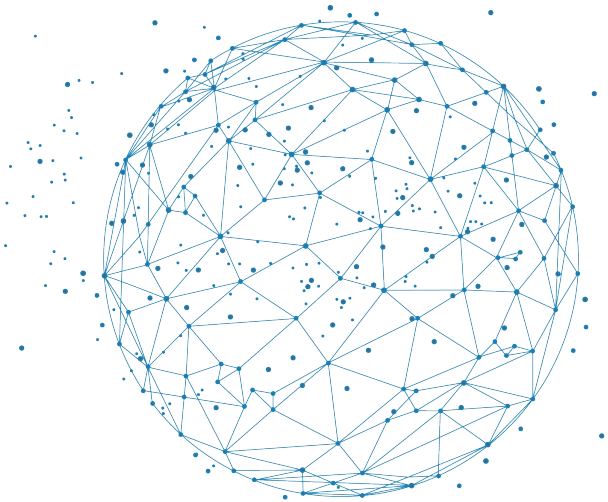
X/Stra - 5 avril 2023



# OpenLink

## Tableau de bord des données
## Passerelle vers la science ouverte

**Julien Seiler**

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea

## Julien Seiler

**IR CNRS**

Directeur des Systèmes d'Information

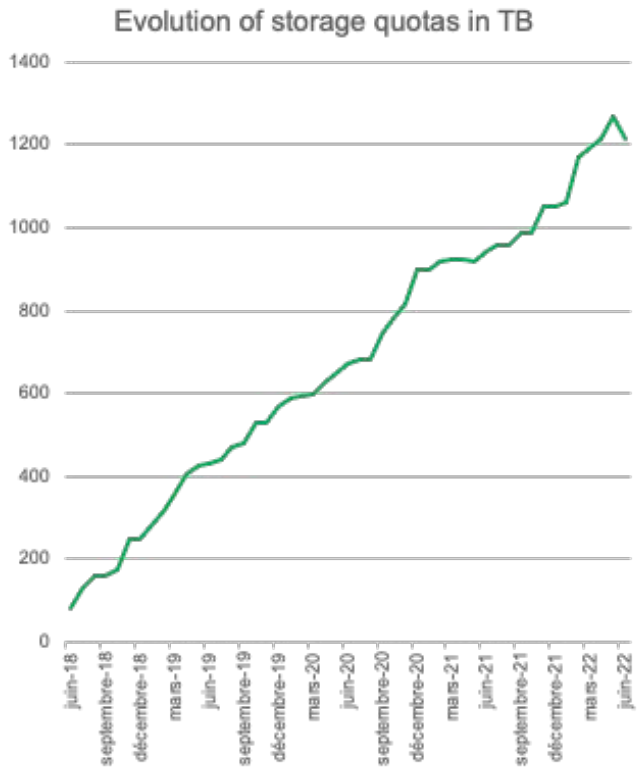
igbmc institut de génétique et de biologie moléculaire et cellulaire

Co-Responsable de l'Infrastructure Nationale de Calcul

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Chef de projet OpenLink

igbmc institut de génétique et de biologie moléculaire et cellulaire IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE




**La réalité du déluge de données à l'IGBMC**

*Institut de Génétique et de Biologie Moléculaire et Cellulaire*

42 équipes de recherche

570 chercheurs, ingénieurs et doctorants

**Passage de 60TB à 2,2PB en 12 ans**

Depuis 2018, l'augmentation moyenne des besoins de stockage est de 24To par mois.

Seulement 30%* des données stockées concernent des projets actifs

**Les chercheurs accumulent les données sans stratégie de conservation à long terme**

**Estimation basée sur un sondage auprès des équipes de recherche IGBMC en 2020*

# L'adoption des principes F.A.I.R.

un chemin semé d'embûches !

**Assurer le suivi des données**

Stockage NAS ? cloud ? cluster ? local ?

**Conserver les méta-données (description des données)**

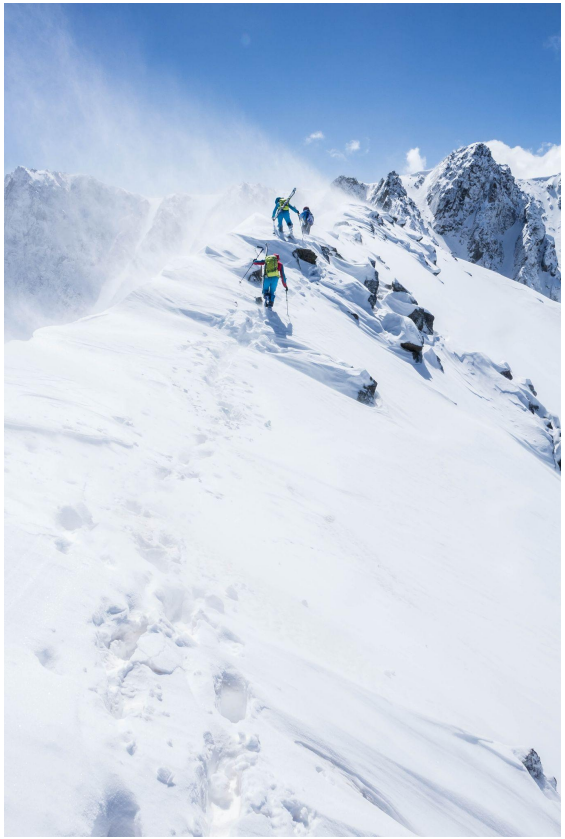
Cahier de laboratoire ? Fichiers compagnons ? Outils spécialisés ?

**Choisir des formats interopérables**

Chaque domaine scientifique, chaque instrument, chaque outil propose son propre format !

**Trouver le bon dépôt pour chaque données**

GenBank · GEO Gene Expression Omnibus · UniProt · IDR · MetaboLights · zenodo · figshare



IFB Institut Français de Bioinformatique · elixir France · République Française Liberté Égalité Fraternité recherche.data.gouv.fr

On ne peut pas s’engager vers la Science Ouverte à l’improviste


✅ Un objectif (choisi ou imposé)

✅ Un bon Plan (de Gestion de Données)

✅ Les bons outils

Et si on pouvait…

Identifier facilement l’ensemble des données associées à un projet de recherche

Accéder au contexte de production et la description de chaque donnée

Accompagner la publication des données

# L'idée d'OpenLink

En 2019, l'appel à projet Flash Science Ouverte de l'ANR permet au projet **OpenLink** de démarrer à l'IGBMC

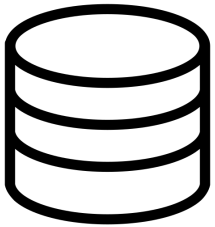
Un consortium 100% IGBMC

- Département informatique (porteur)
- Plateforme d'imagerie
- 3 équipes de recherche

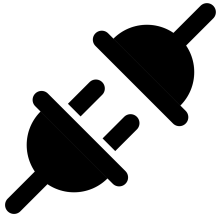
96K€ de financement sur 24 mois

## Objectifs

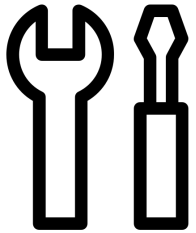
Une **application web open-source** basée sur le framework Django (langage Python)



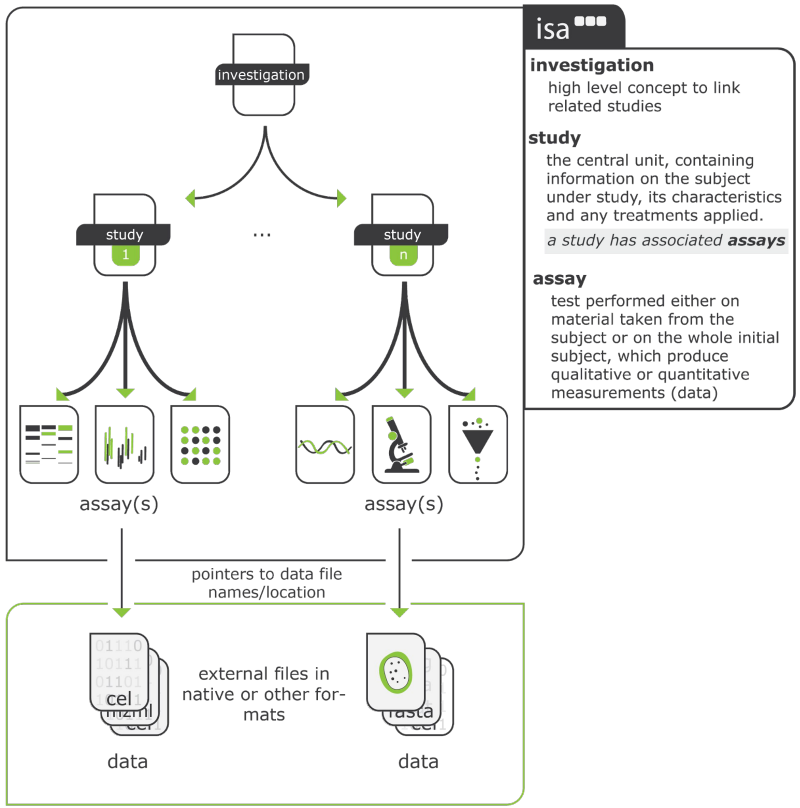
Une base de données permettant de créer des **liens** entre un projet de recherche et de multiple sources de données



Une architecture **évolutive**



Des outils intégrés pour faciliter la **gestion des données**




## Le modèle ISA

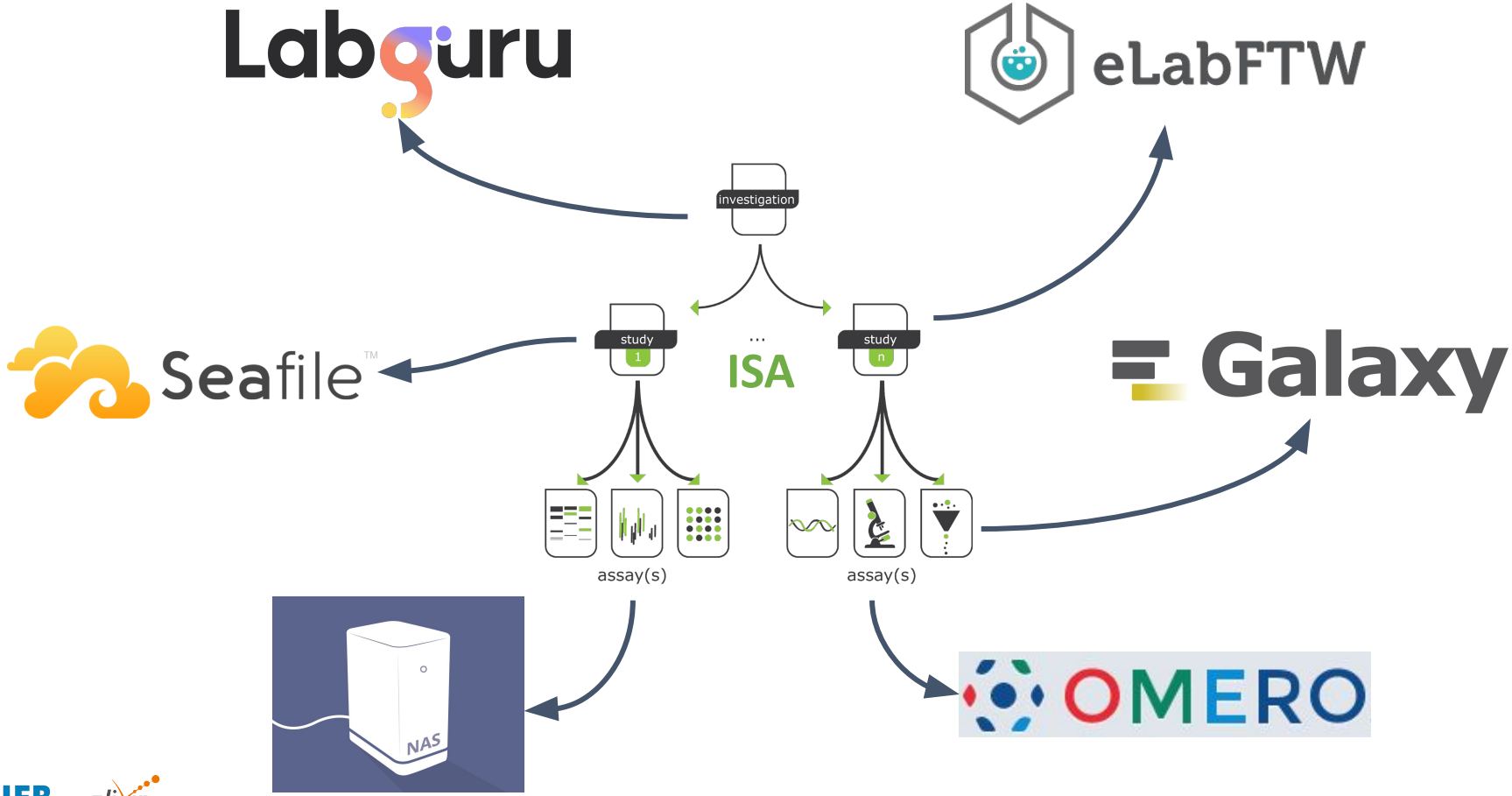
**Investigation** : un objectif du projet

**Study** : une hypothèse biologique que l'on souhaite tester

**Assay** : une expérience, une mesure, un modèle permettant de tester l'hypothèse

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

# Deuxième étape : accéder aux données



# Deuxième étape : accéder aux données


6 connecteurs DataLink :

- LabGuru
- eLabFTW
- Seafile
- Galaxy
- OMERO
- SSHFS

et bientôt plus…

Pour chaque Connecteur :

- Les paramètres de connexion
- La logique permettant l'accès aux données/ressources (via une API ou un autre protocole standard)

`La bibliothèque de connecteurs d'OpenLink est facilement extensible et s'appuie sur le modèle des Apps Django.`

Un autre type de connecteur : les Publishers

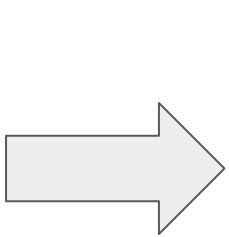
Publisher

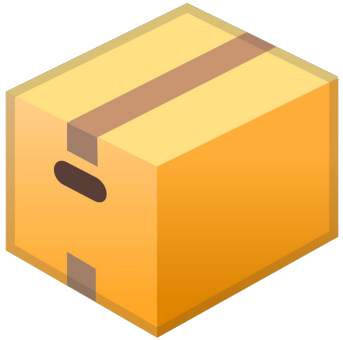
zenodo

**Entrepôt**

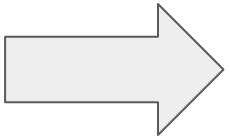
# Troisième étape : publier les données

Télécharger les données



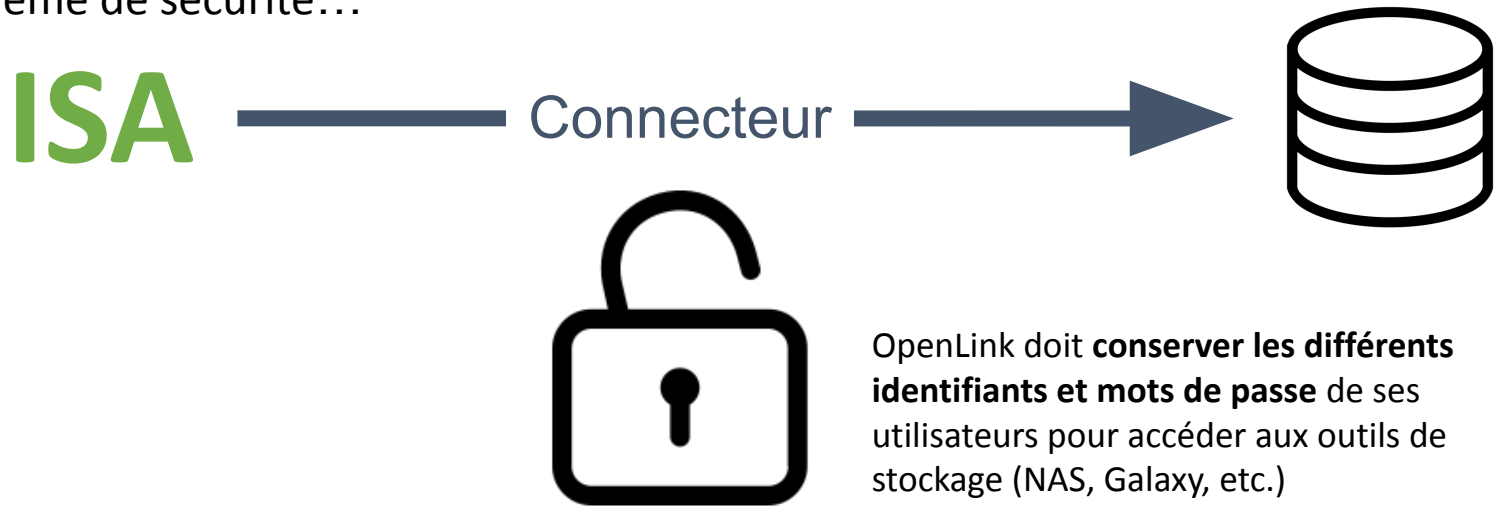


Créer une archive



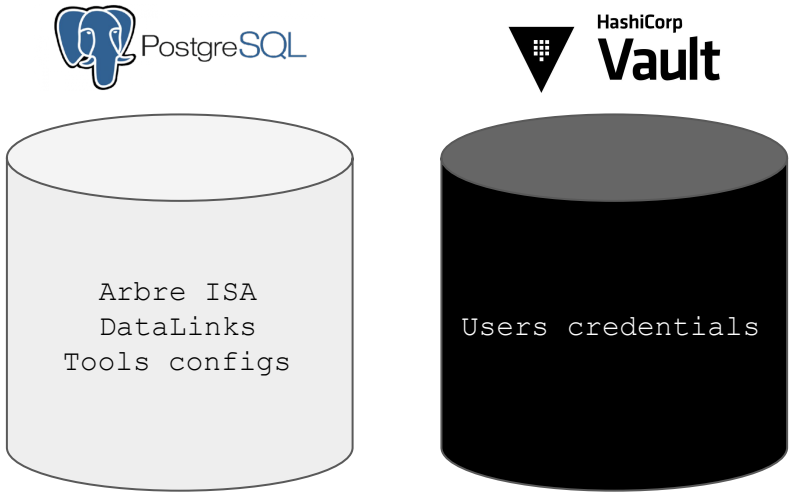
Téléverser sur Zenodo avec les métadonnées

Un problème de sécurité…




OpenLink doit **conserver les différents identifiants et mots de passe** de ses utilisateurs pour accéder aux outils de stockage (NAS, Galaxy, etc.)

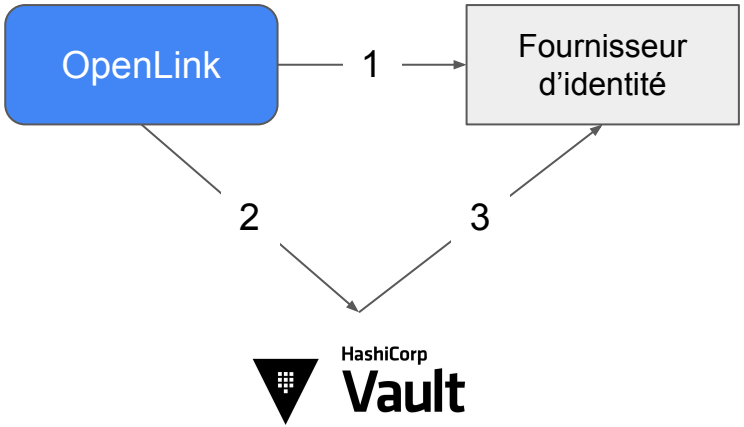
## Les bases de données d'OpenLink




⚠️ OpenLink **ne stocke pas** de données des projets de recherche !

Les données restent où elles se trouvent.

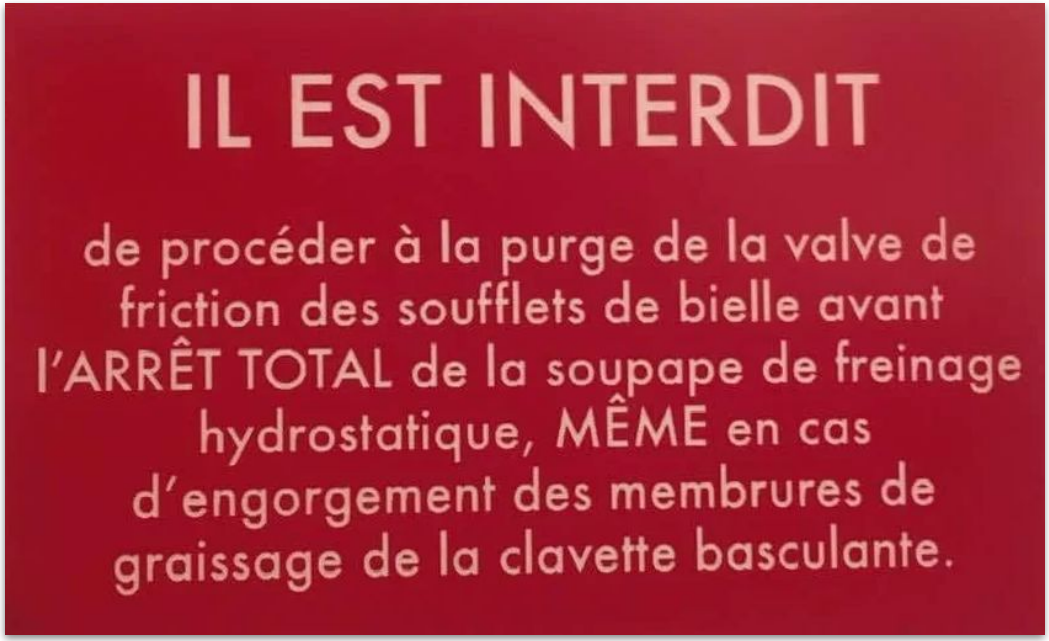
## Limiter l'accès aux données sensibles (identifiants et mots de passe)




1. Déléguer l'authentification à un fournisseur d'identité de confiance
2. Demander l'accès au Vault à l'aide d'un jeton d'identité fourni par le fournisseur d'identité
3. Vérifier l'authenticité du jeton d'identité

- OpenLink n'accède aux mots de passe de l'utilisateur que lorsqu'une session est active
- Les mots de passe sont chiffrés dans le Vault et leur accès est strictement limité à l'utilisateur propriétaire
- Les mots de passe ne sont jamais accessibles à un autre utilisateur

- OpenLink peut vérifier l'accès aux données pour chaque utilisateur




Ca a l'air simple...
Mais on veut bien une démo quand même...

# OpenLink Proto - Démonstration

**OpenLink**
**v1**
2020-2022

*Les erreurs de jeunesse*

**OpenLink**
**v2**
Depuis 2023

*L'app de la maturité*

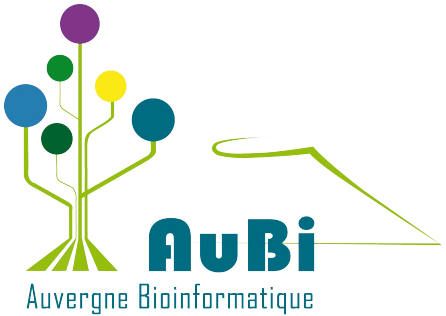
# La poursuite du projet - de nouveaux partenaires et contributeurs

Le projet OpenLink est toujours porté par l'IGBMC avec un groupe de chercheurs participant aux tests du logiciel.

L'IGBMC est membre de la plateforme IFB BiGEst

Le projet OpenLink a été adopté par l'Institut Français de Bioinformatique

OpenLink contribue aux grands axes stratégiques de l'institut (Science Ouverte, Data brokering, Centre de référence thématique)



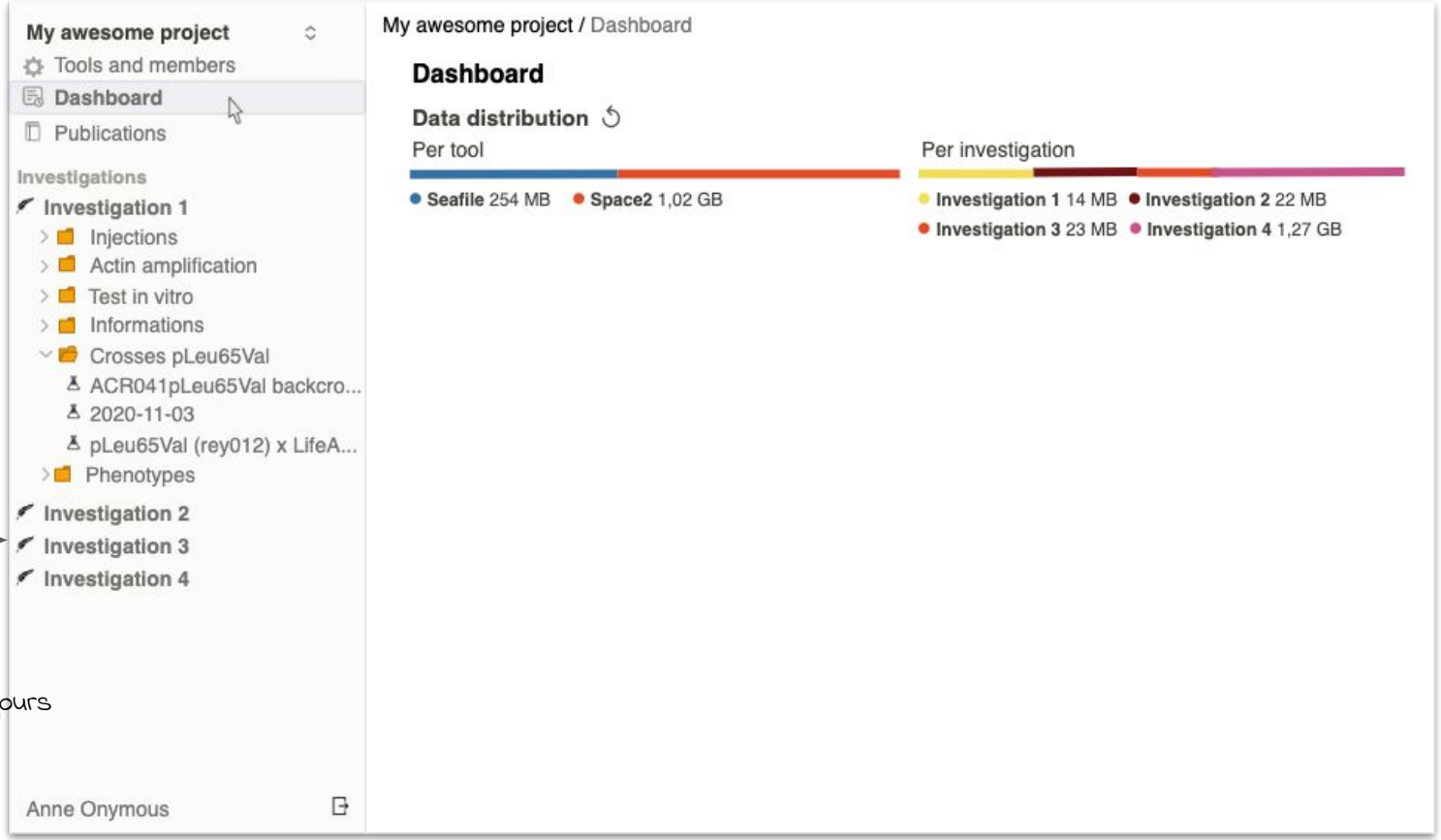

OpenLink bénéficie d'importantes contributions de la plateforme AuBi (Auvergne Bioinformatique)

- Sécurisation des accès avec Vault
- Connecteur Galaxy
- Fournisseur d'identité LDAP
- Et plus encore…

Un groupe de travail accompagne le projet et se réunit **tous les vendredis sur Zoom**

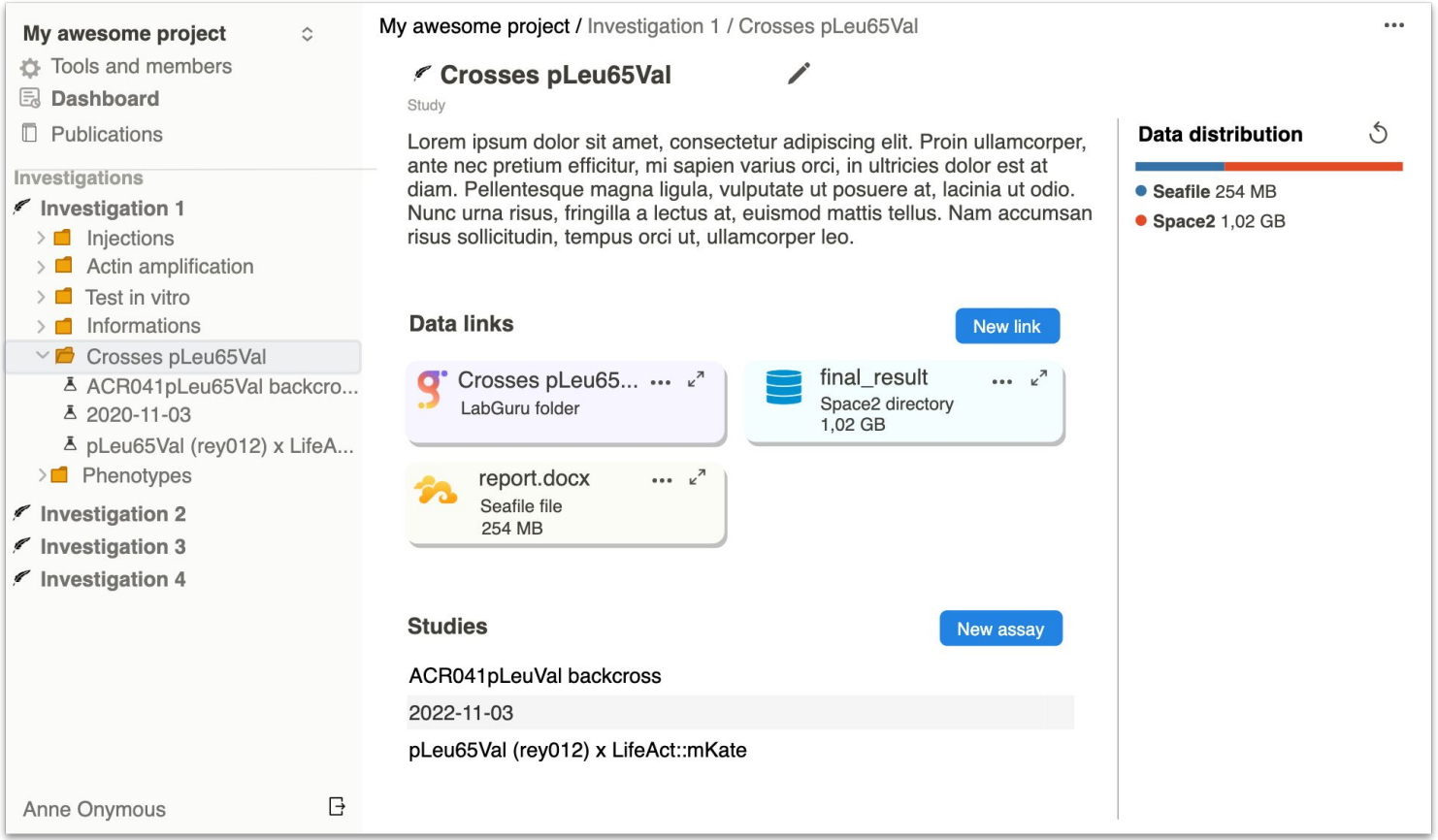
Il est composé de profils hétérogènes (Data Steward, Bioinformaticien, Administrateurs Système, Expert en imagerie, Développeurs webs, etc.) venant de différents horizons (INRAE, Université d'Auvergne, CNRS, INSERM, etc.)

# La poursuite du projet - Refonte graphique




Arbre ISA toujours accessible

# La poursuite du projet - Refonte graphique




NUXTJS


Vue.js


django
REST
framework

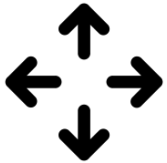
# La poursuite du projet - nouveaux connecteurs



# La poursuite du projet - de nouveaux outils

Vérificateur de format

Audit d'accès



Déplacement de données

Une API ouverte permettra à des applications tiers d'interagir avec OpenLink

1. Accès à l'arbre ISA des projets
2. Accès aux DataLinks
3. Accès aux connecteurs

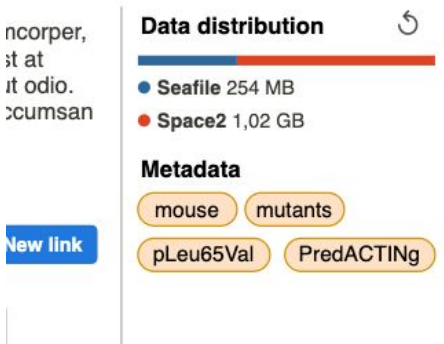
# La poursuite du projet - et les métadonnées ?

Les métadonnées ne sont actuellement pas supportées par OpenLink

Une veille est en cours pour identifier la meilleure approche pour gérer les métadonnées :

- Est-ce le rôle d'OpenLink ?
- Comment retrouver automatiquement les métadonnées associées à chaque DataLink (depuis les fichiers, depuis les outils de stockage, depuis le contexte ISA…)
- Sous quelle forme mettre en place un outil de gestion des métadonnées ?

METARK



# La poursuite du projet - sur le chemin de la production

- OpenLink est actuellement en **phase de test**
  Un groupe de chercheurs de l'IGBMC évalue le logiciel
  Retours utilisateurs essentiels

- Un **chantier important** est en cours pour la refonte graphique du logiciel : plusieurs semaines de travail nécessaires

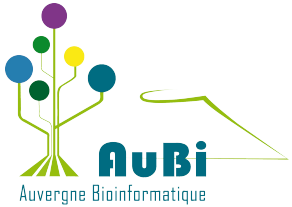
- Une **instance de production** hébergée par l'IFB sera ouverte dès que possible courant 2023

# Contribuer à OpenLink

- Contribuez aux développements : https://gitlab.com/ifb-elixirfr/openlink

- Rejoignez le groupe de **travail** :
  - ~~Slack (https://ifb-openlink.slack.com)~~
- - Mattermost (https://team.forgemia.inra.fr)
  - Réunion Zoom le vendredi (9h à 10h)

# Merci !






*Mateo Hiriart*

*Nadia Goué*

Equipes de recherche et plateforme

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

Service informatique

*Laurent Bouri*
*Guillaume Seith*

*Fred de Lamotte*

*Paulette Lieby*

*Thomas Denecker*