools developed in the frame of the ELIXIR plant science community (again through the H2020 projects ELIXIR-Excelerate, EOSC-Life and ELIXIR-CONVERGE) in collaboration with EMPHASIS to propose extended standards for phenotyping and support the development of MaDMP in relation with the study of the role of microbiote consortia in phenotypes. U2 will ensure that the data remain interoperable with the data developed in U1.

U3 One health and the holobiont- In animals, it has been shown that the variability of the microbiota depends partly on the genotype of the host (between individuals as well as between breeds) and partly on environmental factors, intrinsic, such as sex or age, or extrinsic, such as diet composition or living conditions. In animal breeding, models are being developed to include microbiota data in the prediction equation of breeding values. The more accurate the documentation is for the microbiota, the environmental factors, the host phenotype and genotype, the more efficient the prediction will be. In pigs, modifying microbiota composition appears to be possible by experimental selection on the basis of 16S taxonomic profile. The microbiota composition can be correlated to feed efficiency and immune response of the host, with effects on carriage of gut pathogens (such as salmonella in chickens) or on efficiency of vaccination (in pigs). Understanding the role of the microbiota on gut heath is a common objective for humans and animals, with important applications in nutrigenomics, where probiotics can sustain or restore a proper function of gut epithelia and impact health of the whole organism. Reciprocally, knowing the microbiota of the host can help to design tailored diets for specific populations of hosts. Coupling the microbiota of the animal with that of the diet (raw or transformed) is a field of research yet to be explored. Microbiota has also been studied in relationship with sportive performance in horses, coupling metabolomic data with 16S data showed a link between the microbiota and blood metabolites related to lipid metabolism and glycolysis at basal time. Finally, relationships between microbiota, gut health and animal welfare also need to be better known, but requires to integrate complex data about behaviour of the host, often collected through image analysis at individual and group level, which yields massive data.

References

Daval, S., Gazengel, K., Belcour, A., Linglin, J., Guillerm-Erckelboudt, A.-Y., Sarniguet, A., *et al.* 2020. *Soil microbiota influences clubroot disease by modulating* Plasmodiophora brassicae *and* Brassica napus *transcriptomes*. Pathology.

Massacci, F.R., Tofani, S., Forte, C., Bertocchi, M., Lovito, C., Orsini, S., *et al.* 2020. Host genotype and amoxicillin administration affect the incidence of diarrhoea and faecal microbiota of weaned piglets during a natural multiresistant ETEC infection. *J Anim Breed Genet* **137**: 60–72.

Munyaka, P.M., Kommadath, A., Fouhse, J., Wilkinson, J., Diether, N., Stothard, P., *et al.* 2019. Characterization of whole blood transcriptome and early-life fecal microbiota in high and low responder pigs before, and after vaccination for Mycoplasma hyopneumoniae. *Vaccine* **37**: 1743–1755.

Plancade, S., Clark, A., Philippe, C., Helbling, J.-C., Moisan, M.-P., Esquerré, D., *et al.* 2019. Unraveling the effects of the gut microbiota composition and function on horse endurance physiology. *Sci Rep* **9**: 9620.

Roselli, M., Pieper, R., Rogel-Gaillard, C., de Vries, H., Bailey, M., Smidt, H., *et al.* 2017. Immunomodulating effects of probiotics for microbiota modulation, gut health and disease resistance in pigs. *Animal Feed Science and Technology* **233**: 104–119.

Simon, J.-C., Marchesi, J.R., Mougel, C. & Selosse, M.-A. 2019. Host-microbiota interactions: from holobiont theory to analysis. *Microbiome* **7**: 5.

Xiao, L., Estellé, J., Kiilerich, P., Ramayo-Caldas, Y., Xia, Z., Feng, Q., *et al.* 2016. A reference gene catalogue of the pig gut microbiome. *Nat Microbiol* **1**: 16161.

Deliverables

D5.1. Guidelines and standards for data related to the study of holobionts

D5.2. Specification for an environment for FAIR data management and analysis of microbe collections in relation with holobionts)

D5.3. Fairified example data sets

D5.4. Trainings on good practices

Other partners / collaborators

UMR IGEPP, UMR IRHS, UMR MICALIS , UMR GABI

MIRRI, ELIXIR

A4. Involvement of the personnel

The personal costs include an estimation of the involvement (person.months over the 8 years of the project → 96 PM = 1 FTE) for the personnel required for this project.

1. The members of the current task force of the National Network of Computing Resources (NNCR) who will contribute to consolidate and build new services;
2. The persons who will contribute to its geographic extension and enforcement
3. The persons required for the development, deployment and dissemination of the services on the extended NNCR.
4. The positions that would need to be stabilized during the project.
5. The scientific realisation of the use cases / demonstrators.
6. The coordination of the work packages and actions.
7. The management and administration of the project.

It has to be noted that a large fraction of this personnel is already working together in the task forces of the NNCR and in the other actions of the IFB 2018-2021 work plan.

A4.1. Summaries of the personnel involvement

Figure A4.1.1. Personnel involvement per partner organism

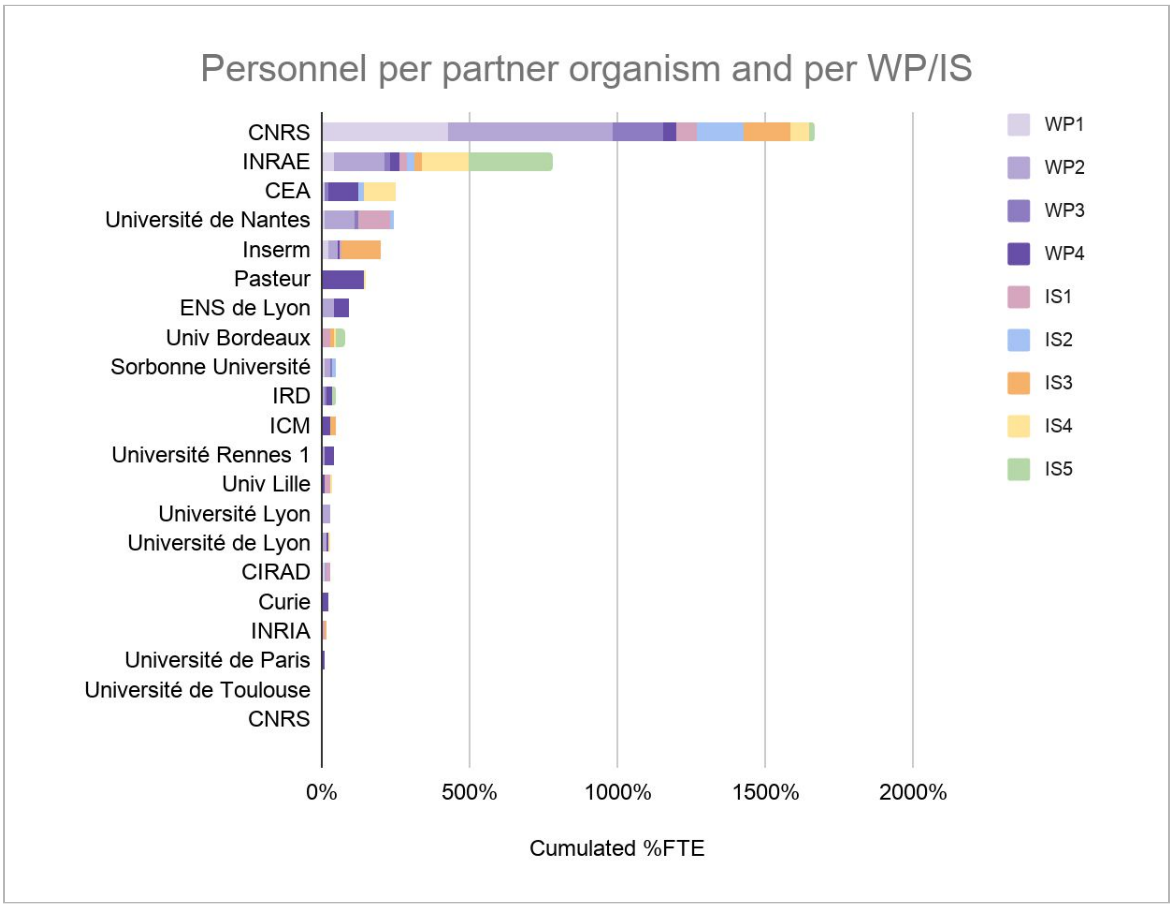


Figure A4.1.2. Personnel involvement per WP and partner organization

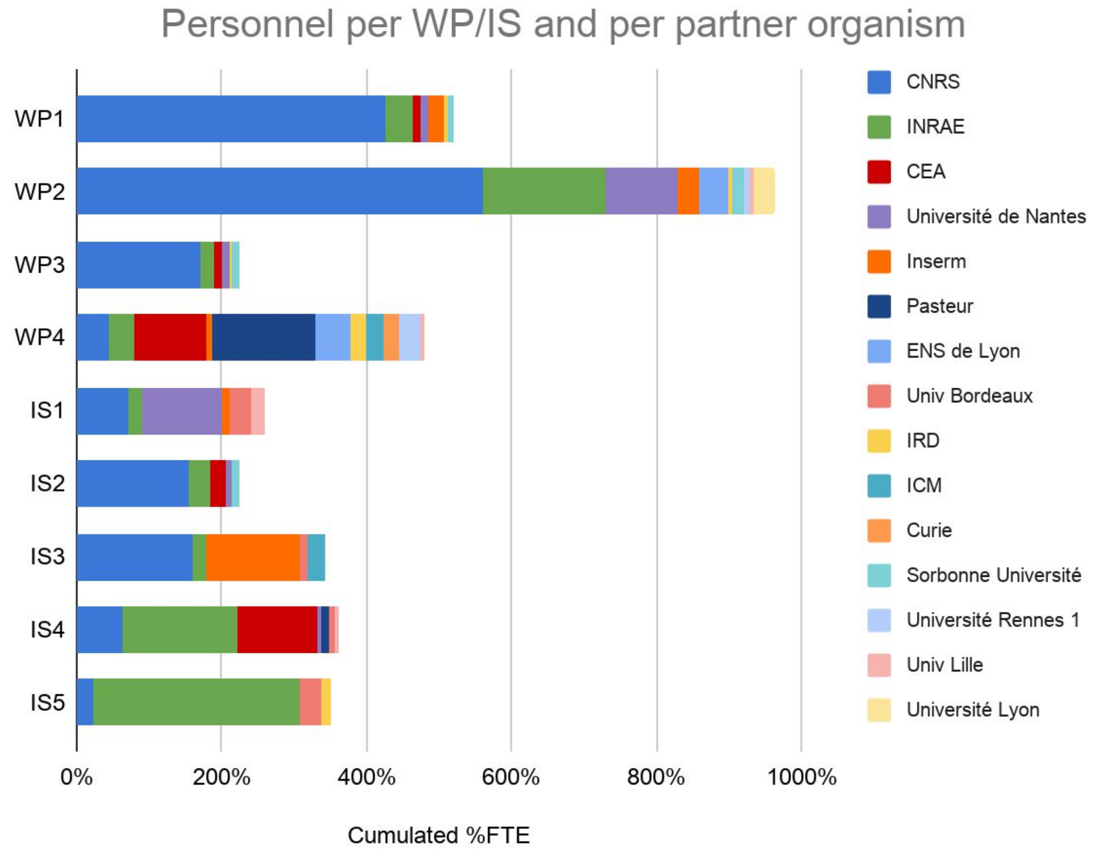
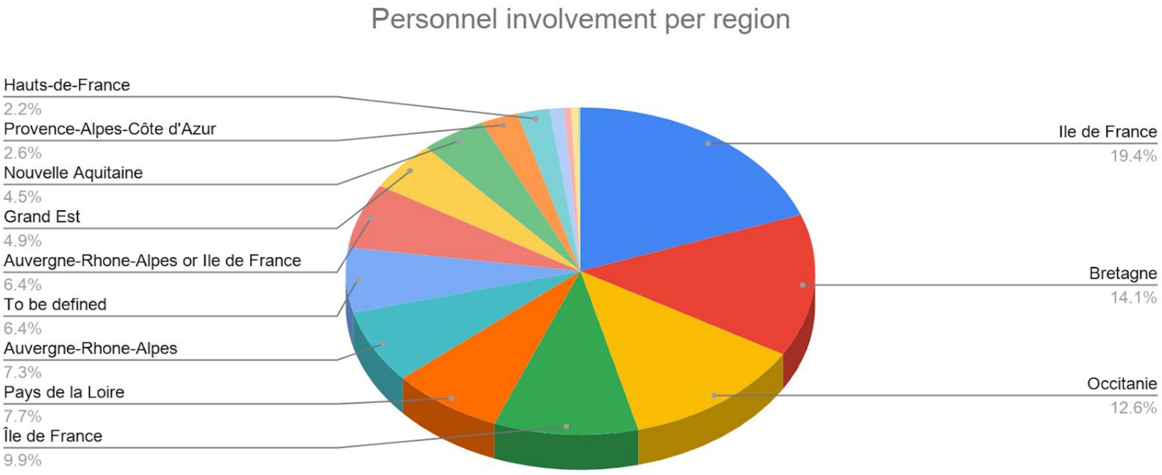
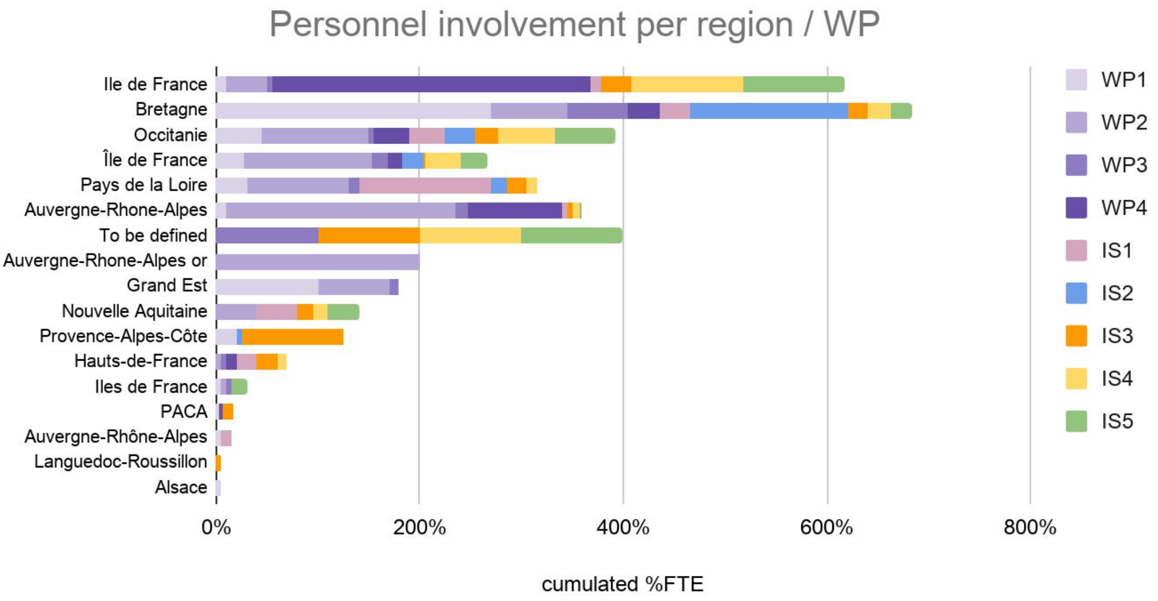


Figure A4.1.3. Personnel involvement per French region



The label "To be defined" refers to personnel that will be recruited on some IFB platform without prior decision on the site (the location will depend on the candidates and hosting team capacities).

Figure A4.1.3. Personnel involvement per French region and WP/IS



The label "To be defined" refers to personnel that will be recruited on some IFB platform without prior decision on the site (the location will depend on the candidates and hosting team capacities).

A4.2. Full list of individual skills brought by the partner units

This table provides the full list of personnel involved in the WPs andISs of MuDiS4LS. It is a complement to table 3.2.2, which was restricted to the co-leads of WPs and ISs.

Prénom Nom	Corps	Partner Unit	Partner organism	Roles / fonction / comments
Anne Françoise Adam Blondom	DR	PlantBioinfoPF	INRAE	Scientific coordinator of the Plant pilar of RARe and deputy Head of nodes for ELIXIR-France. co-coordination of IS5
Michaël Alaux	IE	PlantBioinfoPF	INRAE	Data management in agriculture; leader of a WP on data management in H2020 AGENT. Contribution to data standardization in IS5
Mathieu Almeida	CR	MetaGenoPolis	INRAE	Expertise traitement de données métagénomiques / impliqué dans le PPR antibiorésistance
Abdelkader Amzert	IR	Inserm DSI	Inserm	Co-head of IS3 Bioinfo for health; Head of the demonstrator "Linking electronic lab book with maDMP"
Hugues Aschard	CR expert	Pasteur	Pasteur	Tools development (JASS)
Mouhamadou Ba	IR	MIGALE	INRAE	Text and data Mining. IS5
Emmanuel Barillot	DR	Curie	Curie	co-Head of Platform
Aurélien Barré	IE	CBiB	Univ Bordeaux	Tools development, web development, database management, NGS data analysis, IS3, IS4
David Benaben	IE	CBiB	INRAE	core cluster + infra propre
Magali Berland	IR	MetaGenoPolis	INRAE	Expertise en analyse et modélisation de données metagenomiques, use-case IA
Damien Berry	IE	MIGALE	INRAE	System administrator, NNCR. WP1 & WP2
Thomas Bigot	IR confirmé	Pasteur	Pasteur	Tools development (Pathogene detection pipelines)
Audrey Bihouée	IE	BiRD	Université de Nantes	Head of the BIRD partner unit. NNCR cloud task force.
Christophe Blanchet	IR	IFB-core	CNRS	NNCR-cloud co-head. Biosphere coordinator.
Martial Bornet	IE	DSI ICM	ICM	WP4, cluster administration
Bryan Brancotte	IR confirmé	Pasteur	Pasteur	web development, database managment
Sylvain Brisse	DR	Pasteur	Pasteur	Head of the CRBIP (Biological Resources Center of Institut Pasteur) - IS4 workflow for epidemiological studies
Christophe Bruley	Ingénieur-Chercheur	ProFi	CEA	Mutualisation between national infrastructures for life sciences and health
Guillaume Brysbaert	IR	Bilille	CNRS	Participation to Biosphère

Micaël Calvas	AI	CBP-PSMN	ENS de Lyon	Biosphere cloud computing,
Samuel Chaffron	CR	BiRD	CNRS	IS2 : AO EMBRC partner / data integration. IS4: microbiome modeling
Stéphane Chaillou	IR	DSI ICM	ICM	Coordination WP4, IS3
Nicole Charrière	IE	IFB-core	CNRS	core cluster
Rayan Chikhi	CR expert	Pasteur	Pasteur	methodology development
David Christiany	IE	MIGALE	CNRS	implementation of the national instance of Galaxy (usegalaxy.fr), which is a node of the UE Galaxy network
Valérie Cognat	IR	IBMP	CNRS	WP1 participant
Thomas Cokelaer	IR confirmé	Pasteur	Pasteur	Tools development (Sequana)
Olivier Collin	IR	GenOuest	CNRS	NNCR-cloud co-head. WP1 co-head.
Erwan Corre	IR	ABiMS	CNRS	Economic model. PrincingTarification IS 2. Marine biology data integration and dissemination
Benjamin Dartigues	IE	CBiB	Univ Bordeaux	omics data analysis, image analysis, databases, IS1, IS5
Frédéric de Lamotte	CR	Agap	INRAE	Coordination WP1, training
Sjoerd de Vries	IR	RPBS	Inserm	Future co-head of RPBS partner. Interoperable and reproducible data flows
Patrice Dehais	IE	Genotoul Bioinfo	INRAE	Genotoul bioinfo system administrator, NNCR. WP2.
Stéphane Delmotte	AI	PRABI-Lyon-Grenoble	CNRS	Bioshere cloud girofle sys admin
Marie-Agnès Dillies	IR expert	Pasteur	Pasteur	IFB course coordinatoor
Boris Dintrans	DR	CINES	CNRS	Head of the CINES. Co-head of IS3: bioinformatics for health.
Michel Dojat	DR		INSERM	IS1 FLI-IAM contribution to FAIRification of hterogenous data
Jean-François Dufayard	CDI CIRAD	South Green	CIRAD	Mutualisation between national infrastructures for life sciences and health. S1 co-head.
Yoann Dufresne	IR confirmé	Pasteur	Pasteur	methodology development
Stanley Durrleman	IR	iCONICS	INRIA	WP4 & IS3, biomedical data, machine learning
S. Dusko Ehrlich	DR émérite	MetaGenoPolis	INRAE	PI MGP / articulation avec le projet French Gut et MMHP
Nabila Elarouci	IE	iCONICS	ICM	IS3, biomedical data management
Damien Eveillard	MCU	BiRD	Université de Nantes	IS2 : AO EMBRC partner, omics modeling

Emmanuel Faure	CR	France Bio Imaging	CNRS	IS1+ FBI equipex
Nicolas Francillonne	IE	PlantBioinfoPF	INRAE	Data management in agriculture. Contribution to the aspects on data management in IS5
Sébastien Fromentin	IR	MetaGenoPolis	INRAE	Intégration de données, use-case IA / articulation avec le projet French Gut
Alban Gaignard	IR	BiRD	CNRS	Provenance / health / neuro-imaging use cases
Olivier Gascuel	DREX	Pasteur	Pasteur	Head of Plateform
Christine Gaspin	DR	Genotoul Bioinfo	INRAE	Scientific co-head of GenoToul partner unit. Economic model. PricingTarification. Pilot project: Linking sequencing data within mesocentres in Occitanie (WP2). Contribution to IS4 & IS5.
Franck Gauthier	IE	MetaGenoPolis	INRAE	Pipeline de traitement de données métagénomiques / Calcul HPC
François Gerbes	IE	IFB-core	CNRS	core cluster
Amine Ghozlane	IR confirmé	Pasteur	Pasteur	Tools development (SHAMAN)
Hervé Gilquin	IR	CBP-PSMN	ENS de Lyon	Biosphere cloud computing, WP4 HPC/IA integration and benchmarking
Alexis Groppi	IR	CBiB	Univ Bordeaux	Co-head of CBiB partner Unit. Integrative bioinformatics. Member of FBI + IFB. IS1, IS5.
Justine Guégan	IE	iCONICS	ICM	IS3, biomedical genomics
Loraine Guegen	IE	ABiMS	CNRS	Dev, e-infra ABiMS plateform
Jean-François Guillaume	IE	BiRD	Université de Nantes	BiRD sysadmin. IT storage et computing. NNCR cloud task force + FBI Equipex
Vincent Guillemot	IR confirmé	Pasteur	Pasteur	Tools development (JASS)
Dominique Guyot	IE	PRABI-AMSB	Université de Lyon	candidat à la direction technique du PRABI-AMSB, calcul haute perfomance, formation à l'utilisation des ressources de calcul
Jean-Christophe Haessig	IR	BigEst	CNRS	core cluster
Chiapello Hélène	IR	MIGALE	INRAE	IFB training actions co-coordinator. WP3, IS4 and IS5
Mark Hoebeke	IR	ABiMS	CNRS	Informatician, software developer
Claire Hoede	IR	Genotoul Bioinfo	INRAE	Transcriptomic, metagenomic analysis and pilot of workflows development project (contributing to IS4 and IS5)
Philippe Hupé	IR	Curie	Curie	co-Head of Platform
Bernd Jagla	IR confirmé	Pasteur	Pasteur	Tools development (SchnaPs)
Frédéric Jarlier	IR	Curie	Curie	Informatician, software developer

Hanna Julienne	IR confirmé	Pasteur	Pasteur	Tools development (JASS)
Mehni Kaci	IR	Inserm DSI	Inserm	technical architect for data storage (Scality)
Michael KAIN	IR		INRIA	IS1 FLI-IAM contribution to FAIRification of heterohgenous data
Hervé-Antoine Kerjean	IE	PlantBioinfoPF	INRAE	System Administrator. IS5and links to WP1-3
Erik Kimmel	IE	PlantBioinfoPF	INRAE	Developer of information systems. Contribution to the specifications and developments in IS5
Christophe Klopp	IR	Genotoul Bioinfo	INRAE	Technical co-head of the of GenoToul partner unit, contributing to IS5
Etienne Kornobis	IR confirmé	Pasteur	Pasteur	Tools development (Sequana)
Didier Laborie	IE	Genotoul Bioinfo	INRAE	Genotoul bioinfo system administrator, NNCR and core cluster, WP2.
Pierre Larmande	CR	South Green	IRD	Data management, Artificial intelligence
François Laurent	IR confirmé	Pasteur	Pasteur	Image analysis tools for covid-19 research
Emmanuelle Le Chatelier	CR	MetaGenoPolis	INRAE	Microbiologie / expertise en analyses metagénomiques / antibioresistance / virome humain
Gildas Le Corguillé	IE	ABIMS	Sorbonne Université	MuDiS4LS technical coordinator. NNCR-cluster co-head
Lucas Leclère	CR	LBDV	Sorbonne Université	Co-resp IS2.
Vincent Lefort	IR	ATGC	CNRS	Head of the ATGC partner Unit. Member of the IFB directorate. WP1, WP2, WP4.
Rachel Legendre	IR confirmé	Pasteur	Pasteur	IFB course coordinatoor
Frédéric Lemoine	IR expert	Pasteur	Pasteur	Tools development (NGPhylogeny)
Paulette Lieby	IE	IFB-core	CNRS	Engagement of life science communities in the DMP
Pierre Lindenbaum	IR	BiRD	Inserm	Bioinformatician. Genetics team of Institut-du-Thorax. UMR1087.
Jonathan Lorenzo	IE	IFB-core	CNRS	Biosphère
Valentin Loux	IR	MIGALE	INRAE	Head of MIGALE Partner Unit. IS5 co-head.
Hervé Luga	PR	UFTMiP	Université de Toulouse	Contact Data center DROcc+projet CPER
Nicolas Maillet	IR confirmé	Pasteur	Pasteur	Tools development (RPG)
Christophe Malabat	IR expert	Pasteur	Pasteur	Plateforme deputy director
Fabien Mareuil	IR confirmé	Pasteur	Pasteur	Galaxy administrator, web development
Jérôme Mariette	IE	Genotoul Bioinfo	INRAE	Data integration. Contribution to statistical data integration for IS2 & IS4.

Guillemette Marot	MCU	Bilille	Univ Lille	Scientific Co-Head of bilille - Participation to IS4
Jean-Baptiste Masson	CR expert	Pasteur	Pasteur	Image analysis tools for covid-19 research
Gilles Mathieu	IR	Inserm DSI	Inserm	Link WP1 work with Inserm Reseacher’s digital space
Claudine Médigue	DR	IFB-core	CNRS	Direction of the IFB national infrastructure. Head of MICROSCOPE partner unit. IS4 co-head
Hervé Ménager	IR expert	Pasteur	Pasteur	Head of Web Integration Pole in the Hub
Célia Michotey	IE	PlantBioinfoPF	INRAE	Data management in agriculture. Contribution to the data management aspects in IS5
Cédric Midoux	IE	MIGALE	INRAE	WP3
Denis Milan	DR	France Génomique	INRAE	Articulation with France Génomique. WP2.
Damien Mornico	IR confirmé	Pasteur	Pasteur	web development, IFB course teaching
Ivan Moszer	IR	iCONICS	ICM	Coordination IS3
Samuel Murail	MCU	RPBS	Université de Paris	Future co-head of RPBS partner. GPU calculations for structural bioinfiramtics
Vincent Navratil	IR	PRABI-AMSB	Université de Lyon	Biosphère: développement et déploiement d'applications/ workflows métiers. Candidat à la direction scientifique du PRABI-AMSB, mise à disposition de moyens humains et de calcul à l'interface entre l'environnement et la Santé pour les utilisateurs de la future FR "Environnement, Ville et Eau" de l'université de Lyon. Projet de modélisation (biocuration) et d'analyse des systèmes hôtes/symbiontes (task force COVID-19, labex ecofect Vibraflu etc.).
Macha Nikolski	DR	CBiB	CNRS	Head of CBIB partner Unit. Integrative bioinformatics. Member of FBI + IFB. IS1, IS3 and IS4.
Céline Noirot	IE	Genotoul Bioinfo	INRAE	Contribution to WP1, IS4 and IS5 for workflow development
Christine Oger	IR	PRABI-AMSB	Université de Lyon	Biosphere cloud computing, training
Ongoing recruitment Ongoing recruitment	IE	IFB-core	CNRS	WP2. Admin Sys, permanent position opened by CNRS in 2020 for the NNCR (cloud or cluster)
Ongoing recruitment Ongoing recruitment	IR	Bilille	CNRS	Engineer in bioinformatics recruited in 2020 by CNRS/INSB for the Bilille platform - Participation to IS3
Ongoing recruitment	IR	Genouest	CNRS	ANR-funded 18-month contract for the project maDMP4LS

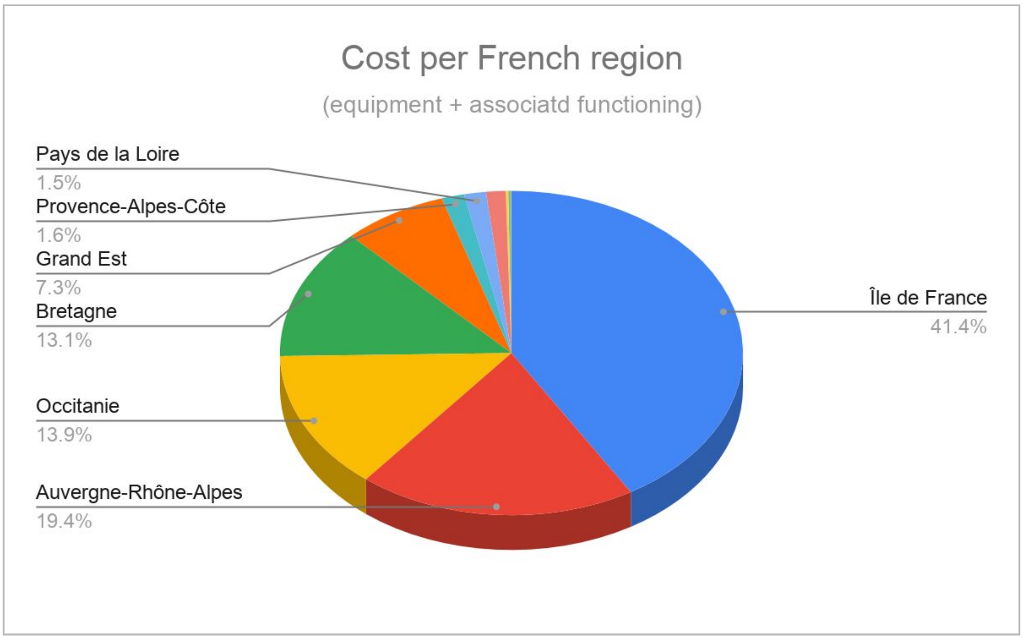
Ongoing recruitment				
Ongoing recruitment Ongoing recruitment	IE		CNRS	IFB action "Bioinformatics for health". Position currently opening, funded by PIA2.
Ongoing recruitment Ongoing recruitment	IR	Bilille	Univ Lille	Engineer in bioinformatics recruited in 2020 by Univ Lille for the Bilille platform - Participation to IS1
Julie Orjuela-Bouniol	IE	South Green	IRD	Tool development, data analysis, training
Adrien Pain	IR confirmé	Pasteur	Pasteur	IFB course teacher
Jérôme Pansanel	IR	BigEst	CNRS	Head ot the SCIGNE platform, Biosphère partner
Stephane Paris	IR	INRAE DSI	INRAE	Appui à la migration vers DROcc (DSI INRAE)
Perrine Paul-Gilloteaux	IR	France Bio Imaging	CNRS	IS1+ FBI equipex
Eric Pellettier	DR	France Génomique	CEA	Co-lead of IS 2. Marine biology data integration and dissemination
Caroline Peltier	IR	iCONICS	ICM	IS3, multimodal data integration
Bertrand Pitollat	IR	South Green	CIRAD	System administrator HPC South Green
Rémi Planel	IR confirmé	Pasteur	Pasteur	Galaxy administrator, web development
Florian Plaza-Onate	IR	MetaGenoPolis	INRAE	Expertise traitement de données métagénomiques / Calcul HPC / programmation GPU
Cyril Pommier	IE	PlantBioinfoPF	INRAE	Architecture of information systems; Semantic for data integration ; co-leader of the ELIXIR Plant Science community and of the development of the information system for EMPHASIS-France;Contribution to the specification of a numeric environment for MaDMP in IS5
Nicolas Pons	IR	MetaGenoPolis	INRAE	IS4
Emmanuel Quemener	IR	CBP-PSMN	ENS de Lyon	WP4 HPC/IA integration and benchmarking
Flores Raphaël-Gauthier	IE	PlantBioinfoPF	INRAE	Developer of information systems, semantic approaches (graph) for data integration; Contribution to the specification and development of a numeric environment for MaDMP in IS5
Richard Redon	DR	BiRD	Inserm	Scientific coordinator of the BiRD partner unit. Partner in IS3 bioinfo for health.
Julien Rey	IR	RPBS	Université de Paris	Service deployment

Eduardo Rocha	DR	Pasteur	Pasteur	Head of the unit "Microbial evolutionary genetics" - IS4, tools for comparative genomics, functional annotation and prediction of mobile genetic elements
Julien Roméjon	IR	Curie	Curie	Informatician, software developer
Thomas Rosnet	IE		CNRS	IFB action "Interoperability", funded by PIA2.
Jérôme Royer	IR	Inserm DSI	Inserm	technical architect and cyber security expert for HDS certified IT environment
Olivier Rué	IE	MIGALE	INRAE	Metagenomic boinformatics analysis, IS4 & IS5
Francois Sabot	DR	South Green	IRD	Co head of Digital Data and Infrastructure Dept at IRD - Genome structure
Valentin Saint-Léger	IE	IFB-core	CNRS	Vérifier si on peut justifier du CDD PIA3 pendant la phase précédant la pérennisation du poste
Jean Salamero	DR	FBI	FBI	Mission Officer "Inter Infrastructures Activities"
David Salgado	IR	MMG-GBIT	Inserm	Co-head of the MMG-GBIT partner unit. Coordination of IS3 Bioinfo for health; Partner in the demonstrator "Linking electronic lab book with maDMP"
Olivier Sallou	IR	GenOuest	Université Rennes 1	nccr cluster et biosphere, participation au projet elixir tools platform
Olivier Sand	IR	IFB-core	CNRS	Innovative projects in integrative bioinformatics
Julien Seiler	IR	BigEst	CNRS	MuDiS4LS technical coordinator. NNCR-cluster co-head.
Guillaume Seith	IR	BigEst	Inserm	core cluster
Nicolas Servant	IR	Curie	Curie	co-Head of Platform
Bruno Spataro	IR	PRABI-Lyon-Grenoble	CNRS	Participation to Biosphere. NNCR-cluster contact
Ndomassi Tando	IE	South Green	IRD	System administrator HPC South Green
Michèle Tixier-Boichard	DR	RARE	INRAE	Director of CRB RARe, IS5 co-lead
To be recruited To be recruited	IE à ans d'ancienneté	IFB-core	CEA	WP4. Engineer in computer sciences to develop Artificial Intelligence methods for life sciences
To be recruited To be recruited	IE à ans d'ancienneté	To be defined	CEA	IS4. Short-time contract to start IS4: microbiology
To be recruited To be recruited	IE à ans d'ancienneté	IFB-core	CNRS	WP2. Admin Sys, permanent position to be opened in the future to enforce the NNCR (cloud or cluster)

To be recruited To be recruited	IE	GenOuest	CNRS	Implementation of user-friendly interfaces to enable the engagement of life science communities using HPC resources
To be recruited To be recruited	IE	ABIMS	CNRS	IS2. Short-time contract to start IS2. marine biology
To be recruited To be recruited	IE	ATGC	CNRS	IS1.integrative bioinfo Imaging + multi-omics; joined action between IFB and FBI, with a request for permanent recruitment in the course of the project
To be recruited To be recruited	IE	Genouest	CNRS	WP1. Orchestrating data fluxes at all the stages of data life
To be recruited To be recruited	IE	To be defined	CNRS	WP3. IR for data brokering
To be recruited To be recruited	IE	Genotoul Bioinfo	INRAE	WP2. DevOps for NNCR, temporary position with perspective of permanent recruitment at INRAE
To be recruited To be recruited	IR	MIGALE	INRAE	Statistics for integrative bioinformatics
To be recruited To be recruited	IE	To be defined	INRAE	IS5. IE/IR for IS5: multi-omics data for agricultrure
To be recruited To be recruited	IR	To be defined	Inserm	IS3. Bioinformatics for Health
To be recruited To be recruited	IR	PRABI-LBBE	Université Lyon	
Rachel Torchet	IR confirmé	Pasteur	Pasteur	web development, UX design
Cyrille Toulet	IE	Bilille	Univ Lille	Participation to Biosphère
Christine Tranchant-Dubreuil	IE	South Green	IRD	Tool development, data analysis, training
Marie-Stephane Trotard	IE	Genotoul Bioinfo	INRAE	Genotoul bioinfo system administrator, NNCR. WP2.
Pierre Tuffery	DR	RPBS	Inserm	Head of RPBS partner.
David Vallenet	DR	MicroScope	CEA	Co-head of MICROSCOPE partner unit. Provide microbial genomics services through IS4.
Jacques van Helden	PR	IFB-core	CNRS	Mutualisation between national infrastructures for life sciences and health. S1 co-head.
Hugo Varet	IR confirmé	Pasteur	Pasteur	IFB course teacher
Stevenn Volant	IR confirmé	Pasteur	Pasteur	Tools development (SHAMAN)
Anna Zukhova	IR confirmé	Pasteur	Pasteur	methodology development

A5. Detailed equipments and costs

Figure A5.1. Equipment and functioning costs (A) Per region. (B) Per data center



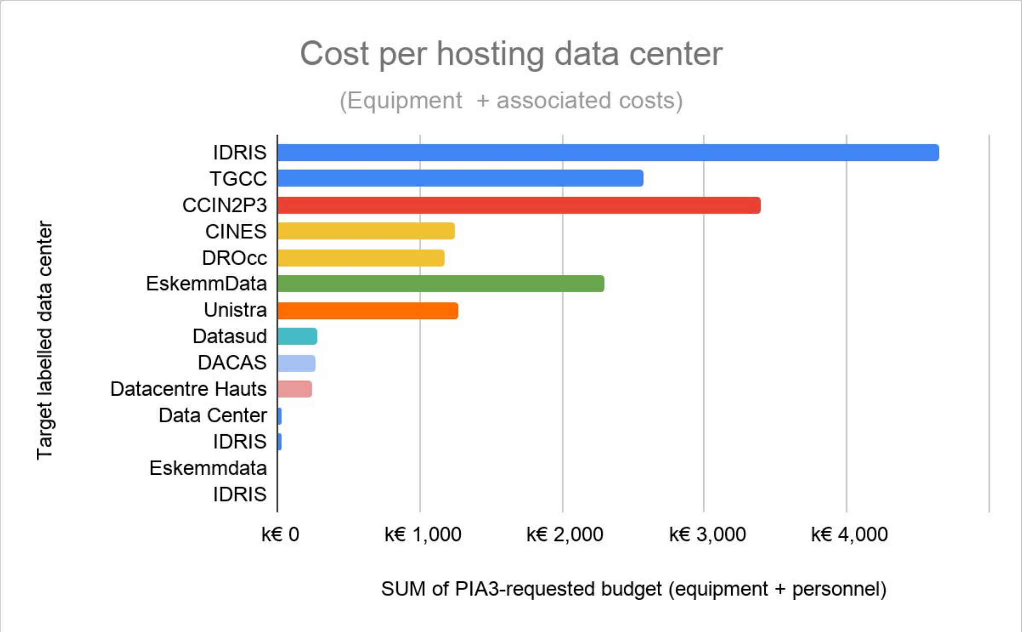


Table A5.2. Total requirements per element (core and regional servers)

Region	Target labelled data center	Nb CPU blades (112 cores)	Nb GPU blades (4 cards)	Nb Fast-access storage arrays (2Po)	Nb Mass storage arrays (2Po)	Nb Backup storage array (2Po)
Element 1 - Compute & Storage equipment for NNCR core resources						
Île de France	IDRIS	30.29	4.62	0.20	1.44	0
Centre-Val de Loire	IDRIS	0.39	0	0.00	0.02	0
Bourgogne-Franche-Comté	IDRIS	0.07	0	0.00	0.00	0
Auvergne-Rhône-Alpes	CCIN2P3	20.42	7.92	0.07	0.63	0
Total		**51.17**	**12.54**	**0.26**	**2.09**	**0**
Element 2 - Compute & Storage equipment for regional IFB platforms						
Provence-Alpes-Côte d'Azur	Datasud	3.78	0	0.00	0.17	0
Pays de la Loire	DACAS	12.46	0	0.08	0.66	0
Occitanie	CINES	2.26	0	0.00	0.10	0
Occitanie	DROcc	30.52	3.3	0.41	1.73	0
Nouvelle Aquitaine	Data Center Régional Nouvelle Aquitaine	3.37	1.32	0.13	0.22	0
Normandie	Eskemmdata	0.07	0	0.00	0.00	0
Île de France	IDRIS	21.51	6.93	0.07	0.55	0
Île de France	TGCC	17.38	0	0.13	0.66	0
Hauts-de-France	Datacentre Hauts de France	6.51	0.66	0.06	0.12	0
Grand Est	Unistra	15.87	2.64	0.09	0.65	0
Bretagne	EskemmData	27.88	6.93	0.14	2.06	0
Auvergne-Rhône-Alpes	CCIN2P3	4.42	0	0.07	0.17	0
Total		**146.02**	**21.78**	**1.18**	**7.08**	**0**
Element 3 - Storage equipment for inter-site data securing						
Provence-Alpes-Côte d'Azur	Datasud	0.00	0.00	0.00	0.00	0.08
Pays de la Loire	DACAS	0.00	0.00	0.00	0.00	0.25
Occitanie	CINES	0.00	0.00	0.00	0.00	0.05
Occitanie	DROcc	0.00	0.00	0.00	0.00	0.67
Nouvelle Aquitaine	Data Center Régional Nouvelle Aquitaine	0.00	0.00	0.00	0.00	0.01
Normandie	Eskemmdata	0.00	0.00	0.00	0.00	0.00
Île de France	IDRIS	0.00	0.00	0.00	0.00	0.55

Hauts-de-France	Datacentre Hauts de France	0.00	0.00	0.00	0.00	0.00
Grand Est	Unistra	0.00	0.00	0.00	0.00	0.33
Centre-Val de Loire	IDRIS	0.00	0.00	0.00	0.00	0.01
Bretagne	EskemmData	0.00	0.00	0.00	0.00	0.51
Bourgogne-Franche-Comté	IDRIS	0.00	0.00	0.00	0.00	0.00
Auvergne-Rhône-Alpes	CCIN2P3	0.00	0.00	0.00	0.00	0.95
Total		**0.00**	**0.00**	**0.00**	**0.00**	**3.41**
Element 4 - Health data hosting and secured research environments						
Occitanie	CINES	0		0	0.66	0
Total		**0**		**0**	**0.66**	**0**
Element 5 - BioDataVerse						
Auvergne-Rhône-Alpes	CCIN2P3	0	0	0	0.33	0.33
Total		**0**	**0**	**0**	**0.33**	**0.33**

Table A5.3. Cost per element (core and regional servers)

Region	Target labelled data center	Total equipment	Total functioning equipment	Total Sub-contracting cost
Element 1 - Compute & Storage equipment for NNCR core resources				
Île de France	IDRIS	€1,841,851.90	€102,008.42	€0.00
Centre-Val de Loire	IDRIS	€20,453.23	€1,077.76	€0.00
Bourgogne-Franche-Comté	IDRIS	€3,473.19	€183.02	€0.00
Auvergne-Rhône-Alpes	CCIN2P3	€1,102,133.82	€47,939.91	€0.00
Total		**€2,967,912.14**	**€151,209.11**	**€0.00**
Element 2 - Compute & Storage equipment for regional IFB platforms				
Provence-Alpes-Côte d'Azur	Datasud	€197,971.79	€10,432	€0.00
Pays de la Loire	DACAS	€756,449.73	€44,629	€0.00
Occitanie	CINES	€118,474.35	€6,243	€0.00
Occitanie	DROcc	€2,108,025.04	€129,691	€19,800.00
Nouvelle Aquitaine	Data Center Régional Nouvelle Aquitaine	€315,696.87	€20,342	€0.00
Normandie	Eskemmdata	€3,473.19	€183	€0.00
Île de France	IDRIS	€1,067,002.20	€43,285	€4,494.60
Île de France	TGCC	€918,720.00	€48,739	€1,001,088.00
Hauts-de-France	Datacentre Hauts de France	€286,675.13	€11,622	€0.00
Grand Est	Unistra	€899,996.17	€46,687	€0.00
Bretagne	EskemmData	€2,144,520.92	€134,411	€0.00
Auvergne-Rhône-Alpes	CCIN2P3	€250,470.00	€14,062	€0.00
Total		**€9,067,475.36**	**€510,326**	**€1,025,382.60**
Element 3 - Storage equipment for inter-site data securing				
Provence-Alpes-Côte d'Azur	Datasud	€22,515.57	€28,148	€0.00
Pays de la Loire	DACAS	€65,461.94	€81,839	€0.00
Occitanie	CINES	€13,474.23	€16,845	€0.00
Occitanie	DROcc	€177,491.16	€221,894	€0.00
Nouvelle Aquitaine	Data Center Régional Nouvelle Aquitaine	€2,501.73	€3,128	€0.00
Normandie	Eskemmdata	€395.01	€494	€0.00
Île de France	IDRIS	€145,583.13	€182,004	€0.00

Hauts-de-France	Datacentre Hauts de France	€570.57	€713	€0.00
Grand Est	Unistra	€86,704.70	€108,396	€0.00
Centre-Val de Loire	IDRIS	€2,326.17	€2,908	€0.00
Bretagne	EskemmData	€136,300.40	€170,399	€0.00
Bourgogne-Franche-Comté	IDRIS	€395.01	€494	€0.00
Auvergne-Rhône-Alpes	CCIN2P3	€253,398.92	€316,792	€0.00
Total		**€907,118.52**	**€1,134,053**	**€0.00**
Element 4 - Health data hosting and secured research environments				
Occitanie	CINES	€573,444.96	€57,752	€379,500.00
Total		**€573,444.96**	**€57,752**	**€379,500.00**
Element 5 - BioDataVerse				
Auvergne-Rhône-Alpes	CCIN2P3	€161,739.60	€222,580	€0.00
Total		**€161,739.60**	**€222,580**	**€0.00**
Element 6 - Fast track access to the Jean Zay Supercomputing facility				
Île de France	IDRIS	€0.00	€0	€558,400.00
Total		**€0.00**	**€0**	**€558,400.00**

A6. Support letters

A6.1. Support letters from partner organisms

Le Directeur Général Délégué à la Science

Monsieur Jacques Van Helden
Inserm U1090 - TAGC
163, Avenue de Luminy
13288 Marseille cedex 09

Paris, June, 15th, 2020

Subject: support letter to the *Equipex+* MuDiS4LS

Within the scope of the Call for proposals "ESR-EquipEx+", the French National Centre for Scientific Research (CNRS) will submit a limited number of projects. These choices are based on several criteria:

- quality of the project and its adequacy with the CNRS priorities
- structuring and unifying impact of the equipment
- equipment's potential for joint utilisation

The project MuDiS4LS meets all these criteria and it is ranked as a priority for the CNRS. The CNRS strongly supports this project and its contributions are estimated, at the time of submission, at 6 097 156,80 €.

The objective of this proposal is to provide the scientific community within the fields of life science and health with an integrated tool to store, to compute, and to share the massive flow of digital data produced by high resolution imaging, next generation sequencing and more. This is a timely project, integrating the growing need of interaction between the experts in the field, with the requirement of easy-access computer power and know-how from the biology and health science sector. It will also be an important step ahead in terms of standardizing exchange protocols for easy access and dissemination of digital scientific data. This project is supported by the national research infrastructure "Institut Français de Bioinformatique" which is one of the components of the European Elixir Infrastructure. This project is a high priority for the CNRS which strongly supports it. Two engineer's positions will be created to guarantee its success and to ensure its sustainability.

Alain Schuhl
Directeur Général Délégué à la Science

Copies :
Monsieur André le Bivic, Directeur de l'institut des sciences biologiques (INSB) – CNRS
Madame Marie-Hélène Papillon, Déléguée régionale CNRS - Ile-de-France Gif-sur-Yvette
Monsieur Ali Charara, Directeur de l'institut des sciences de l'information et de leurs interactions (INS2I) - CNRS

INRAE

La Directrice Générale Déléguée
Science et Innovation

A qui de droit

Objet : Projet MuDiS4LS_IFB

Paris, le 16 juin 2020

INRAE a choisi de ne soutenir en priorité qu'un nombre limité de projets dans lequel il est partenaire et répondant aux critères d'innovation et de structuration de l'Appel à Manifestation d'Intérêt « Équipements structurants pour la recherche / EquipEx+ ». Parmi ceux-ci, « MuDiS4LS_IFB» est un projet stratégique pour INRAE qui a fait de la question des sciences et de la gestion des données une des priorités de son document d'orientation.

Pour INRAE, ce projet concernera principalement les 3 infrastructures scientifiques collectives dédiées à la bioinformatique, labellisées par INRAE et IBISA, ainsi que ses actifs dans l'Infrastructure de recherche (IR) nationale IFB, dont INRAE est l'un des principaux contributeurs en termes de postes permanents. Ce projet est articulé avec d'autres IR nationales, notamment France Génomique, RARe, MTH, IBISBA-Fr. Il est un support important pour le projet MicroWord Discovery dédié à la culturomique et la métagénomique qui sont des thématiques majeures pour INRAE. Il comporte également un volet consacré aux données génétiques et multi-omiques pour l'agriculture.

En développant un cadre qui s'appuiera sur les centres de données nationaux et régionaux labellisés, pour permettre aux scientifiques d'orchestrer les flux de données biologiques, ce projet renforcera le Réseau National de Ressources Informatiques (RNR) développé par l'IFB depuis 2017, dans lequel INRAE est particulièrement impliqué. Il contribuera à apporter des services autour des données et de leur « FAIRisation ». Ce projet est structurant pour INRAE et permettra d'accélérer le développement d'une e-infrastructure fédérative concertée et interconnectée avec les e-infrastructures nationales (DataTerra pour les échelles (supra-)individu et IFB pour les ressources génétiques, l'imagerie et les omiques) et l'insertion du dispositif informatique INRAE dans l'organisation nationale soutenue par le MESRI.

C'est pourquoi, INRAE soutient la demande de subvention et confirme la mobilisation des ressources mentionnées dans les annexes financières, nécessaires à la réalisation du projet. Par ailleurs, outre la mobilisation déjà importante des ressources humaines actuelles, INRAE s'engage à recruter 2 postes d'ingénieurs dédiés. Ce projet complète un projet CPER en région Occitanie pour lequel INRAE s'est également engagé à hauteur de 250 k€.

Le projet MuDiS4LS_IFB constitue donc un des projets prioritaires de INRAE.

Christine Cherbut

Inserm

La science pour la santé
From science to health

Le Président-directeur général

Dossier suivi par :
Mme Morgane Vincent
Chargée de mission
Pôle Partenariats et Politique de Site
Département Partenariats et Relations extérieures
Tél. +33 (0)1 44 23 67 97
E-mail : morgane.vincent@inserm.fr

N/réf. MV/NaB 2020-227

Pr Jacques van Helden
Co-directeur IFB
U1090 (Université Aix-Marseille -Inserm)
63, Avenue de Luminy,
13288 MARSEILLE

Paris, le 1 8 JUIN 2020

Objet : Soutien Projet Equipex + MuDiS4LS

Monsieur,

J'ai pris connaissance avec intérêt de la candidature que vous allez soumettre dans le cadre de l'appel "Equipex +" du programme des Investissements d'Avenir, pour lequel vous avez sollicité le soutien de l'Inserm.

Le projet MuDiS4LS que vous proposez de coordonner est très structurant au niveau national. Il offre aux équipes Inserm un accès à de nouveaux services essentiels pour la bio-informatique médicale et constitue un cadre de cohérence nationale entre des infrastructures nationales majeures en biologie et des centres de calcul et de données nationaux et régionaux. Il permettra aux chercheurs d'orchestrer l'ensemble des flux des données biologiques, depuis leur source de production jusqu'à leur mise à disposition via des dépôts nationaux ou internationaux, en passant par le stockage à chaud et la sécurisation intermédiaire pendant la phase d'analyse et d'exploitation. Rassemblant 4 grands organismes de recherche nationaux et plusieurs universités sur les différents sites, le projet permettra une application concrète des principes FAIR, constituera un outil de pilotage pour rationaliser les moyens humains en calcul et bio-informatique, et contribuera à renforcer la position de la France dans les initiatives européennes tel que EOSC ou ELIXIR.

Le projet est en cohérence avec la politique nationale de l'Inserm. Dans ce cadre, l'Inserm a décidé de participer en co-pilotant un workpackage et en s'investissant dans un second. L'Inserm souhaite contribuer aux besoins de ce projet de haut niveau et s'engage à recruter un ingénieur de recherche supplémentaire spécialisé en bioinformatique en santé pendant la durée du projet Equipex + (8 ans). Cet ingénieur pourra coordonner tous les aspects santé de l'Equipex dans la phase d'exploitation du projet.
L'Inserm soutient ce projet prioritaire qui renforce les investissements pour ses équipes et propose ainsi le meilleur environnement possible à la recherche d'excellence en sciences de la vie et de la santé.

Vous félicitant pour le travail que vous avez effectué et vous formulant mes vœux de pleine réussite dans votre projet, je vous prie d'agréer, Monsieur, mes cordiales salutations.

Dr Gilles Bloch
PDG de l'Inserm

Copies : Franck Lethimonnier, Directeur de l'Institut Technologie pour la Santé
Dominique Nobile, Délégué régional Inserm PACA

Saclay, le 18 juin 2020
N/Réf. : DRF/DIR-20-0242
Objet : Letter of support for MuDiS4LS

To whom it may concern,

The MuDiS4LS project is ambitious, complex yet realistic, since it takes root in a robust infrastructure carried by a strong federation of bioinformatics facilities supported by the main French research organization as well as several medical institutes and universities. The main goal of MuDiS4LS is to develop a framework that will rely on the national and regional data centers to enable scientists controlling the flow of biological data, from their origin (data-producing national infrastructures) to their public release in national or international repositories, while ensuring their mid-term securing during the intermediate phases of analysis and exploitation.

This project is in line with several strategic axes of the CEA

- The project is led in close collaboration with the national infrastructure France Génomique, in charge of developing sequencing facilities
- IFB collaborates with the N4HCloud project promoted by CEA
- The project develops innovative actions to face the challenge of applying Artificial Intelligence approaches to heterogeneous data resulting from the integration of large-scale biology data produced with different high-throughput technologies.

Several implementation studies (IS) are directly involving CEA teams and related to CEA projects;

- IS2 “Marine biology data integration and dissemination”, led in collaboration with TARA and EMBRC-FR.
- IS3 “ Bioinformatics solutions to handle health data”, aiming at developing solutions to handle health and personal genomics data, in collaboration with the Inserm DSI and the CAD of the Plan France Médecine Génomique 2025

This MuDiS4LS project is of great importance to CEA because IFB is the only national infrastructure able to address the challenges of integrative bioinformatics, encompassing a broad range of application including omics data, imaging, structure, dynamical modelling of complex biological systems, etc. This integration has of course to be done in collaboration with the other national infrastructures oriented towards the production and treatment of specific data types, and with whom IFB is already engaged in several collaborations.

For all these reasons, CEA fully supports the MuDiS4LS project, for its high potential impact and scientific interest.

Yours sincerely,

Elsa Cortijo,
Director of the Fundamental Research Division of CEA

Inria

Paris, 18 June 2020

To Whom It May Concern:

Inria is a contributor of the "Institut Français de Bioinformatique" (IFB) INBS, as partner of this infrastructure. Our involvement in IFB is important for us because of our expertise in software development and maintenance. Our main involvement in IFB is through the GeneOuest platform, IFB's main platform for computing and storage resources and bioinformatics software deployment. GeneOuest also provides IFB with a significant number of datasets. Our project-teams Genescale and Dyliss, in particular, are strongly implicated in the development of software tools and datasets available on the GeneOuest platform.

In this framework, we consider that MuDiS4LS is a crucial step towards higher standards in data science for French biological and health sciences. Life and medical sciences increasingly depend on the availability of interoperable and open data sets and of the software tools necessary to analyze those data. Because new analysis tools, incl. AI, demand data volumes that are larger and larger, mutualization of data sets and software pieces has become a necessary goal in the field. However, to deliver its full potential, this approach needs high quality data qualification and standards. Data collection, aggregation and exchange between depositaries will also have to be automated and normalized. The objectives of MuDiS4LS with its machine-actionable data management plans seem a solid proposal to enforce these goals. MuDiS4LS should permit the emergence of a national actor for the organisation, distribution and storage of data for life sciences.

For those reasons, Inria judges MuDiS4LS will be a crucial asset for the French academic community in life science research because it will endow it with state-of-the-art data exchange organization. Therefore, we fully support MuDiS4LS application to the PIA3 ANR call.

Sincerely yours,

Stéphane Ubéda
Directeur du Centre Inria Rennes Bretagne Atlantique

Le Directeur du centre de recherche
INRIA Rennes-Bretagne Atlantique
Stéphane UBÉDA

GENCI

Paris, June 19th, 2020

To whom it may concern

O/Ref. : CD200611

Subject : Support letter to MuDiS4LS project

Dear Sir,

As part of its new strategic plan for 2019-2023, GENCI, the French large scale Research Infrastructure for high performance computing has the mission to establish itself as the reference solution to respond to new computing and data processing needs of public sector research institutions ; it is also committed to develop and adapt the capabilities to new challenges through the continued diversification of the resources access modes to computing and data ; one of its main goal is then to support end-to-end complex workflows of data from instruments to data centers.

In this respect, GENCI is supportive of MuDiS4LS project (*Mutualised Digital Spaces for FAIR data in Life and Health Science*) to foster the collaboration with the French research infrastructure IFB on the building of an integrated platform of distributed data and FAIR services deployed across and supported by a continuum of 9 science-driven infrastructures (the so-called NNCR for *National Network of Computing Resources* launched by IFB in 2017).

Thanks to the partnership with GENCI represented in the project MuDiS4LS by CINES, IDRIS and TGCC, that is the three leading national HPC centers, IFB and its partners will then have access to means of calculation, data processing and AI-based analysis combining high-performance computing, high flow rate computing exploiting HPC and Cloud architectures with accelerators, farms of data processing, and containerization.

We do appreciate the opportunity to express our support for this proposal and wish its team a successful outcome from the proposal selection process.

Yours Sincerely,

Philippe LAVOCAT
CEO of GENCI

Institut de Recherche pour le Développement
FRANCE

44 boulevard de Dunkerque - CS 90009
13572 Marseille cedex 02 - FRANCE
tél. +33 (0) 4 91 99 95 10
fax +33 (0) 4 91 99 92 15
d2s@ird.fr

Le Directeur Délégué à la Science

Marseille, le 17 juin 2020

Monsieur Jacques van Helden
Institut français de Bioinformatique (IFB)
IFB-Core, Génoscope
2 rue Gaston Crémieux,
91057 - ÉVRY – CEDEX

N/Réf. : D2S-2020-n°7
Objet : Lettre de soutien au projet MuDiS4LS: Mutualised Digital Spaces for FAIR data in Life and Health Science.

Monsieur le responsable du projet Gaia Data, cher collègue,

L'objectif principal de MuDiS4LS est de développer un cadre qui s'appuiera sur les centres de données nationaux et régionaux pour permettre aux scientifiques d'orchestrer les flux de données biologiques, depuis leur source (infrastructures nationales productrices de données) jusqu'à leur diffusion publique via des dépôts nationaux ou internationaux, tout en assurant leur sécurisation à moyen terme lors des phases intermédiaires d'analyse et d'exploitation.

L'Institut de recherche pour le développement (IRD) est un organisme pluridisciplinaire reconnu internationalement, travaillant principalement en partenariat avec les pays méditerranéens et intertropicaux. Les priorités de l'IRD s'inscrivent dans la mise en œuvre, associée à une analyse critique, des Objectifs de développement durable (ODD) adoptés en septembre 2015 par les Nations unies, avec pour ambition d'orienter les politiques de développement et de répondre aux grands enjeux liés aux changements globaux, environnementaux, économiques, sociaux et culturels qui affectent la totalité de la planète.

La création d'une politique de gestion de données de l'IRD, notamment via l'implémentation progressive des principes FAIR, s'inscrit dans la gouvernance globale des données de la recherche de l'institut et dans le cadre d'un partenariat équitable avec les partenaires du Sud. En particulier, les principes d'interopérabilité et de réutilisation des données sont une priorité pour l'institut car ils sont essentiels afin de faciliter les approches interdisciplinaires et de promouvoir la science de la durabilité. Par ailleurs, afin de prendre en compte l'empreinte environnementale des données produites par la science, l'IRD a à cœur d'étendre les principes FAIR à la dimension environnementale (FAIRS, Sustainable). Pour toutes ces raisons l'IRD soutient fortement le projet MdDiS4SL et ses objectifs de mutualisation d'espaces numériques.

Avec mes amicales salutations,

Philippe CHARVIS

copie : Mme Valérie Verdier, Présidente-Directrice Générale,
M. Franck Carenzi, Directeur de la Mission d'appui au partenariat et à la science (MAPS)

A6.2. Support letters from European Nucleotide Archive (ENA) for data brokering

EMBL-EBI

Guy Cochrane, Team Leader and
Head of European Nucleotide
Archive
Tel.: +44 (0) 1223 492564
E-mails cochrane@ebi.ac.uk

17 June 2020

Dear Dr. Médigue and Prof. van Helden,

I have been very interested to hear from you the details of your proposed "Espaces numériques mutualisés pour des données FAIR en biologie-santé (MuDiS4LS)" project that you will submit to the Investissements d'Avenir programme. I write to express my enthusiastic support for the proposal and to state my plans with regard to the important collaboration that this will enable between our respective groups.

As the Head of the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home), the ELIXIR Core Data Resource for sequence and associated data, I have very specific interests in the project. Indeed, the fullest and most impactful delivery of the ENA – achieved through deep compliance with community data standards, the following of FAIR principles and active support of data generating communities - will increasingly rely on projects such as yours.

The data brokering elements of your project will bring our most direct collaboration: here, our established programme of thematic and domain-specific data submissions brokering to ENA and current drive to engage brokers at the national level, will directly connect to the brokering infrastructure that you will build. We will make available the relevant ENA brokering tools, support in the use of these tools and actively engage with staff working on these elements of your project. Here I expect to welcome to EMBL-EBI visitors from the project, work closely together in practical workshops and enjoy other forms of close collaboration. I note in particular your wise plans to staff the brokering work with both technical experts able to track our brokering interfaces and data structures as they evolve over time, and also experts in specific biological domains able to "translate" between end users and informatics services.

Further areas of synergy include our mutual interest in data standards and the application of knowledge around these during early development stages of scientific projects to achieve FAIR data. I hope that we will work closely on data standards around sequencing technologies and push together for the inclusion of standards into data management planning and its formalisation.

I wish you luck with this important proposal and look forward to our fruitful collaborations.

Yours sincerely,

Guy Cochrane, PhD

Head of the European Nucleotide Archive (ENA)

A6.3. Support letters from other ESR/Equipex+ projects

INVESTISSEMENTS D'AVENIR METABOHUB PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR ANR

Toulouse June 18th 2020

Prof Dominique Rolin
Directeur de MetaboHUB (2013-2020)
71 Avenue E Bourlaux, Campus Vert de la Grande Ferrade
33882 Villenave d'Ornon
Tel : 05 57 12 26 90 Email : rolin@bordeaux.irnae.fr

Dr Fabien Jourdan
Directeur de MetaboHUB2.0 (2021-2025)
UMR1331 TOXALIM
180 Chemin de Tournefeuille
Tel : 05 82 06 63 95 Email : fabien.Jourdan@irnae.fr

To:
Claudine Médigue et Jacques
van Helden
Directrice et cp-directeur
IFB

Object: MuDiS4LS Project Support Letter from MetaboHUB

Dear C. Médique and J. van Helden,

We, the undersigned Prof. Dominique ROLIN and Dr. Fabien JOURDAN, are writing to express our support to the MuDiS4LS project, coordinated by IFB (Institut Français de Bioinformatique). You have informed us that you are applying for the "ESR/EquipEx+" call for the development of the project entitled "Mutualised Digital Spaces for FAIR data in Life and Health Science".

Regarding your Equipex+ project, we have noted that your intention is to reinforce the development of open science through a FAIR strategy including machine actionable DMP, streamlining a secured, national computational infrastructure, mutualizing data spaces and structuring the bioinformatics community. The strong expertise and the know-how of your consortium are clear elements which will guarantee the realization of this ambitious project.

MetaboHub, national infrastructure of metabolomics and fluxomics, is developing cutting edge methods and tools in the field since 2013. We already have the chance to develop shared project among which the internationally acknowledged galaxy workflow for metabolomics (Workflow4Metabolomics, W4M). Beyond this shared method development, MetaboHub is strongly interacting with IFB on data storage and high-performance computing.

MetaboHub project MetEx+ submitted to the Equipex+ call aims at developing data interoperability in the field of metabolomics. This is fully in synergy with MuDiS4LS and in particular the open science support which this project will offer. Notably MetaboHub will be a relevant partner to develop and deploy machine actionable DMP.

For all these reasons we fully support the MuDiS4LS project and wish every technical and economic success to be able to use the platforms at the end.

Sincerely,

Prof D. Rolin Directeur de MetaboHUB (2013-2020) Rolin	Dr F. Jourdan Directeur de MetaboHUB2.0 (2021-2025)

Institut de Génomique
Fonctionnelle de Lyon

Lyon, June 18th 2020

Dr. Yad Ghavi-Helm, Group leader
Institut de Génomique Fonctionnelle de Lyon (IGFL)
ENS de Lyon / CNRS / Univ Lyon
46 Allée d'Italie
69364 Lyon cédex 07, France
04 26 73 13 50
yad.ghavi-helm@ens-lyon.fr

Letter of support for the MuDiS4LS project

To whom it may concern,

The members of the Spatial-Cell-ID project, federating 11 research units (IGFL, RDP, LBMC, SBRI, CRNL, INMG, CIRI, LBTI, LEM, MAP), 4 service units (UMS Biosciences, SFR Lyon-Est, PSMN, CBP) in Lyon and three industrial collaborators (Cellenion, Vidium and Leica) propose to create a national spatial transcriptomics equipment to identify and characterize the transcriptome of each cell and the spatio-temporal dynamics of cellular heterogeneity in tissues from animal and plant organisms, whether adult or in development, healthy or sick.

Beyond setting up the experimental imaging and genomics platforms, the research laboratories behind Spatial-Cell-ID are faced with the challenge:

- to organize the flow of data from the instruments (i) to appropriate local HPDA equipments for their analysis and exploitation, (ii) to the analysis and storage platforms used by visiting colleagues, who come to Lyon to carry out their experiments and (iii) to national or international repositories for the eventual public release of the data.
- to leverage the experience and competence of colleagues from all over France in finding a way to install their data analysis pipelines on machines with direct access to Spatial-Cell-ID data. The success of our project critically depends on the integration of the most advanced data analysis tools to handle the extra layer of complexity from the concomitant generation of sequences and images from the same experimental system.

In this task, we are locally supported by the *Pôle Scientifique de Modélisation Numérique* (PSMN) and the *Centre Blaise Pascal* (CBP) of the ENS de Lyon. These resource and competence centers in computational science across all disciplines have a large experience in running HPC, HPDA and cloud services. Currently they operate 700 highly mutualized servers with more than 15,000 cores in the state-of-the-art Datacenter of the ENS de Lyon, which is part of the regional CINAuRA ESR data center.

The national bioinformatics infrastructure pioneered by the *Institut Français de Bioinformatique* provides an extremely useful framework for connecting Spatial-Cell-ID to researchers from laboratories all over France. The simultaneous involvement of the PSMN and the CBP as the local computing facility in Spatial-Cell-ID and as a HPDA and cloud node in IFB's MuDiS4LS project guarantees an optimal integration of Spatial-Cell-ID into the IFB network. Through the MuDiS4LS project, national and regional infrastructures will be continuously upgraded to fit the evolution of life sciences and confront the tremendous increase of the demand for computing, storage and data transfer. This point will be especially important for Spatial-Cell-ID since it will help to face yet unknown challenges in computing and storage resources.

We also share MuDiS4LS' aim to encourage biologists to adopt information technologies for their research, which in line with our own efforts to disseminate a numerical culture that should help biologists to become fluent in analysis and modeling of their own datasets. The members of the Spatial-Cell-ID consortium fully subscribe to the MuDiS4LS proposal and its strategic axes of developing Findable, Accessible, Interoperable and Reusable (FAIR) software for life and health science, of hosting local computing resources in state-of-the art regional data centers, of mutualizing these resources and data spaces in a national network (NNCR), of developing synergies between mesocenters like the PSMN/CBP and national computing and data centers, of developing mutualized services, and of joining international projects like the European Science Cloud.

For all of the above reasons, the Spatial-Cell-ID consortium warmly supports the MuDiS4LS project: it is centered on a very timely and needed structuration at the national scale, which fits the national roadmap. It is also well integrated in the larger national network of imaging and omics. It will represent a very valuable asset in the full development of the potential of the Spatial-Cell-ID project for generation of ground-breaking discoveries in cell and developmental biology.

Best regards,

Yad Ghavi-Helm, scientific coordinator of the Spatial-Cell-ID ESR/EquipEx+ project

UNIVERSITÉ
CÔTE D'AZUR

Aix*Marseille
université
Initiative d'excellence

Nice, June 17th, 2020

Subject: Letter of support to the project Mutualised Digital Spaces for FAIR data in Life and Health Science (MuDiS4LS)

To whom it may concern,

By the present letter Université Côte d'Azur (UCA) and Aix-Marseille Université (AMU) are pleased to endorse the project MuDiS4LS coordinated by Institut Français de Bioinformatique (IFB).

The Universities are particularly interested in developing new collaborations with IFB through both Equipex projects submitted to this call: 4D-Omics, coordinated by Université Côte d'Azur and MuDiS4LS coordinated by IFB.

The main goal of MuDiS4LS (Mutualised Digital Space for Life Sciences) is to develop a framework that will rely on the national and regional data centers to enable scientists to orchestrate the fluxes of biological data, since their source (data-producing national infrastructures) to their public release via national or international repositories, whilst ensuring their mid-term securing during the intermediate phases of analysis and exploitation.

To be more specific, UCA and AMU would be willing to provide the IFB with an access to our resource to implement specific projects. In parallel, the Universities are particularly interested in the IFB expertise regarding the development of data management plans, data FAIRification, and by solutions regarding large-scale computing.

The Universities are confident that such collaboration will be possible as we can already rely on an ongoing collaboration through the management of the OMERO database. Indeed, the implementation of the database at IN2P3 and its usage in Nice though the Imaging platform was made possible partly thanks to the support of the IFB. On this subject, it is interesting to point out, that the database should soon be moved from IN2P3 to Data Center Sud. Finally, it can be mentioned that this work was already recognized nationally, with the CNRS Cristal award given to 5 engineers (2 from the MICA platform in Nice, 2 from Villefranche and 1 from IFB) involved in the creation and operation of the tool, which confirms the potential of the model.

As such, through this letter, Université Côte d'Azur and Aix-Marseille Université express their interest in the activities to be conducted in the MuDiS4LS project and, should the proposal be selected, wishes to be regularly updated on the project achievements and resulting services. In addition, good coordination can only be beneficial to both projects and therefore 4D-OMICS invited IFB to participate in its Scientific Advisory Board.

Sincerely,

Pour le Président d'Université Côte d'Azur
et par délégation,
Le Vice-Président
Recherche et Innovation

Noël DIMARCQ

Jeanick Brisswalter President of UCA	Eric Berton President of AMU