OpenLink

A data management dashboard for research teams

Clermont Ferrand University

January 17, 2023

Laurent Bouri, Nadia Goué, Mateo Hiriart, Julien Seiler

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

BiGEst

AuBi
Auvergne Bioinformatique

cnrs INRAE Inserm cea

About IFB : Missions and actions

Digital infrastructure

- National Network of Computing Resources (**NNCR**)
- High performance computing : **cluster** + **cloud**
- Data protection (**GDPR**)

11 Po storage
21K cores

Software

- Specialized software **development**
- **Packaging** (conda)
- **Virtualization** (appliances, containers)
- **Best practices** in software engineering
- **Collaborative development, open code**

More than **900 tools** available

Users support

- A single point of entry for all IFB services and platforms
- Support for projects from the design stage
- Data management plan (**DMP**)
- Design and implementation of **workflows**
- **Data Science**

400 experts
>200 FTE

9115 users
310 collaborative projects

Mutualised Task Forces

Knowledge bases

- Development of data and knowledge bases
- Quality standards
- Technical support for DB development
- Deployment of databases developed in France

Formations

- **Thematic schools** (NGS, multi-omics, phylogeny, biostat, programming, workflows…)
- **Bring Your Own Data** (BYOD) training
- Webinars, MOOC
- **Newx: FAIR-data and FAIR-bioinfo training**
- Permanent adaptation to the evolution of the demand

136 trainings
2326 people trained
341 days of training

Foresight and innovation

- Identification of emerging bioinfo needs
- Pilot projects to meet needs:
  - Integrative bioinformatics
  - Health data integration
  - Artificial intelligence for biology and health

Open Science & Interoperability

- Data management
- Interoperability, standards, ontologies...
- Data brokering
- Machine actionable DMP
- Cooperation with data-producing infrastructures

About IFB : The scope of IFB in 2022

A distributed digital infrastructure

- 36 platforms and research teams
- wide geographic coverage
- encompass all areas of bioinformatics expertise

22 member-platforms
7 contributing platforms
8 associated tems
>400 experts (~200 FTE)

3 membership applications in progress

Pooled resources

- Infrastructure financing:
  PIA1 RENABI-IFB (22.8M€), PIA3 MUDIS4LS (16.5M€)
- missions articulated around “Task Forces” pooled between platforms

Collective projects

- EMERGEN, ABRomics, PEPR

Coordination: IFB-core (UAR 3601)

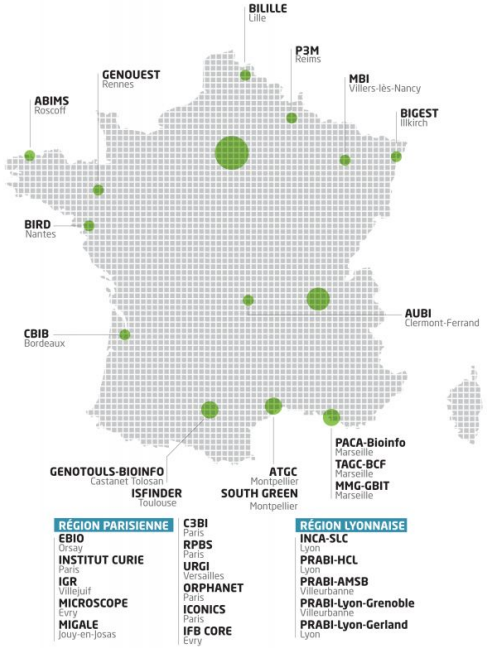
- multi-tutelle (CNRS, Inserm, INRAE, CEA)
- mission : coordination and management of pooled resources
- ELIXIR french node (ELIXIR-FR)

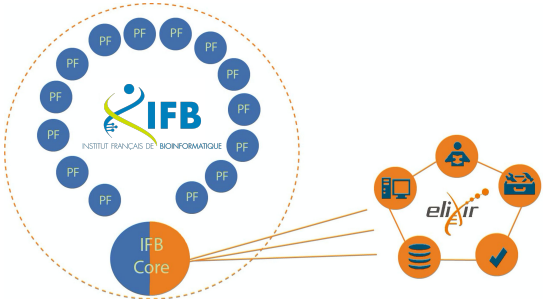
Reaching all communities

- biology, health, agriculture, environment

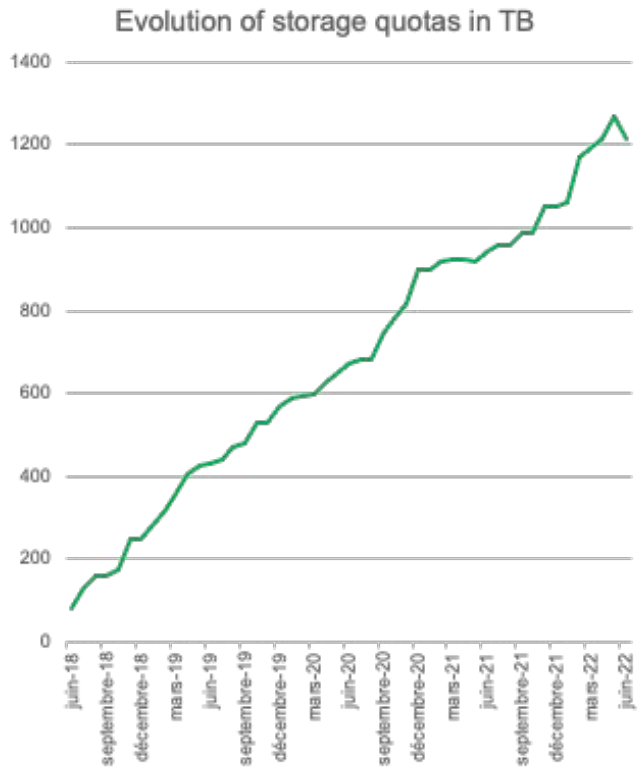
Strong link with Europe

- french node of ESFRI ELIXIR
- entry point for european projects





A project born at IGBMC / BiGEst



IGBMC

Institut de Génétique et de Biologie Moléculaire et Cellulaire

Member of the BiGEst IFB Platform

42 research teams

570 researchers and engineers

From 60TB to 2,2PB in 12 years

Since 2018, the average increase in storage requirements is 24TB per month

Only 30%* of data stored concern active projects

*Estimate based on a survey of research teams in 2020

Real life Open Science is hard

Adopting the F.A.I.R. principles is a bumpy road

Keep track of data

Different instruments / infrastructures / data management software

*Research projects need **clear organisation** from the ground-up*

Keep track of data description (meta-data)

Electronic lab notebook ? DMP ? Companion files ?

*We need **better tools***

Use interoperable formats

Each domain (instrument) has its own format

Real life Open Science is hard

Publish data to the right repository

Domain specific or generalist repositories

Not always designed for "mere mortals".

*We need **data brokering solutions***

zenodo

figshare

Generalist Repositories

GenBank

GEO Gene Expression Omnibus

IDR

UniProt

MetaboLights

Domain specific Repositories

What if… 🤔

Keep track of data description

Keep track of data

Use interoperable formats

Publish data to the right repository

2019 : ANR Open Science Flash Call

-> Accelerate F.A.I.R. and Open Science adoption in all communities

Let's propose something :

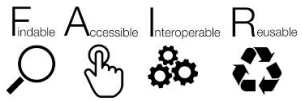
- IT department
- Imaging center
- 3 research teams

OpenLink

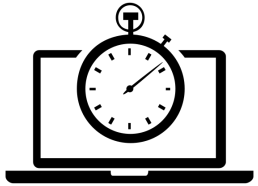
What is **OpenLink**

Goals

A **clear view of the data associated** with each research project



Reduce barriers to adoption of FAIR principles



Limit the impact of FAIR on data management time

Assist researchers in **publishing data**

What is OpenLink

Solution

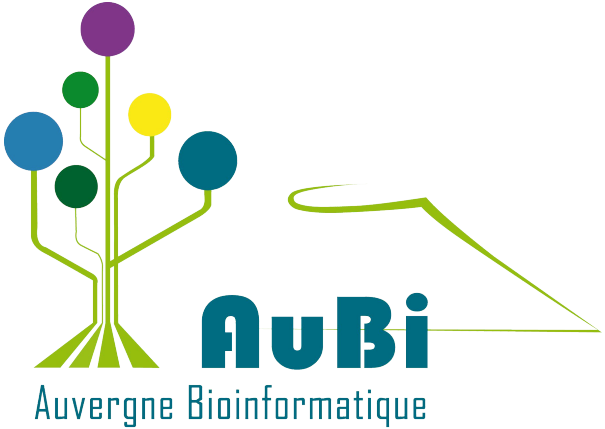
An **open-source web application** based on the Django framework (Python language)

A database to create **links** between a research project and multiple data sources

Extensible and pluggable

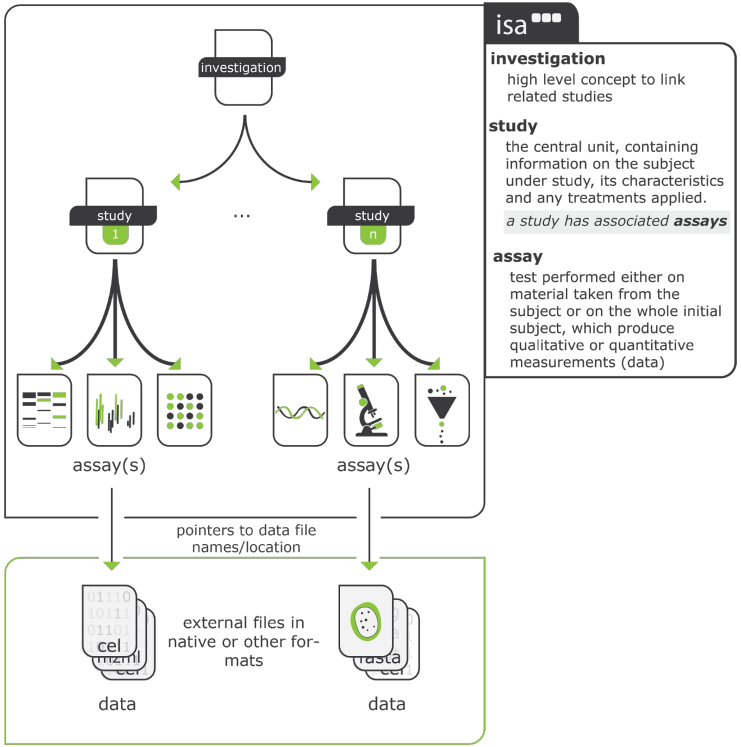
Integrated tools to facilitate **data manipulation**

New supports



ISA model to describe an Openlink Project

https://isa-specs.readthedocs.io/en/latest/isamodel.html



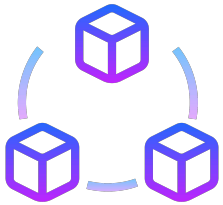
Investigation: The main objectives of a project

Study: A particular biological hypothesis, which you plan to test in different ways

Assay: Experiments, measurements or models

Connector concept in **Openlink**

Connectors describe **how to access a Tool** via his **API**



Connectors are **modular**

2 kinds of **connector**:
Mapper and **Publisher**

Connectors currently available in **Openlink**

Labguru
an electronic lab notebook

Omero
an image visualization, management and analysis tool

Galaxy
a set of tools for manipulation and analysis of genomic data

Seafile
a file synchronization and sharing solution

SSH
a network protocol giving a user a secure way to connect to a computer

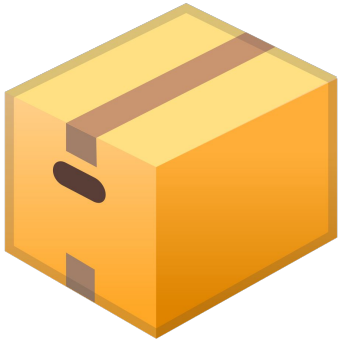
Zenodo
a universal repository for scientific research data

Mapper

Publisher

Publish to Zenodo

Download data



Create archive

Upload to Zenodo
with metadata

Openlink Demo

OpenLink

Home

About

Logout

Ibouri

Admin

Contact

Version 1.0.2.rc5

Demo openlink | Edit | Manage Users | Manage Tools | Add investigation

› Demo openlink

Data distribution

Omero | Galaxy

Omero: 265.67 KB / Galaxy: 31.91 KB / total: 297.58 KB

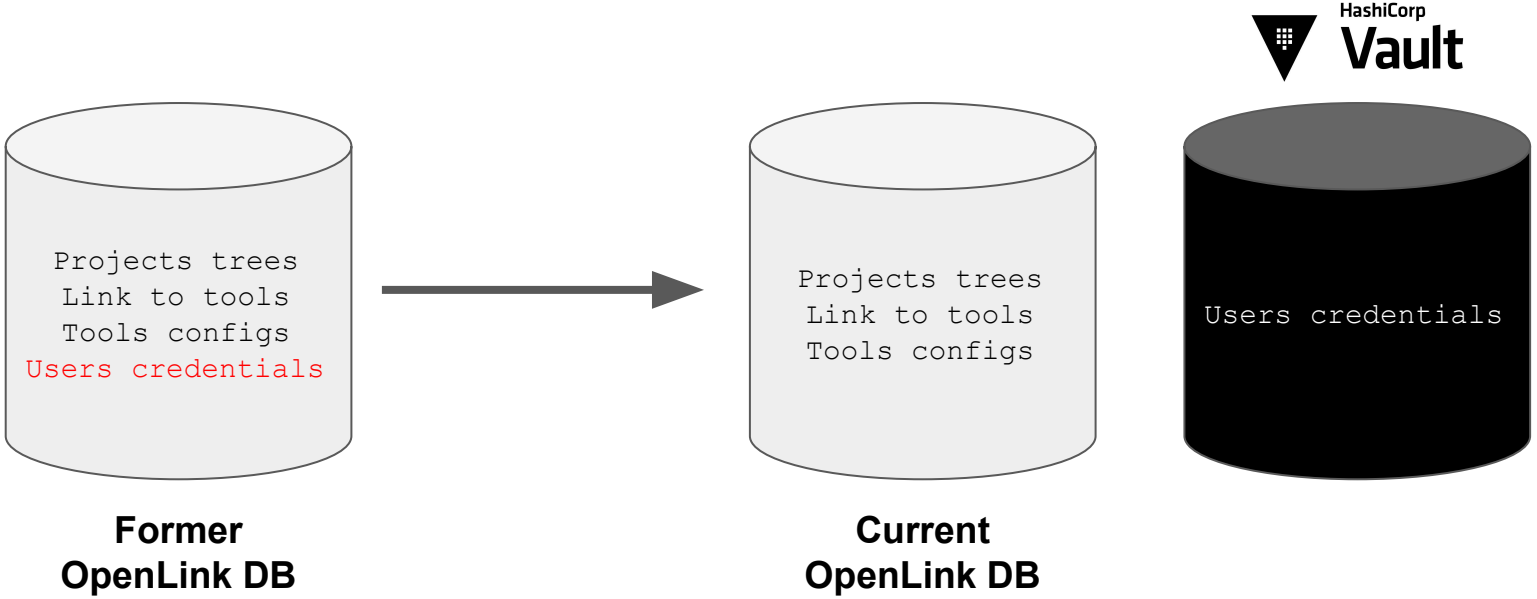
- Growth control of the eukaryote cell [labguru]
  - Metagenomes and Metatranscriptomes of p… [labguru]
    - metagenome sequencing [labguru] [galaxy]
      - Arabidopsis_thaliana.fasta [galaxy]
    - eukaryotic cells imagery [labguru] [omero]
      - img cells data [omero]

Expand All | Collapse All

Asynchronous tasks

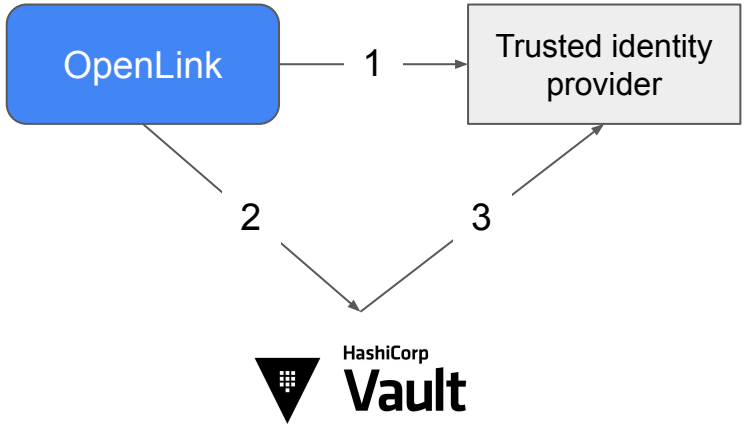
Recent improvements

- Handle **user access** to sensitive data



Recent improvements

- Handle **user access** to sensitive data

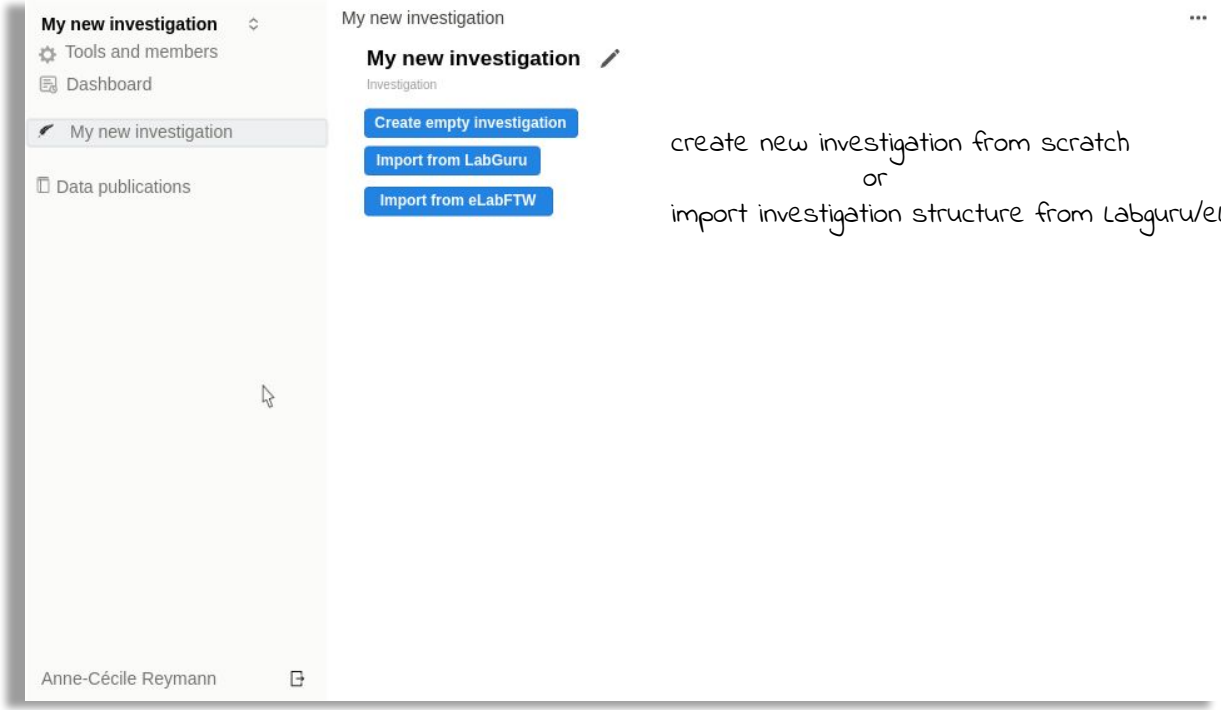


1. Delegate authentication to trusted identity provider
2. Request access to vault with identity token
3. Verify identity token against identity provider

- OpenLink can't access users credentials outside a user session
- Encryption and strict policies for users credentials
- No credentials sharing between users
- OpenLink can enforce data access verification

Upcoming improvements

- Refactor user interface (introducing Openlink JS client en OpenLink API)



create new investigation from scratch
or
import investigation structure from Labguru/eLabFTW

Upcoming improvements

- Refactor user interface (introducing Openlink JS client en OpenLink API)

Screen splitted in two panels

ACT-2 Mutants-PredACTINg

Tools and members
Dashboard
ACT-2 Mutants-PredACTINg
Injections
Actin amplification
Test in vitro
Informations
Crosses pLeu65Val
ACR041pLeu65Val backcro...
2020-11-03
pLeu65Val (rey012) x LifeA...
Phenotypes
Data publications
Anne-Cécile Reymann

ACT-2 Mutants-PredACTINg / Crosses pLeu65Val rey012

Crosses pLeu65Val rey012
Study

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin ullamcorper, ante nec pretium efficitur, mi sapien varius orci, in ultricies dolor est at diam. Pellentesque magna ligula, vulputate ut posuere at, lacinia ut odio. Nunc urna risus, fringilla a lectus at, euismod mattis tellus. Nam accumsan risus sollicitudin, tempus orci ut, ullamcorper leo.

Data links New link

Crosses pLeu65... LabGuru folder
final_result Space2 directory 1,02 GB
report.docx Seafile file 254 MB

Assays New assay

ACR041pLeuVal backcross
2022-11-03
pLeu65Val (rey012) x LifeAct::mKate

Data distribution
Seafile 254 MB
Space2 1,02 GB

Investigation Tree

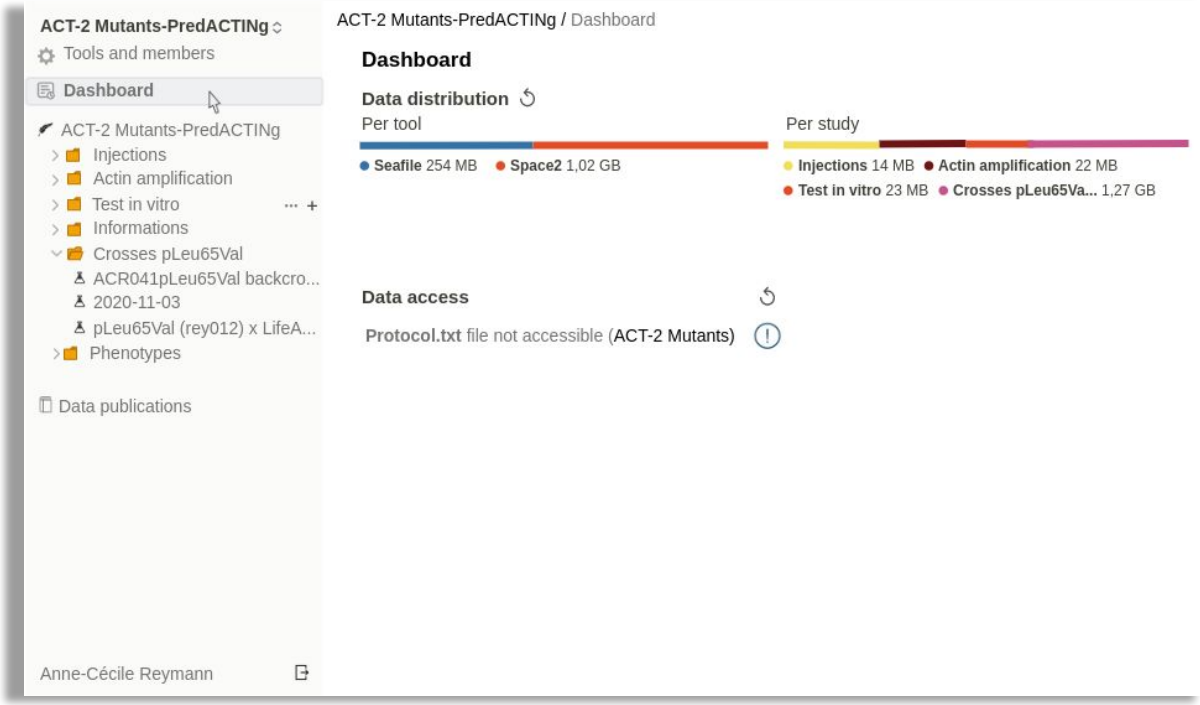
Secondary information area

More visible data links

Direct access to sub-objects

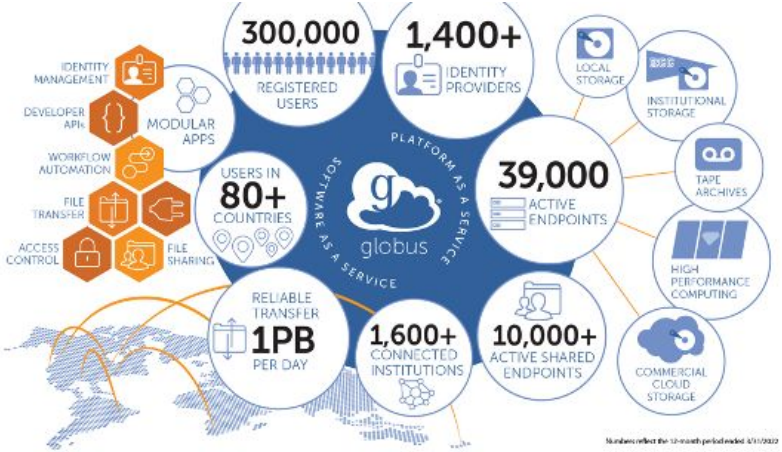
Upcoming improvements

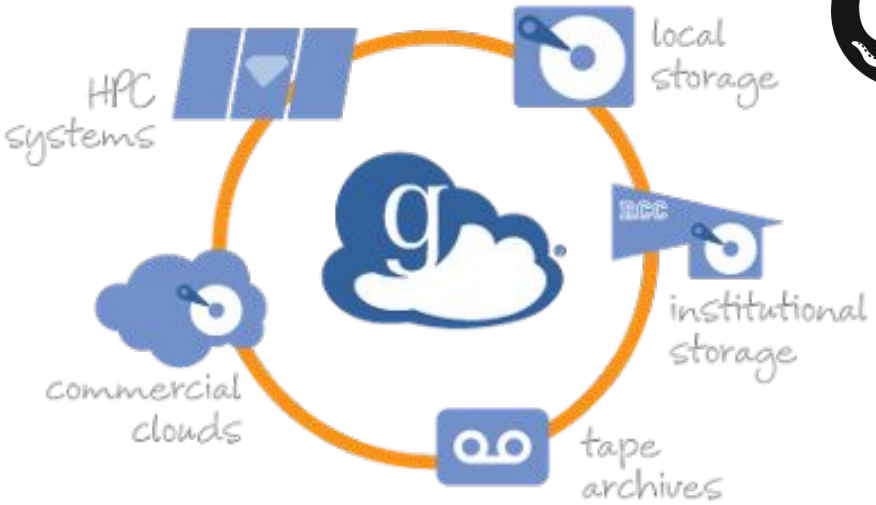
- Refactor user interface (introducing Openlink JS client en OpenLink API)



Upcoming improvements

- Refactor user interface (introducing Openlink JS client en OpenLink API)
- Develop new Openlink **connectors** for Globus, Git and more tools…





Upcoming improvements

- Refactor user interface (introducing Openlink JS client en OpenLink API)
- Develop new Openlink **connectors** (Globus, Git, ...)
- Develop a **REST API** for accessing the OpenLink database
- Interface Openlink with **Data Brokering tools** proposed by the **IFB**
- Improve **metadata Management**

About metadata

Metadata are currently not handled by OpenLink

Several ways to retrieve metadata :

- From data files (fasta/fastq, tiff, etc.)
- From data storage tools (Omero, Galaxy, …)
- From production context (OpenLink links each dataset to the project ISA model)

Several types of metadata :

- Sample description (species, location, date, …)
- Experiment description (provenance, protocol, project name, participants, …)
- Data acquisition (hardware parameters, creator, creation date, …)
- Data processing (processing pipeline description, tools versions and parameters, etc.)

Metadata are currently very fragmented and OpenLink could play a role in centralizing and classifying metadata.

The main challenge is not metadata description (we have many ontologies) but how to handle and store them.

On the way to production

- Openlink is currently in **Test Phase:**

  IGBMC Imaging researchers *as group of beta testers*

  First users feedbacks are essentials **!**

- **Production instances will be available "soon" :**
  - AuBi
  - BiGEst
  - IFB Core Cluster

- Add **ELIXIR Authentication and Authorization Infrastructure (AAI)**

  Use your university or institute account to access Openlink

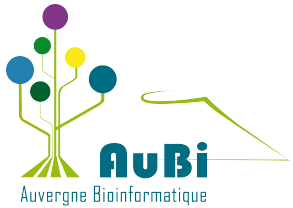
Support Openlink

- Contribute to code : https://gitlab.com/ifb-elixirfr/openlink

- Fund the project :
  - Institut Français de Bioinformatique
  - Cellule Science Ouverte de l'UCA / AuBi

- Join the Openlink workgroup :
  - Slack (https://ifb-openlink.slack.com)
  - Friday meeting on Zoom

Thanks



Mateo Hiriart

Nadia Goué

Research team and platforms

Juliette Godin
Bertrand Vernay
Anne-Cecile Reymann
Erwan Grandgirard
Elvire Guiot
Nicolas Torquet

IT department

Guillaume Seith

Fred de Lamotte

Paulette Lieby