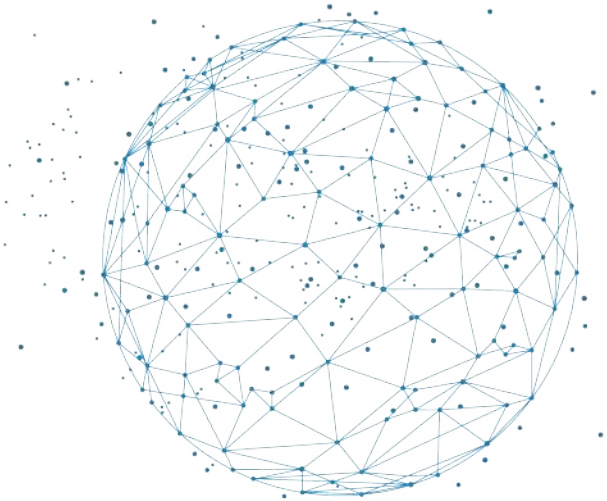
# Madbot

Thomas Denecker, Julien Seiler

Roscoff- 12 mars 2024



# Madbot

## Une passerelle vers la science ouverte

**Julien Seiler**
IGBMC & IFB - https://orcid.org/0000-0002-4549-5188

**T. Denecker**
IFB - https://orcid.org/0000-0003-1421-7641

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

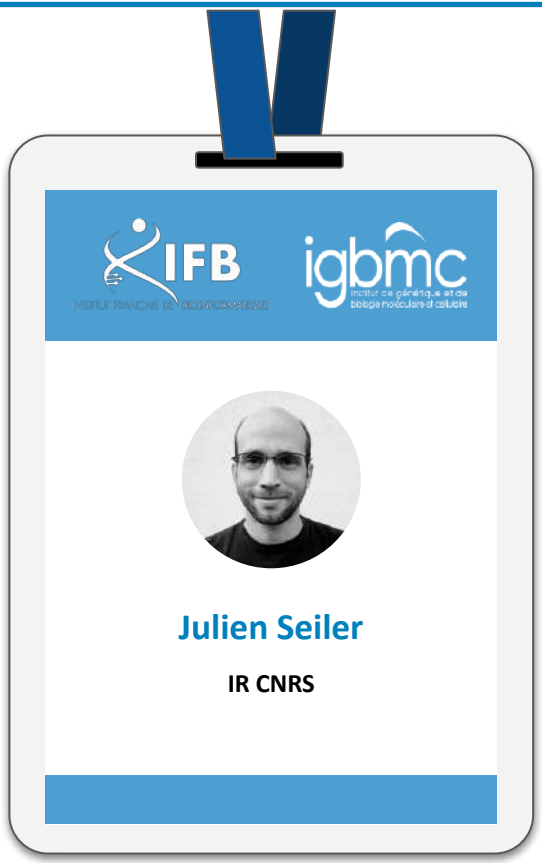
elixir FRANCE

cnrs INRAE Inserm cea

# Plan

- Contexte de la science ouverte
- Pourquoi nous en sommes arrivés à développer madbot ?
- Qu’est-ce que c’est madbot ?
- La suite du projet

# Qui sommes-nous ?



**Julien Seiler**

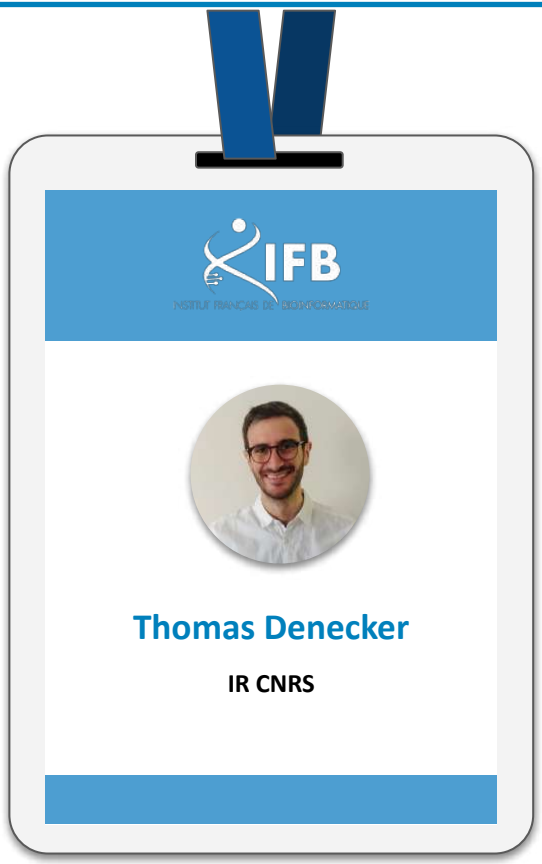
**IR CNRS**

Directeur des Systèmes d'Information (IGBMC)

Co-Responsable de l'Infrastructure Nationale de Calcul (IFB)

Chef de projet Madbot (IFB & IGBMC)

# Qui sommes-nous ?



**Thomas Denecker**

**IR CNRS**

Co-Responsable “Développement et déploiement logiciel”

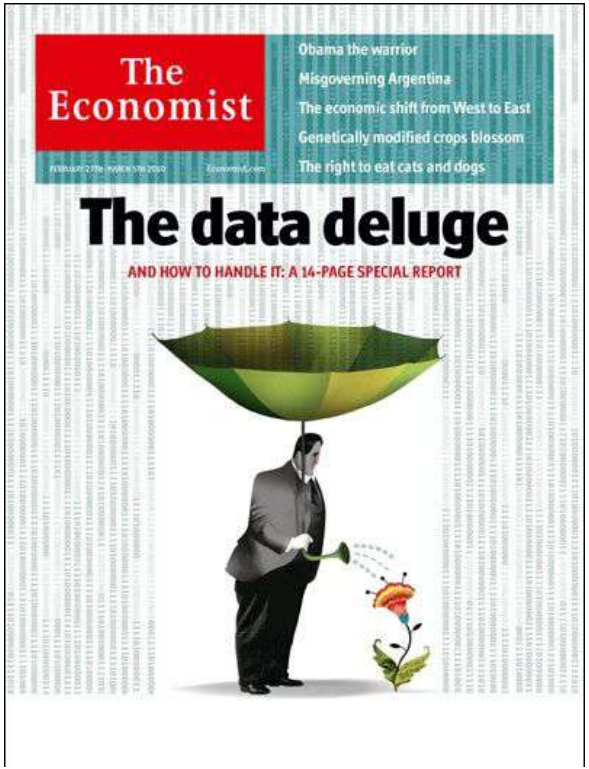
Chef de projet Madbot

Formateur sur la thématique FAIR

# Contexte

## Science ouverte

# De plus en plus de données

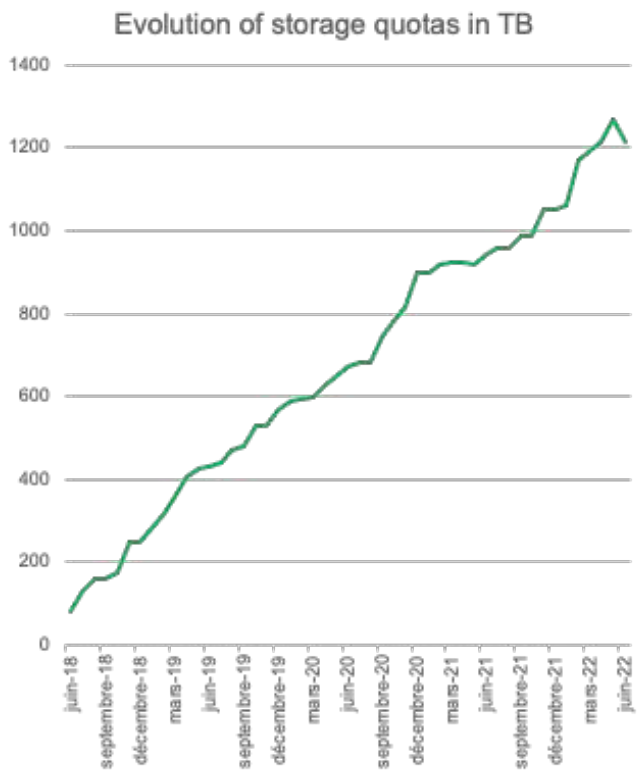



*Data is the new oil*

Clive Humby

*Data is the new oil? No: Data is the new soil.*

David Mccandless




**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

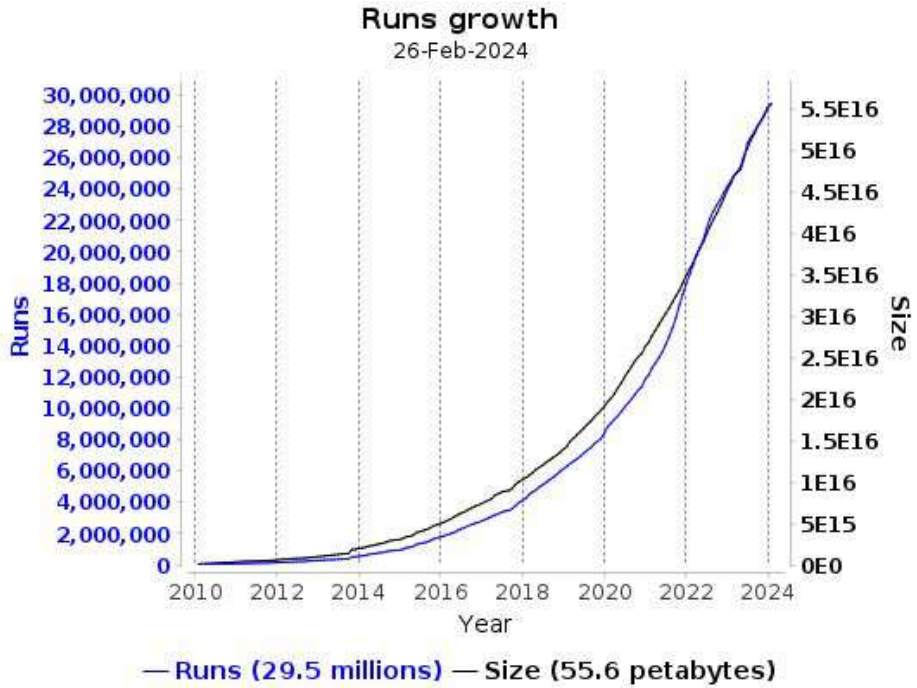
**Les chercheurs accumulent les données sans stratégie de conservation à long terme**

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

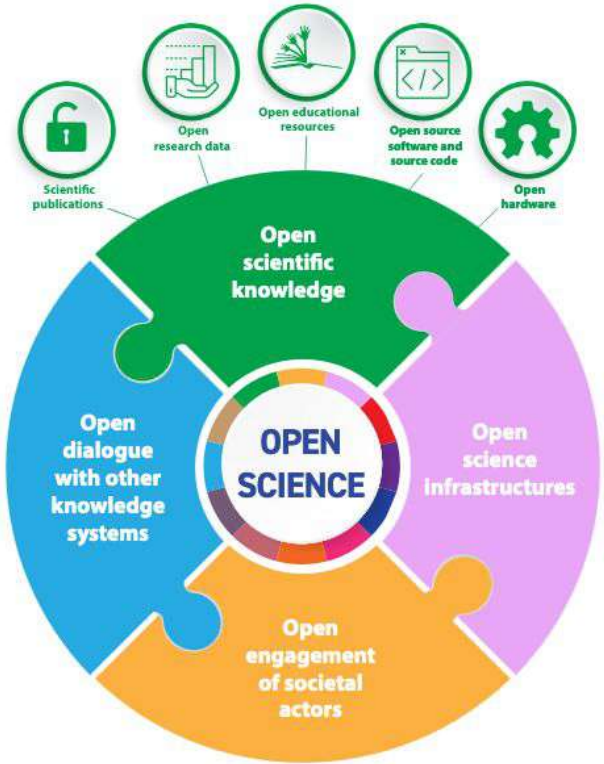
ENA
European Nucleotide Archive

https://www.ebi.ac.uk/ena/browser/about/statistics

Et c'est pareil dans PRIDE, SRA,...



# Recommandations de l'UNESCO pour la science ouverte

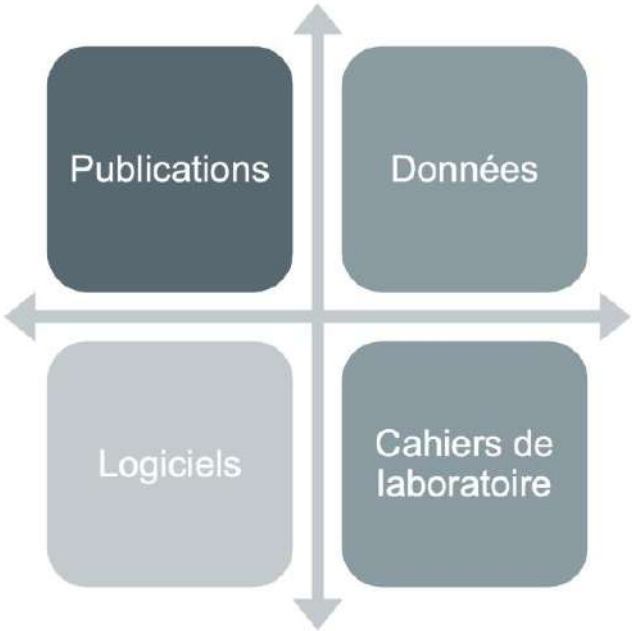



**UNESCO Recommendation on Open Science, nov 2021**
https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

Objectif : rendre la recherche accessible à tous

- Pas seulement l'accès à la **connaissance** elle-même
- Tout le processus de **création** et de **dissémination** de celle-ci
- La possibilité de **réutilisation**
- Ouverture au **dialogue** avec tous les acteurs, interdisciplinarité
- Engagement de et vers la **société**

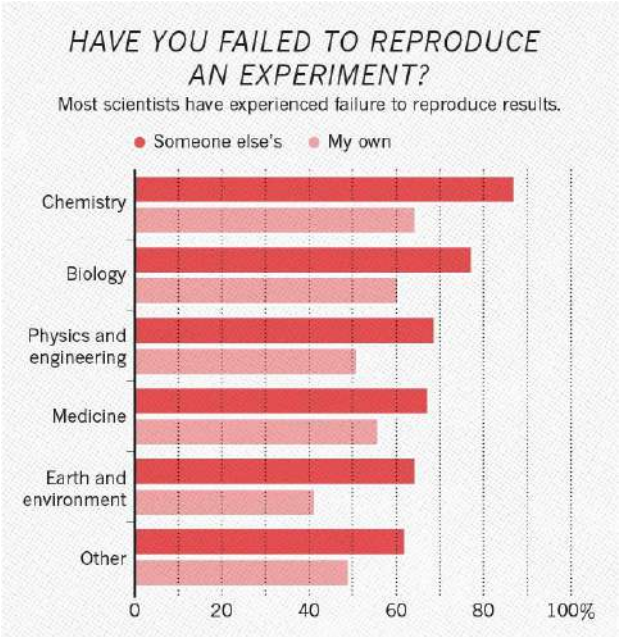
- **Démocratiser l'accès aux savoirs**
- **Rendre la science plus cumulative**, plus fortement étayée par les données, plus transparente
- **Augmenter l'efficacité de la recherche** en évitant de dupliquer les efforts, en ré-utilisant des données ou du matériel scientifique
- **Favoriser les avancées scientifiques** et l'**innovation**
- Favoriser la **confiance des citoyens dans la science**



# 70 %

des analyses en biologie expérimentale ne sont

**pas reproductibles**




Monya Baker, 2016

**Les données de la recherche sont des informations publiques :**

- Principe d'ouverture par défaut et de libre utilisation (Loi Lemaire - LPRN 2016)
- Principe de gratuité (Loi Valter 2015) :
  - Seule une liste fermée d'administrations peuvent fixer des redevances de réutilisation (IGN, Météo France)
  - Articulation possible avec le dépôt de brevets et d'autres formes de valorisation

**« aussi ouvert que possible,**
**aussi fermé que nécessaire »**

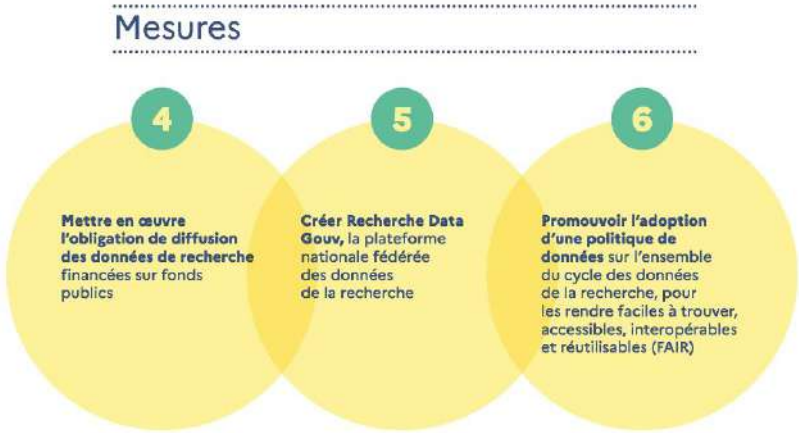
# Deux plans nationaux pour la Science Ouverte

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
*Liberté*
*Égalité*
*Fraternité*

Premier plan national pour la science ouverte : poser les principes de la science ouverte


2018-2021


2021-2024



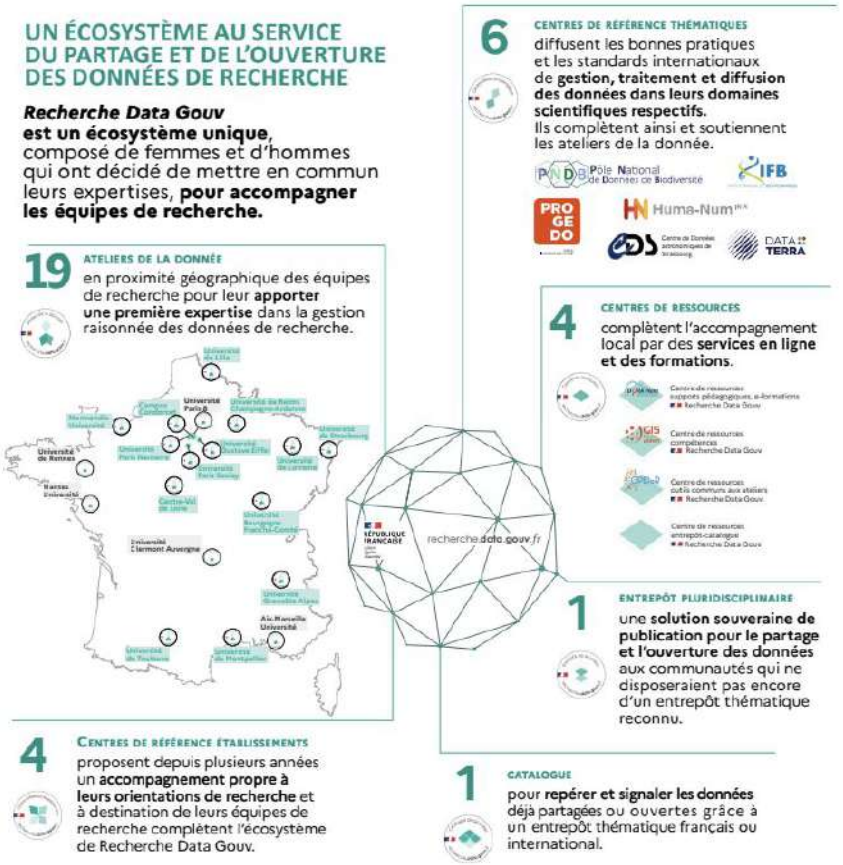

Deuxième plan national pour la science ouverte : périmètre étendu aux algorithmes et codes sources et déclinaisons thématiques

RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

recherche.data.gouv.fr

**Objectifs**

- Soutenir les équipes de recherche dans leur travail de structuration des données pour les rendre Faciles à trouver, Accessibles, Interopérables, Réutilisables (FAIR)
- Devenir un service de l'European open science cloud (EOSC), offrant un accès au patrimoine des données partagées et ouvertes de la recherche pour favoriser leur réutilisation




IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
elixir FRANCE

# La vie d’un projet

## Montage de projet

Paragraphe sur la gestion des données

Budgéter la science ouverte (APC, data manager, stockage, …)

## Début de projet

Mise en place d’un plan de gestion des données (PGD)

## Pendant le projet

Gestion des données suivant les principes FAIR

Mise à jour du PGD

## Fin du projet

Partage des données “aussi ouvert que possible, aussi fermé que nécessaire”

Entrepôts de données

Open Access des publications

# La vie d’un projet

## Montage de projet

Paragraphe sur la gestion des données

Budgéter la science ouverte (APC, data manager, stockage, …)

## Début de projet

Mise en place d’un plan de gestion des données (PGD)

## Pendant le projet

Gestion des données suivant les principes FAIR

Mise à jour du PGD

MADBOT
Metadata And Data Brokering Online Tool

## Fin du projet

Partage des données “aussi ouvert que possible, aussi fermé que nécessaire”

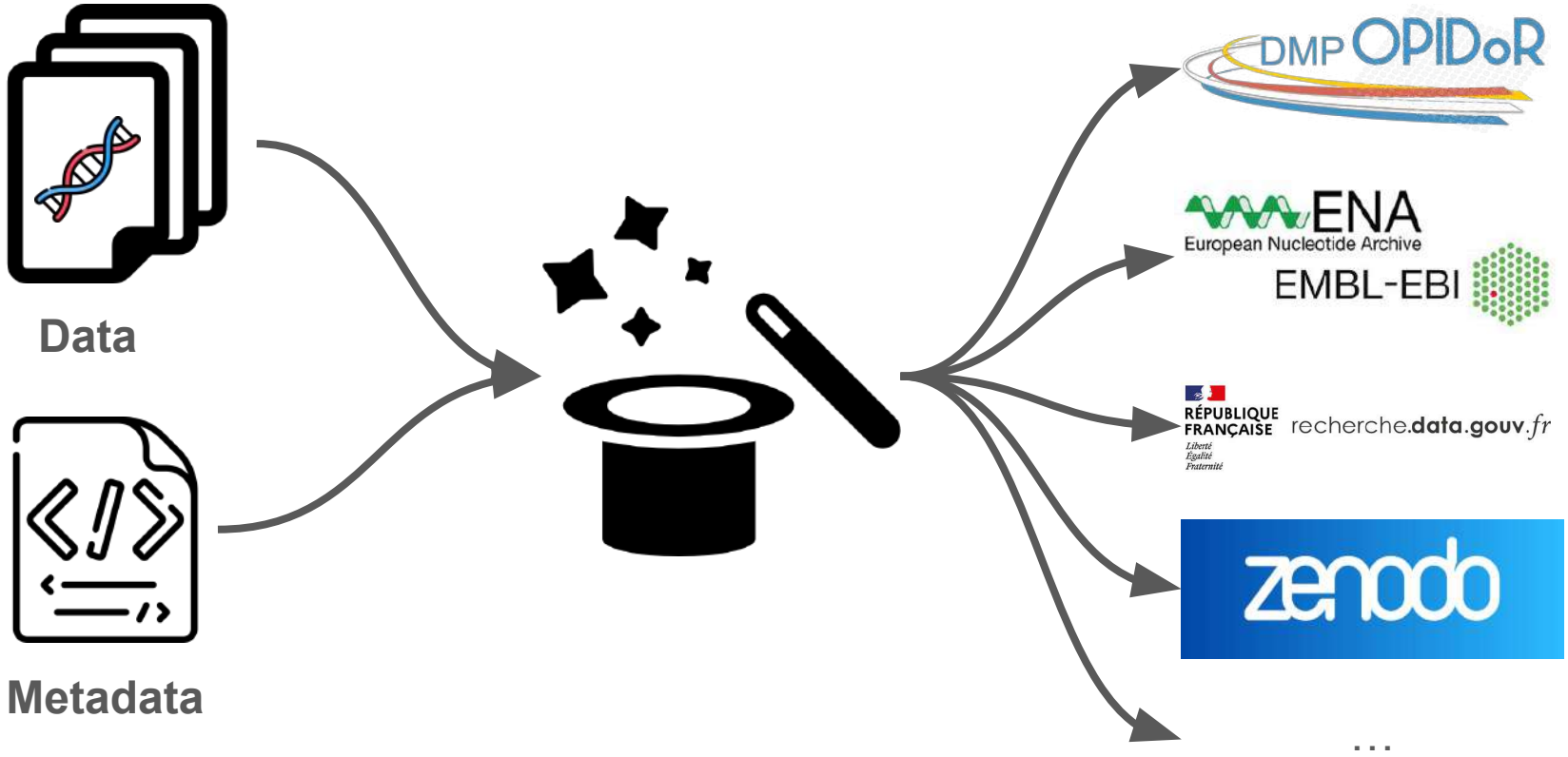
Entrepôts de données

Open Access des publications

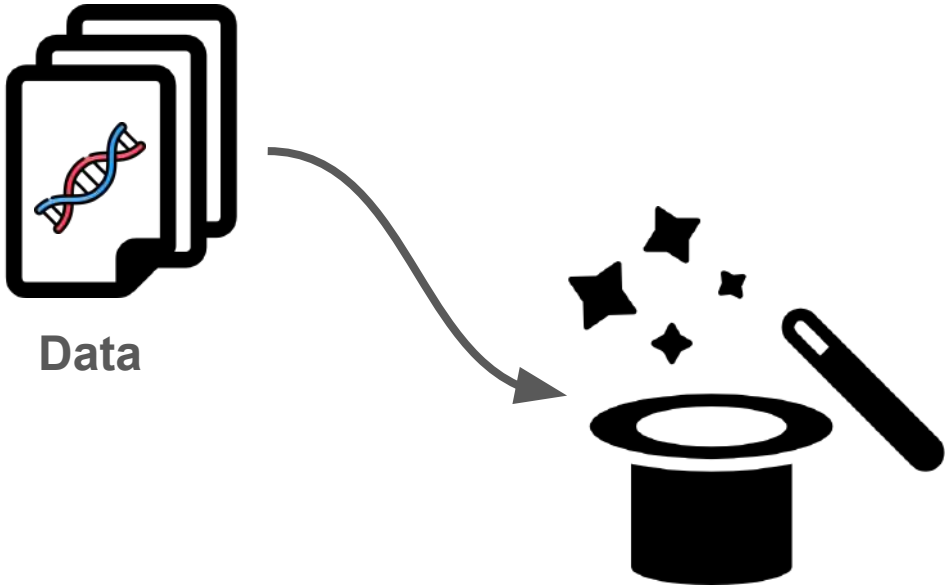
# Pourquoi Madbot ?

## Revenons aux sources

# Notre objectif



# Notre objectif



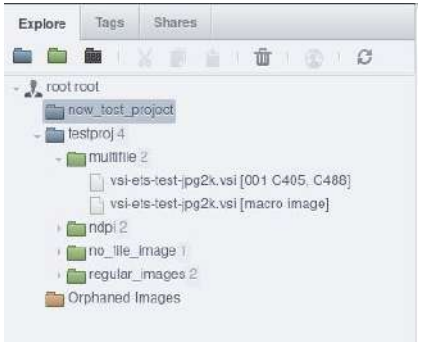
# Et les données ? N solutions de stockages, N organisations…

Galaxy

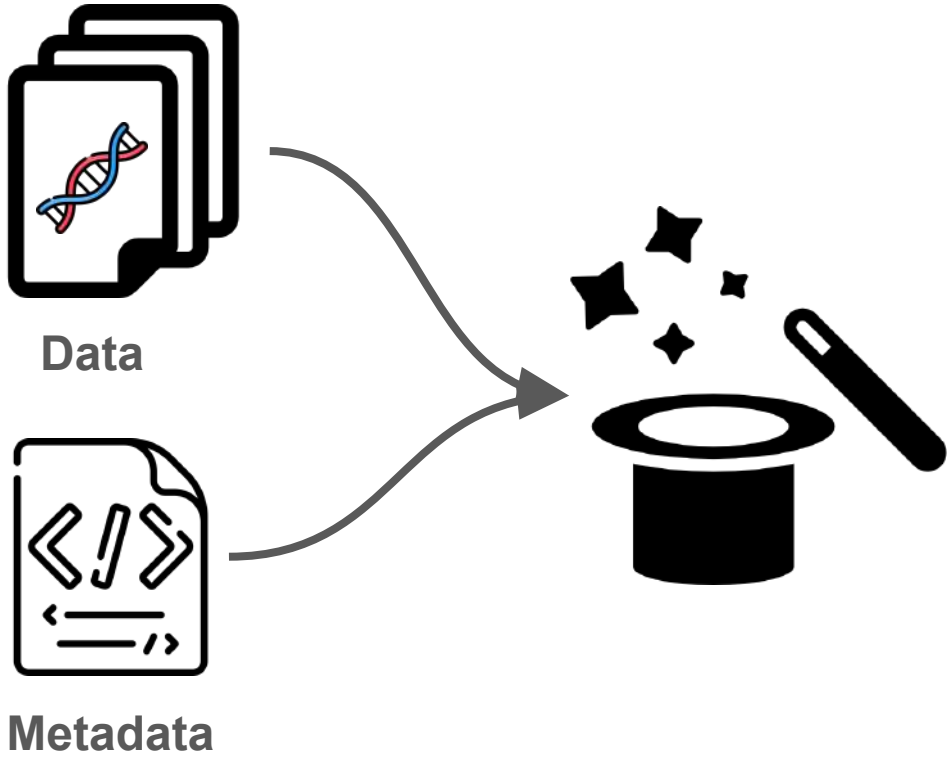
IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

OMERO





# Notre objectif



# Comment décrire au mieux les données? Avec un standard

In essence, a standard is an **agreed way of doing something**.

A standard provides the **requirements**, **specifications**, **guidelines** or **characteristics** that can be used for the **description**, **interoperability**, **citation**, **sharing**, **publication**, or **preservation** of all kinds of **digital objects** such as data, code, algorithms, workflows, software, or papers.

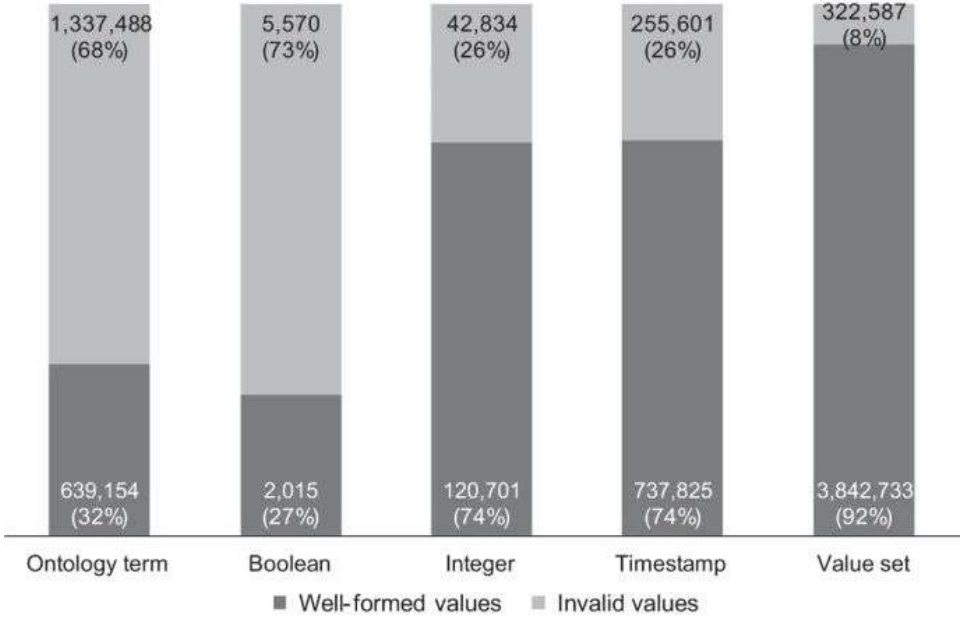
*source: https://fairsharing.org/educational/*

# Standards, mais...

La soumission dans les entrepôts publics est souvent une tâche complexe

Les procédures de soumission sont hétérogènes.

**Les métadonnées sont souvent incomplètes, incohérentes, redondantes ou insuffisamment informatives**.

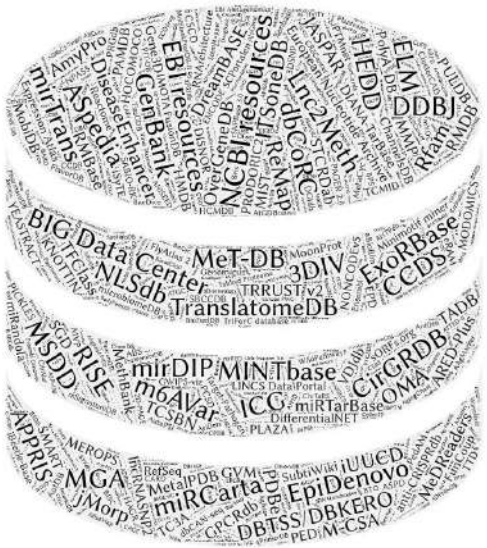



**Quality of dictionary attributes in NCBI BioSample according to their type, in** Gonçalves et al., 2019

# Standards, mais...

Il existe des milliers de bases de données, de logiciels et de ressources en biologie dont le **niveau d'adoption des standards est inégal**.

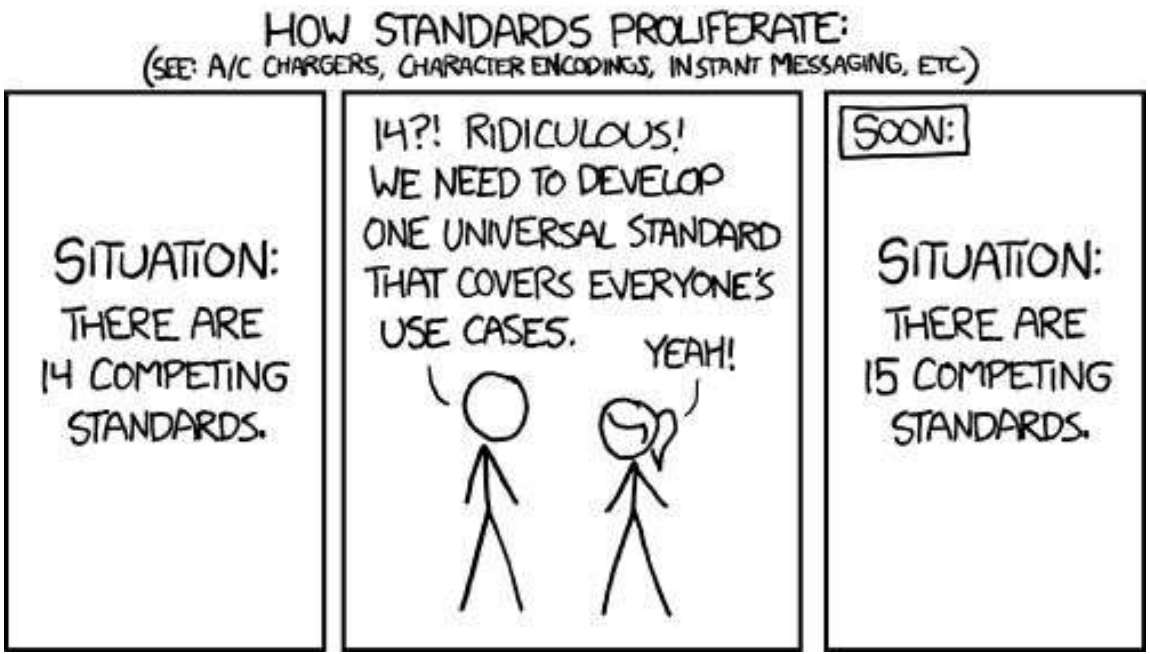
Il n'est pas toujours facile pour les biologistes et les bioinformaticiens d'identifier et d'utiliser les standards appropriés….




**1641** databases in NAR Database 2021

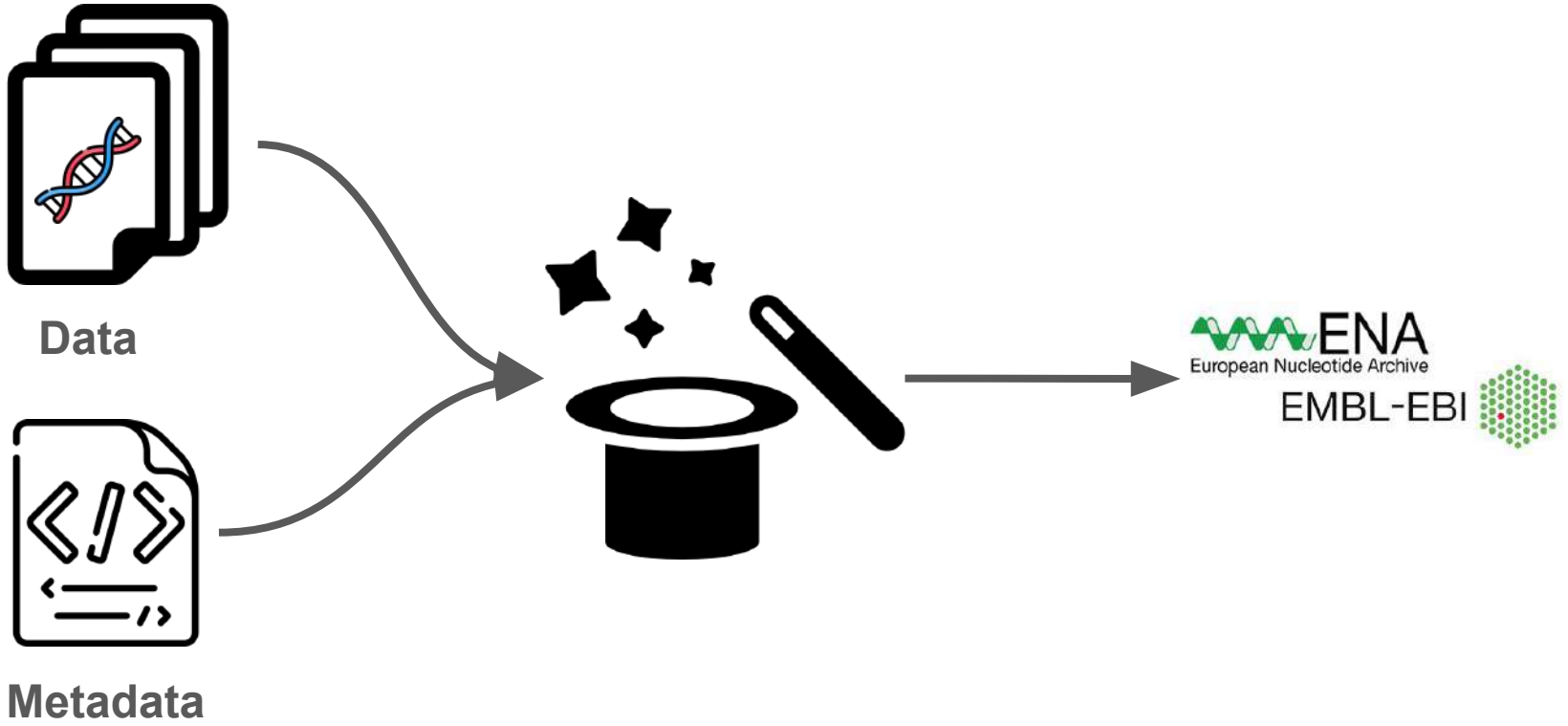
Rigden *et al*. 2021

# Standards, mais...



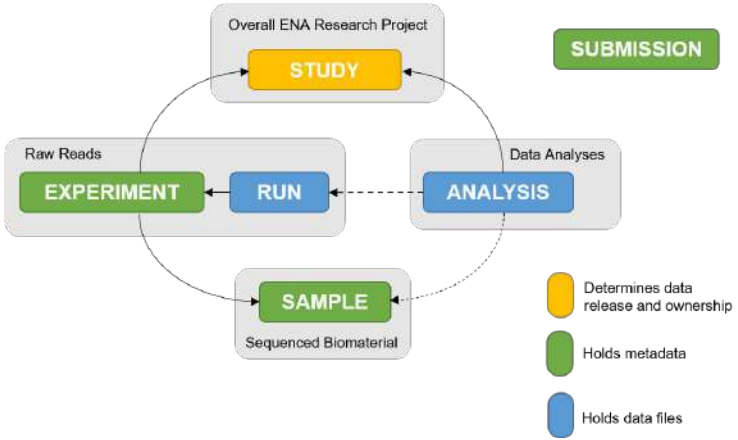

Source: https://xkcd.com/927/

# Notre objectif



# J’ai mes données, mes métadonnées, j’ai fait le plus gros (😂)

La European Nucleotide Archive (ENA) est une plateforme ouverte pour la gestion, le partage, l'intégration, l'archivage et la diffusion des données de séquence.




| | Interactive | Webin-CLI | Programmatic |
|---|---|---|---|
| Study | Y | N | Y |
| Sample | Y | N | Y |
| Read data | Y | Y | Y |
| Genome Assembly | N | Y | N |
| Transcriptome Assembly | N | Y | N |
| Template Sequence | N | Y | N |
| Other Analyses | N | N | Y |

```
<PROJECT_SET>
   <PROJECT alias="iranensis_wgs">
      <NAME>WGS Streptomyces iranensis</NAME>
      <TITLE>Whole-genome sequencing of Streptomyces iranensis</TITLE>
      <DESCRIPTION>The genome sequence of Streptomyces iranensis (DSM41954) was obtained using
      <SUBMISSION_PROJECT>
         <SEQUENCING_PROJECT/>
      </SUBMISSION_PROJECT>
      <PROJECT_LINKS>
         <PROJECT_LINK>
            <XREF_LINK>
               <DB>PUBMED</DB>
               <ID>25035323</ID>
            </XREF_LINK>
         </PROJECT_LINK>
      </PROJECT_LINKS>
   </PROJECT>
</PROJECT_SET>
```

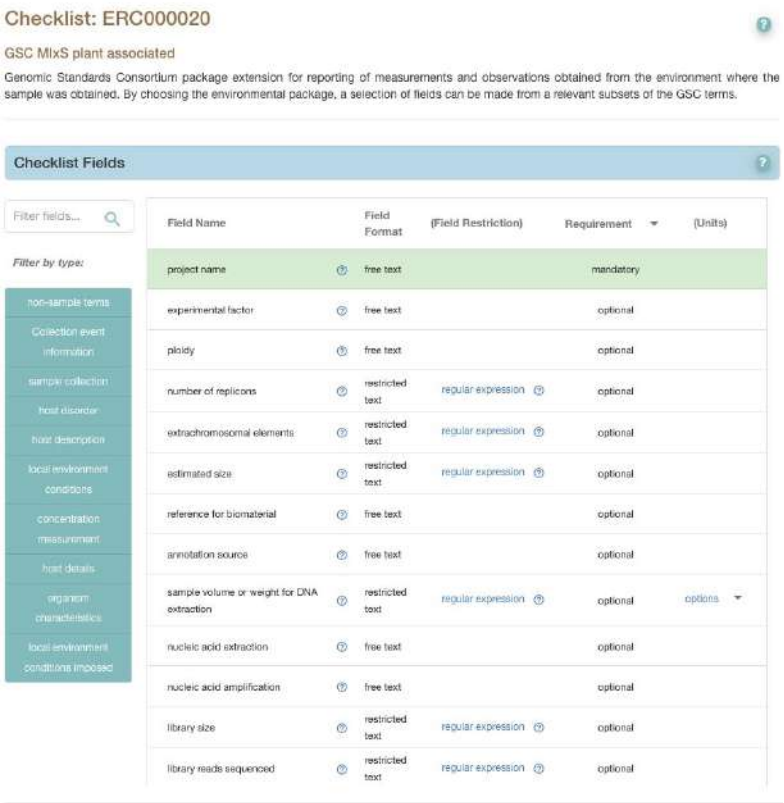
**Source**: https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html

# Les checklists pour les métadonnées

Une **checklist** définit les **métadonnées minimales et optionnelles** attendues pour décrire un échantillon biologique

Elles sont basées sur les recommandations du **Genomic Standards Consortium (GSC)**

La **checklist la plus appropriée** dépend du type d'échantillon : https://www.ebi.ac.uk/ena/browser/checklists



# Pour faciliter les soumissions, il existe des solutions




# Bilan

## C’est pas si simple…

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

# Idée !

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


MADBOT
Metadata And Data Brokering Online Tool
IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

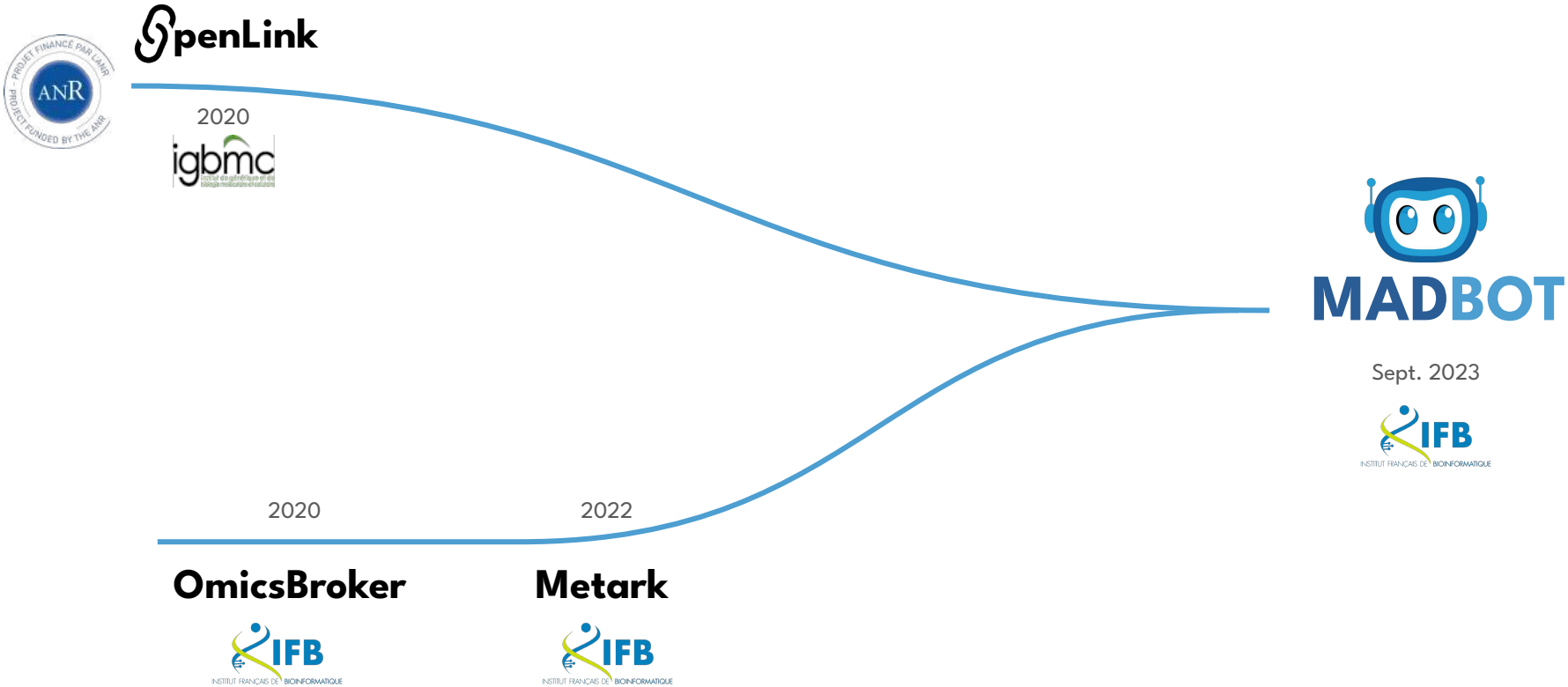
# Qu’est-ce que MADBOT ?

**MADBOT** (Metadata And Data Brokering Online Tool) est une application web qui fournit **un tableau de bord pour la gestion des données et des métadonnées de recherche**.

L’application permet d’**agréger les métadonnées** autour de projets scientifiques et d’**identifier précisément où sont stockées les données** au travers de nombreux connecteurs.

Elle accompagne également les chercheurs pour la **publication des données et des métadonnées** vers un grand nombre d’entrepôts généraliste (RGD, Zenodo, etc.) et thématique (ENA, GEO, etc.)

# Un jeune projet sur des bases solides



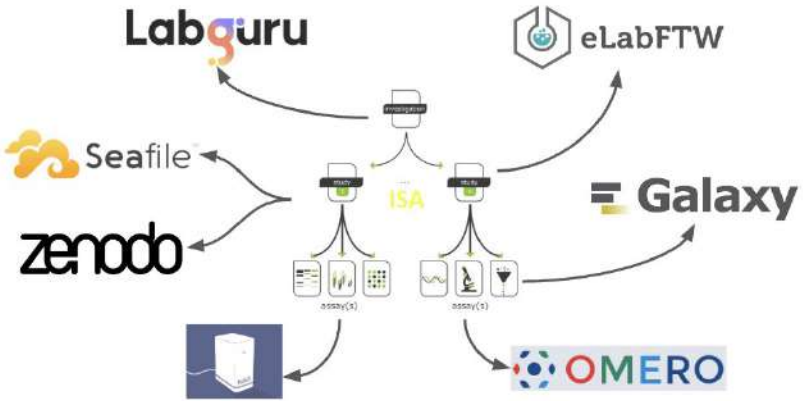
# OpenLink

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

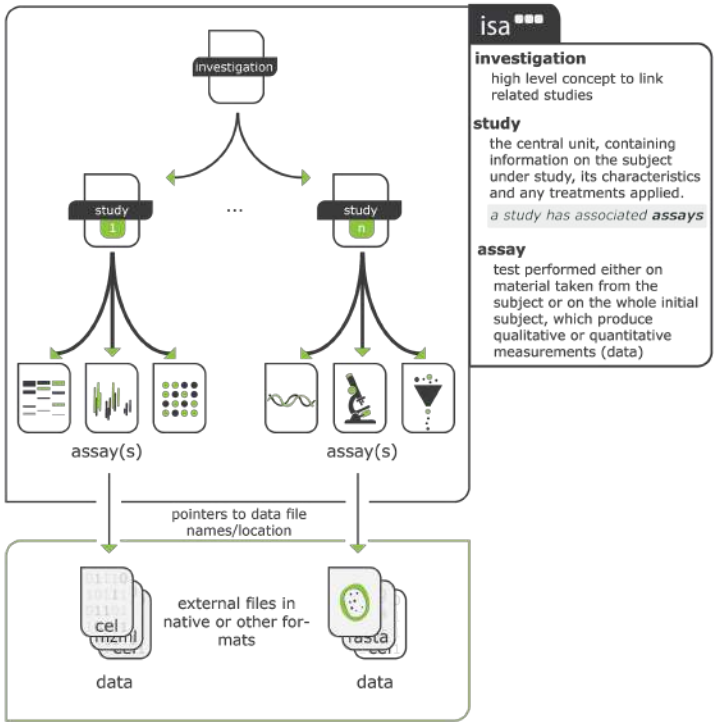


# Le modèle ISA

Un standard pour les données des sciences de la vie

Un modèle pour représenter les métadonnées expérimentales à travers 3 entités de base :

- **I**nvestigation : le contexte du projet
- **S**tudy : une expérimentation en un seul lieu
- **A**ssay : une mesure spécifique qui cible un trait avec une méthode et une échelle.




Sources: https://isa-tools.org and : https://isa-specs.readthedocs.io/en/latest/isamodel.html

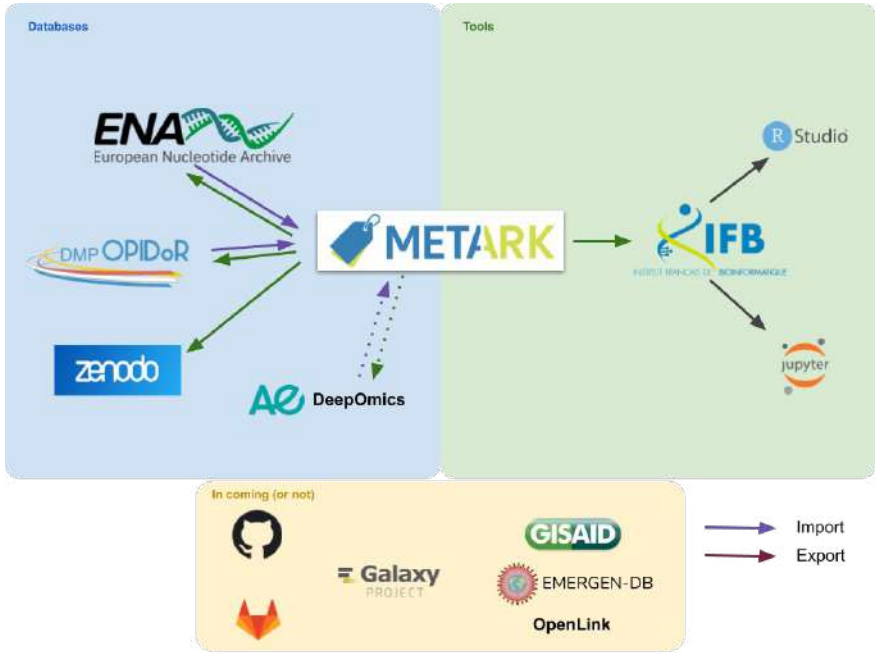
# METARK

Un projet qui a débuté juste avant la collaboration IFB avec SpF pour la COVID-19

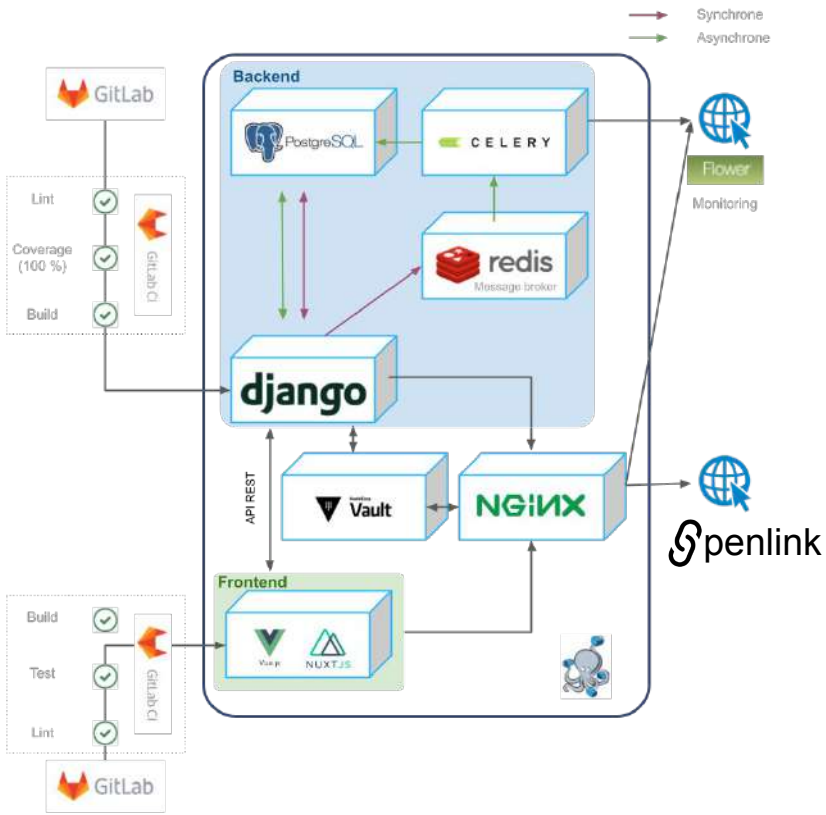
Metark un outil de data brokering pour faciliter la soumission des données et des métadonnées  des sciences de la Vie dans des entrepôts internationaux

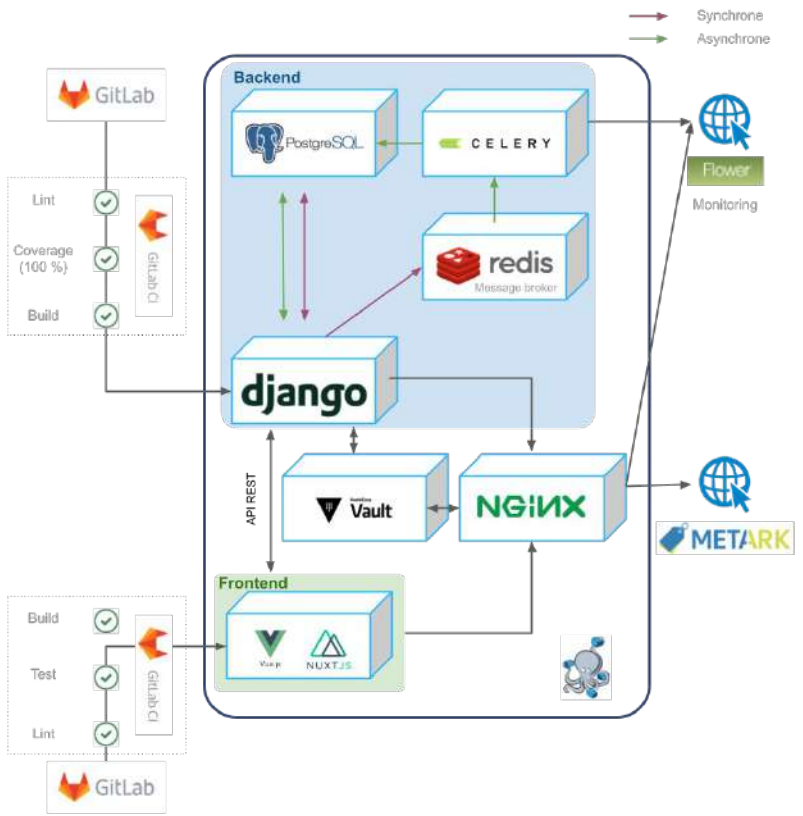
Sur une demande de l'ENA pour que l'IFB deviennent le data broker national.

Basé aussi sur un modèle ISA



# Une architecture commune



# Fusion !

OpenLink

Spécialisé dans la gestion des données et qui souhaitait intégrer les métadonnées

METARK

Spécialisé dans la gestion des métadonnées et qui ne savait pour où étaient les fichiers de données à soumettre


Vas-y mollo je crois que les biologistes les aiment bien


Ok, je commence par le tout vert, ça va calmer tous ceux en collant

# Une nouvelle équipe de développeurs

**Thomas DENECKER**
Full stack developer & Scrum master

**Julien SEILER**
Full stack developer & scrum master

**Laurent Bouri**
Full stack developer

**Imane MESSAK**
Full stack developer

**Baptiste ROUSSEAU**
Full stack developer

- **Description des projets de recherche** au travers d'une structure arborescente, dont le modèle ISA (Investigation, Study, Assay)
- **Identification des données** associées aux projets
- Gestion et suivi des **métadonnées**
- Gestion et suivi des **échantillons biologiques**
- **Tableau de bord** de gestion des données
- **Publication** des données

Au sein de **MADBOT**, chaque projet de recherche est décrit sous la forme d'un ensemble de nodes qui peuvent être des investigations, des études (study) et d'expériences (assay) sous la forme d'un arbre hiérarchique (modèle ISA).

Mais nous ne nous limitons pas qu'au modèle ISA et nous prévoyons de mettre en place de nouvelles collections de nœuds très prochainement comme projets, workspaces, etc.

Chaque élément de cet arbre peut être associés à des données, des échantillons biologiques (qui sont également des nodes) et des métadonnées.

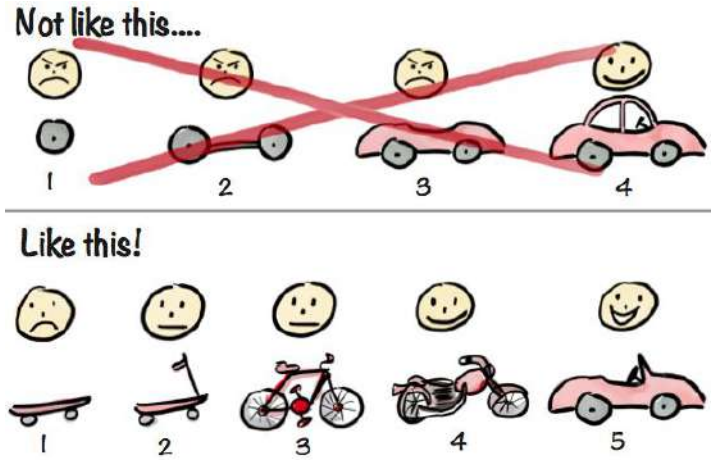
# Notre feuille de route

Nous souhaitons produire le plus rapidement possible une application fonctionnelle.

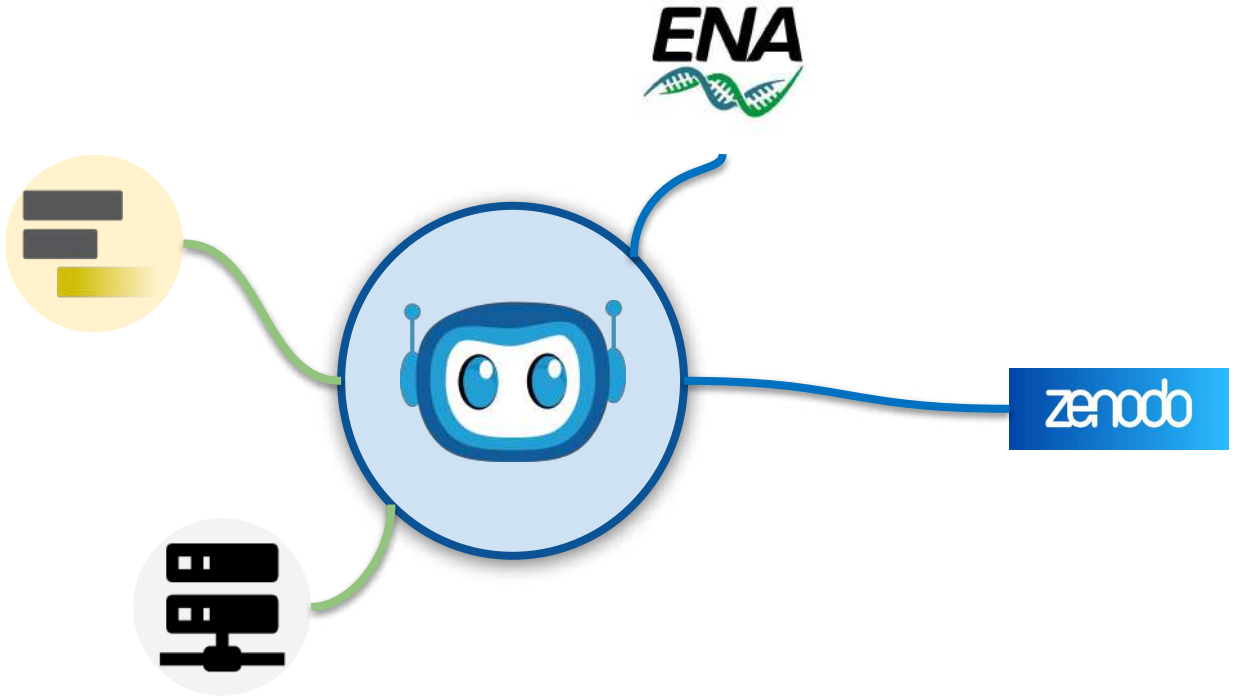
**Objectifs principaux**

- Démontrer que nous pouvons associer à un arbre ISA des données localisées sur différents systèmes de stockage (Galaxy et SSHFS)
- Démontrer que nous pouvons renseigner des métadonnées pour un arbre ISA et ses données
- Démontrer que nous pouvons publier ces données sur des entrepôts de références thématiques (ENA) et non thématiques (Zenodo).

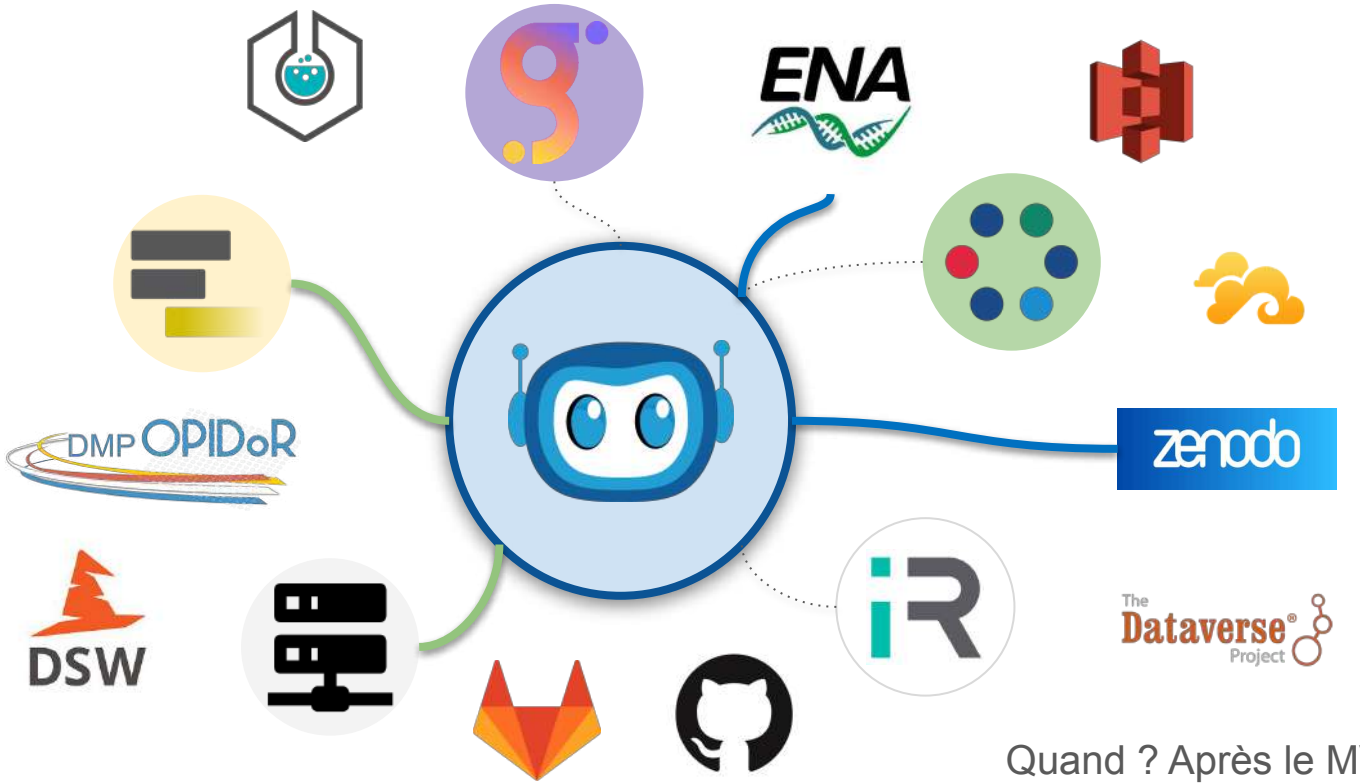
L'application Madbot capable de réaliser ces objectifs est notre **M**inimal **V**iable **P**roduct (MVP)



# Les connecteurs de Madbot pour le MVP



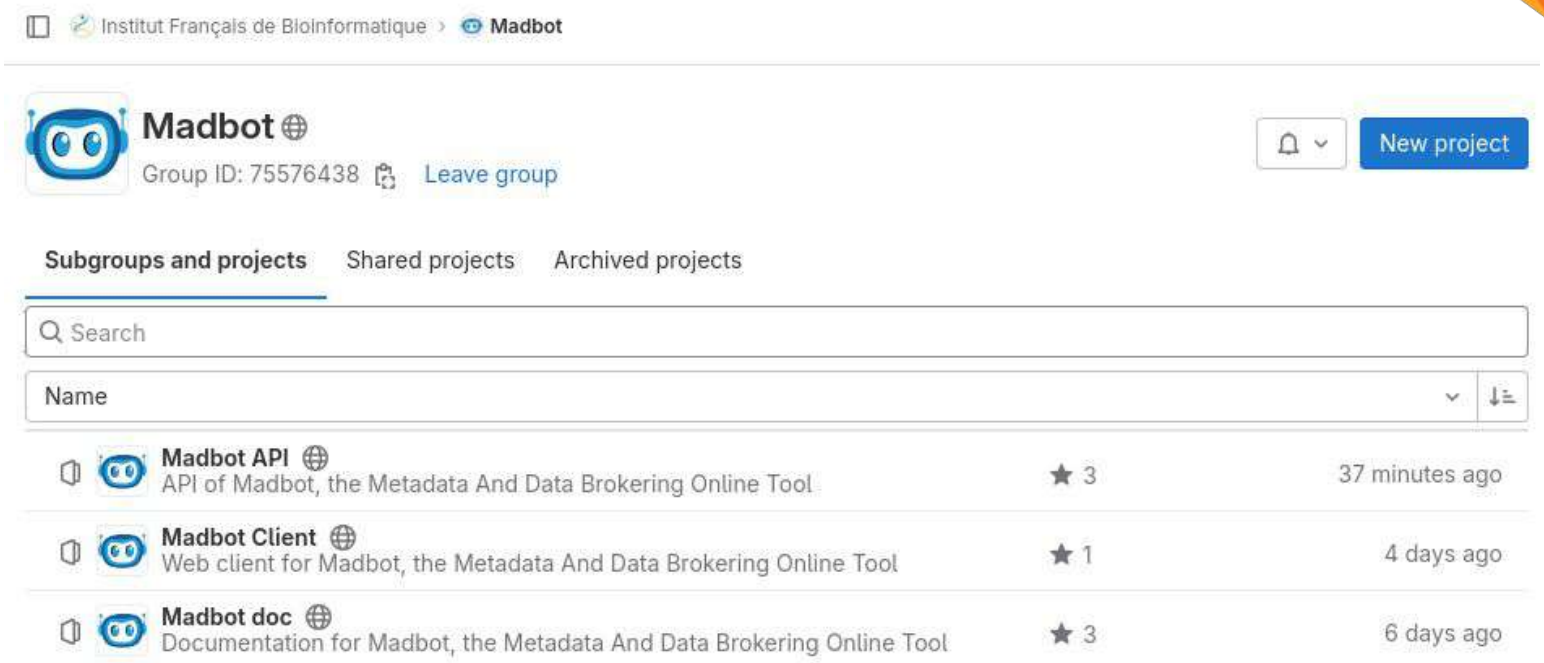
# Les connecteurs pour le futur de Madbot



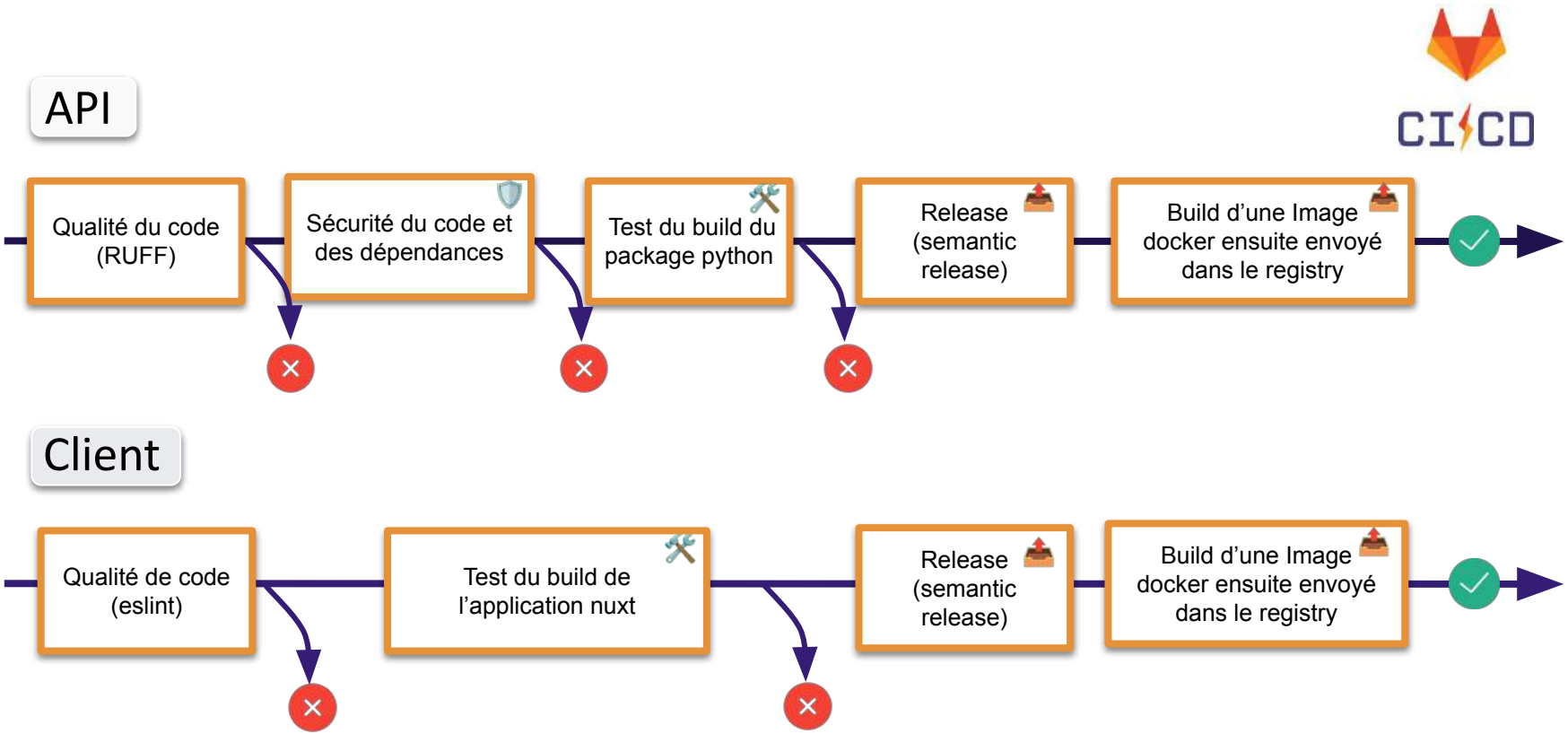

Quand ? Après le MVP ou par contribution

# Un projet Open Source sous licence

Le **code source** est hébergé sur **Gitlab** : https://gitlab.com/ifb-elixirfr/madbot

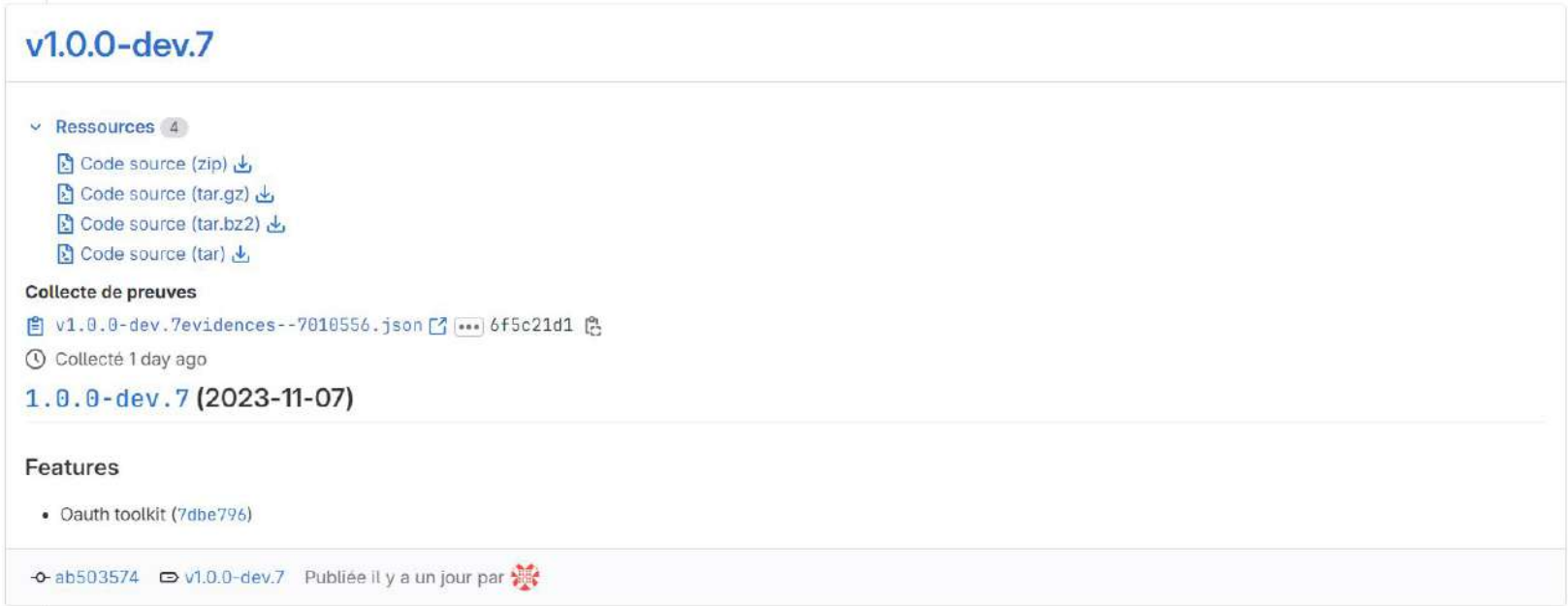


# Le code est testé et contrôlé automatiquement

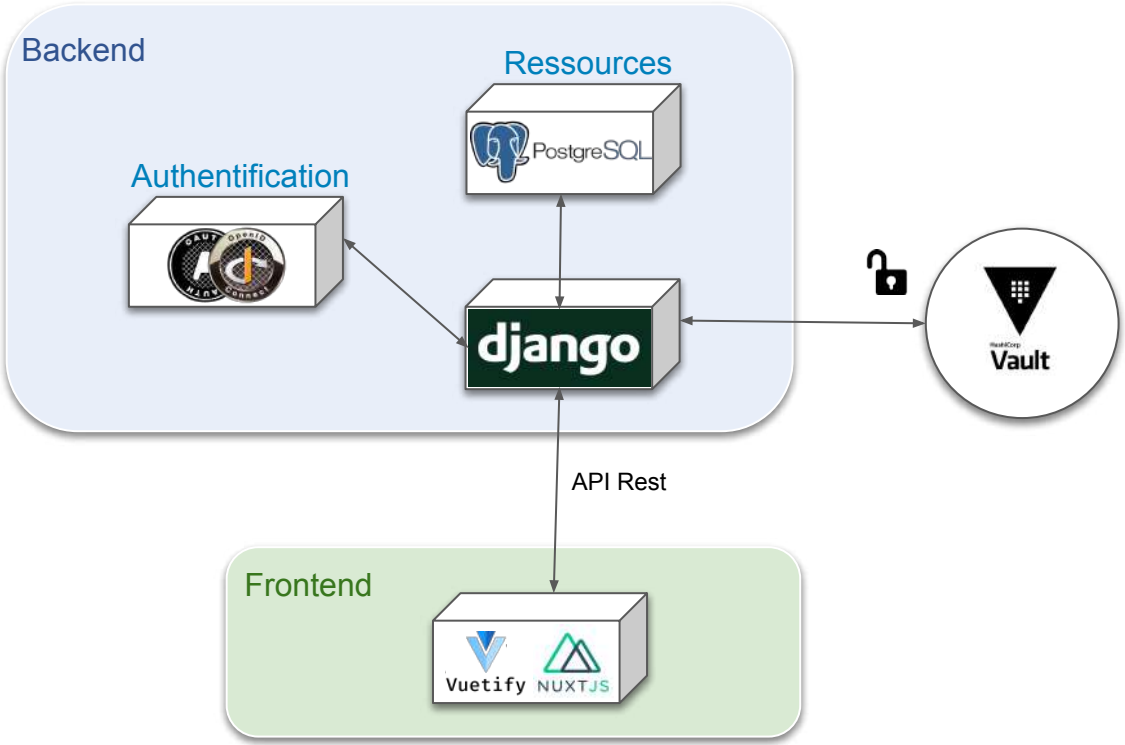


# Le code est releasé automatiquement

**Semantic release**:
Gestion des **versions et publication des packages** entièrement automatisées



# Architecture de Madbot




*Une couche d'asynchronie va bientôt être ajoutée.*

# 2 éléments clés dans cette architecture

Authentification

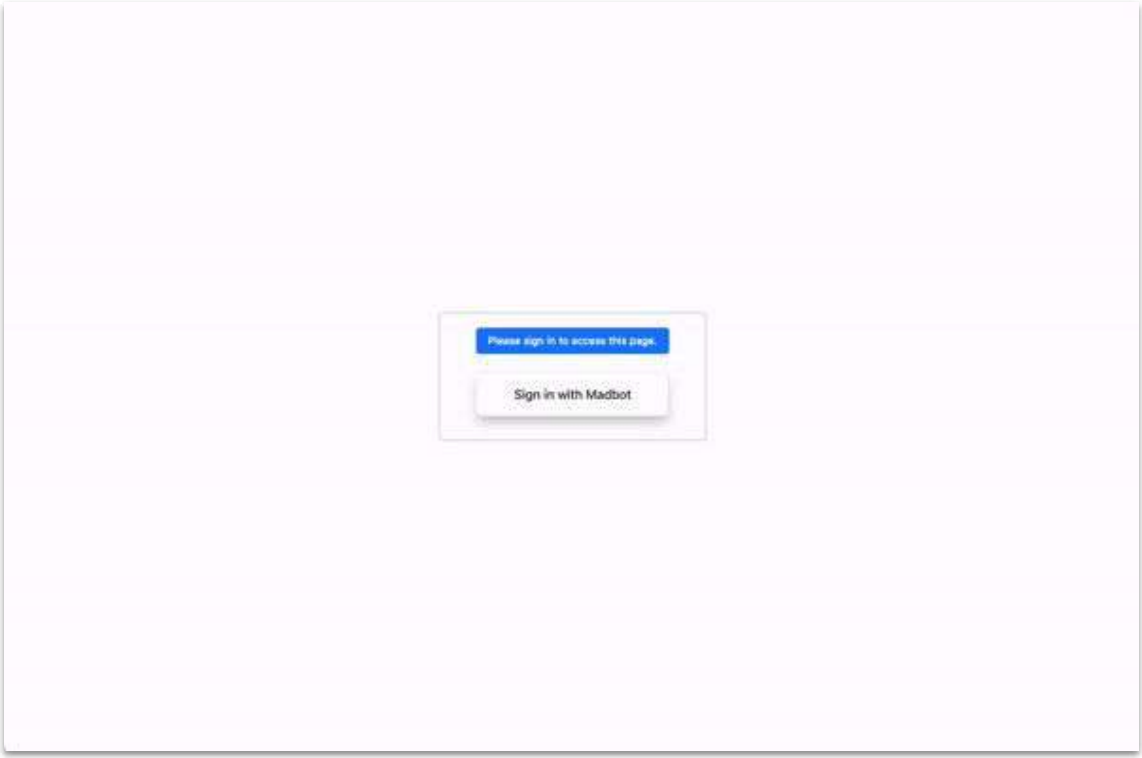
Madbot propose un système d'authentification oauth 2 et une possibilité d'interfacer n'importe quels identity provider.


## Outil de gestion et de protection des identifiants

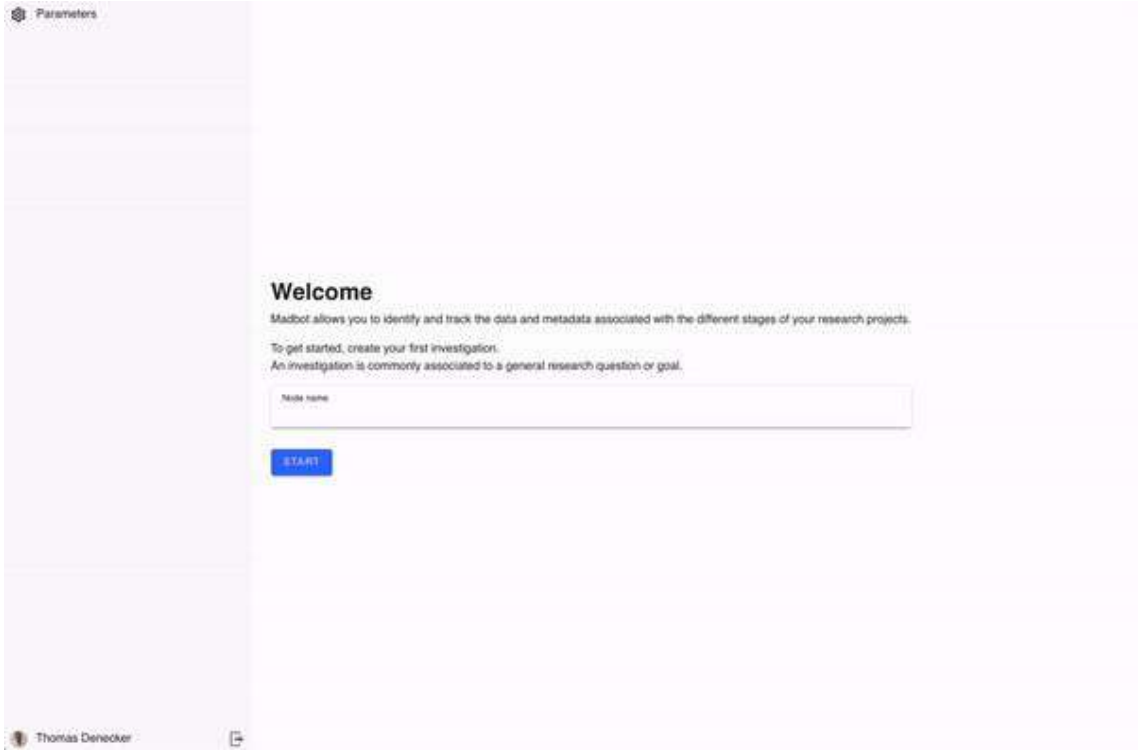
**Pourquoi stocker les identifiants ?**

- Éviter de demander de se connecter sans arrêt à des outils extérieurs
- Permettre de réaliser des tâches asynchrones (publication, mise à jour des métadonnées)
- Permettre de proposer un service de data brokering

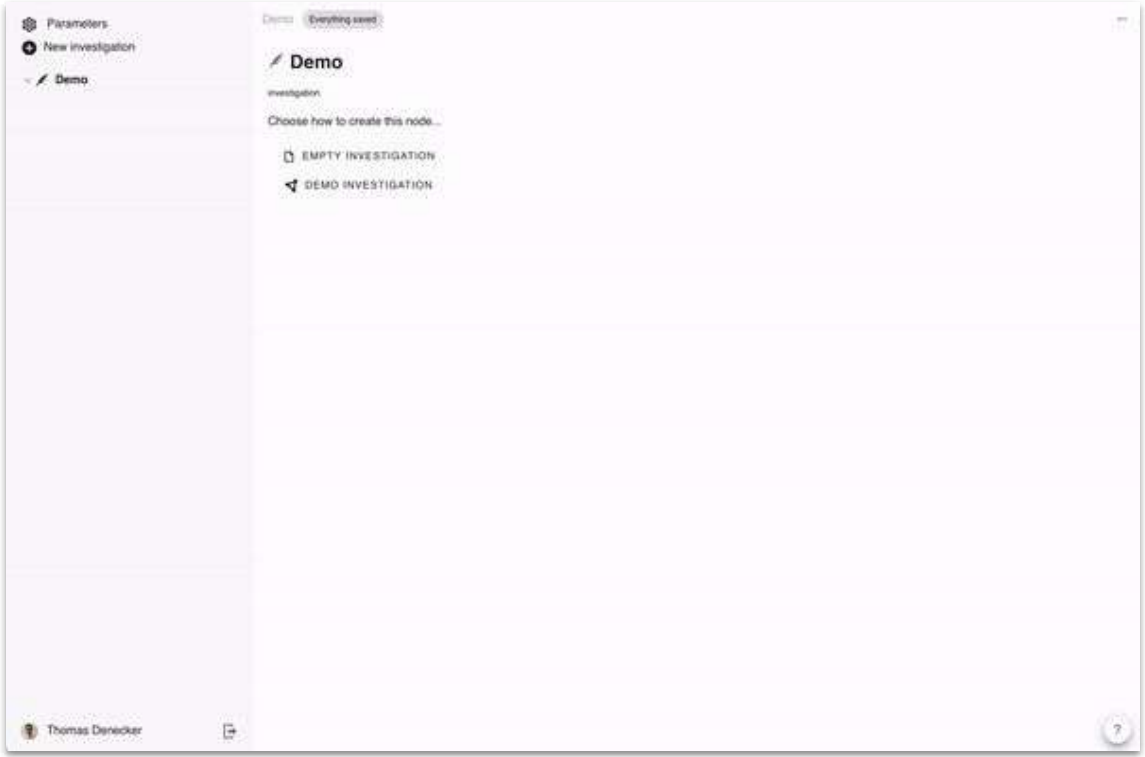
1. Connection



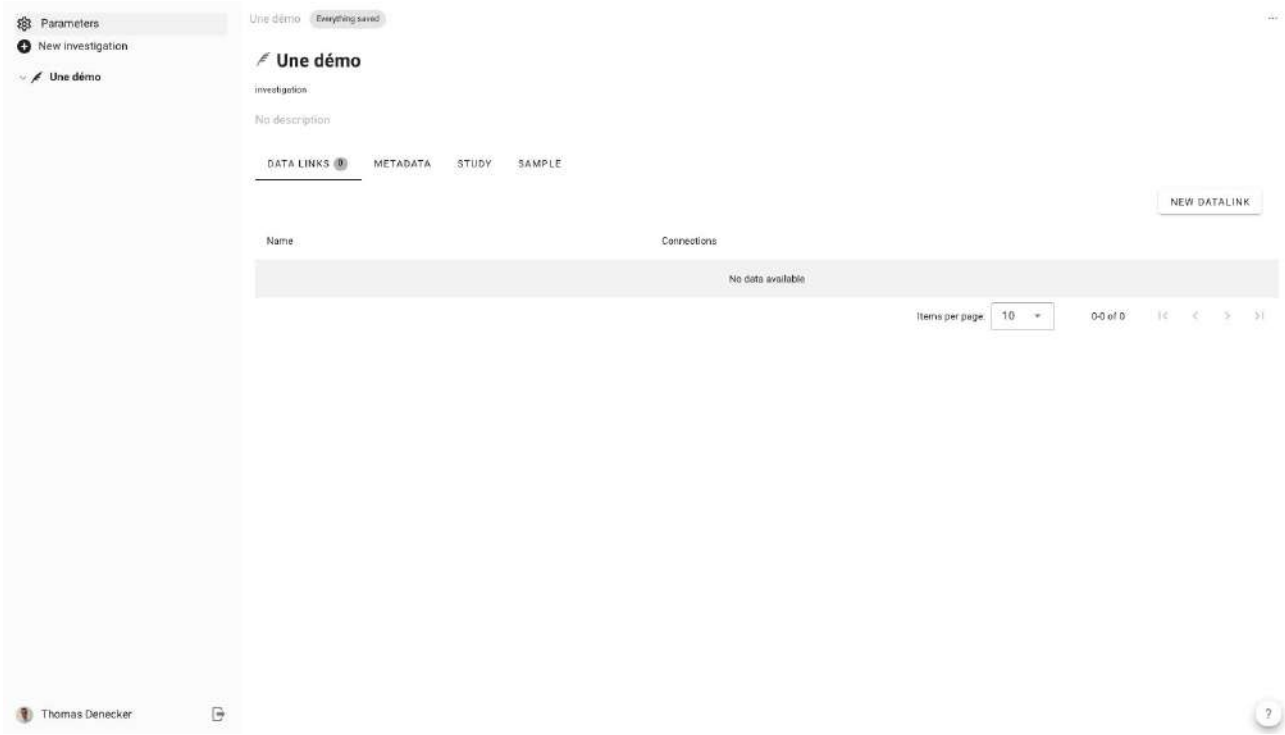
## 2. Création d'un premier node



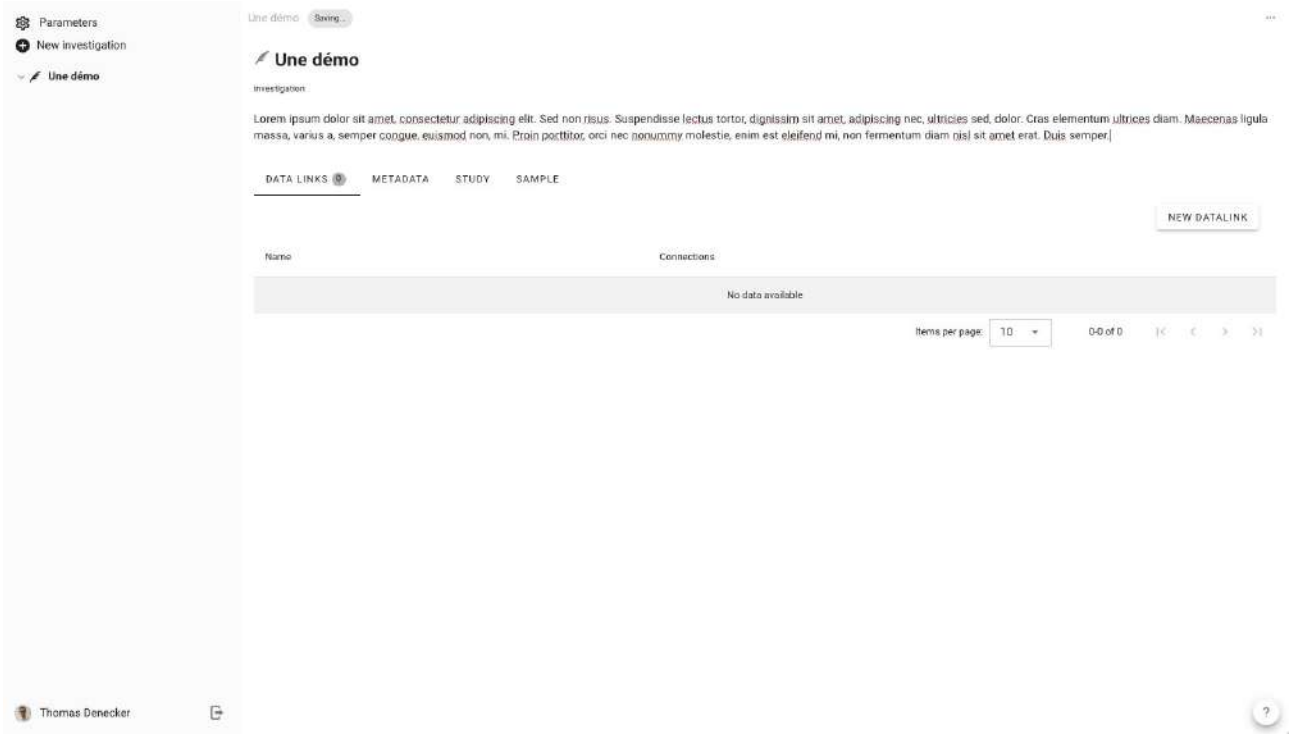
## 2. Choisir comment créer le node



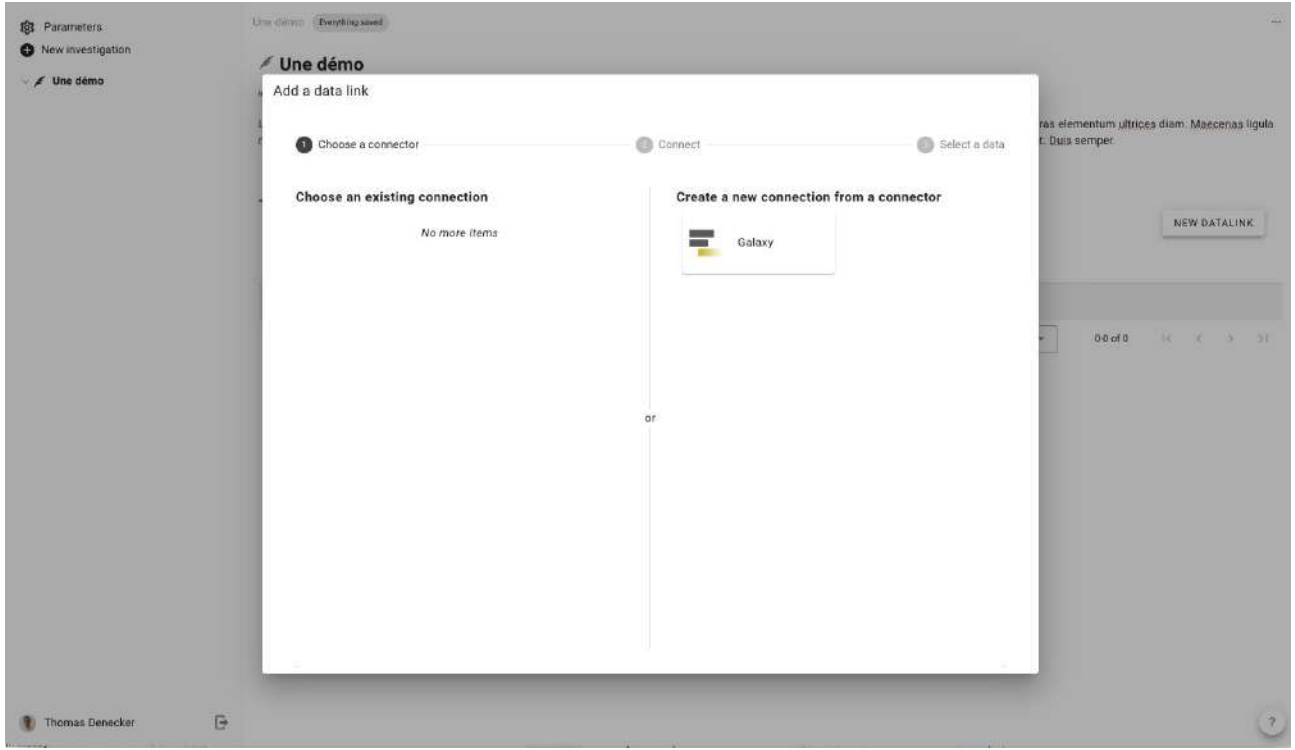
## 3. Page du node



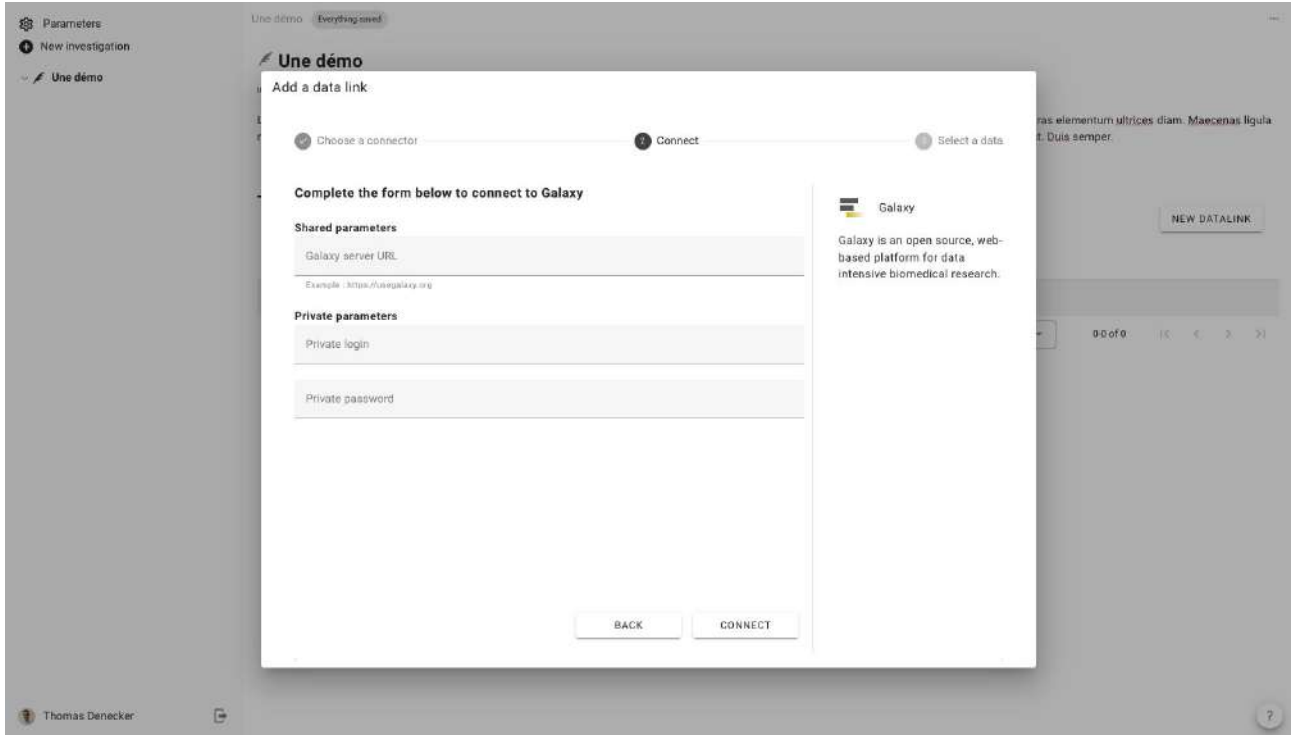
## 4. Ajout d'une description



## 4. Création d'un datalink

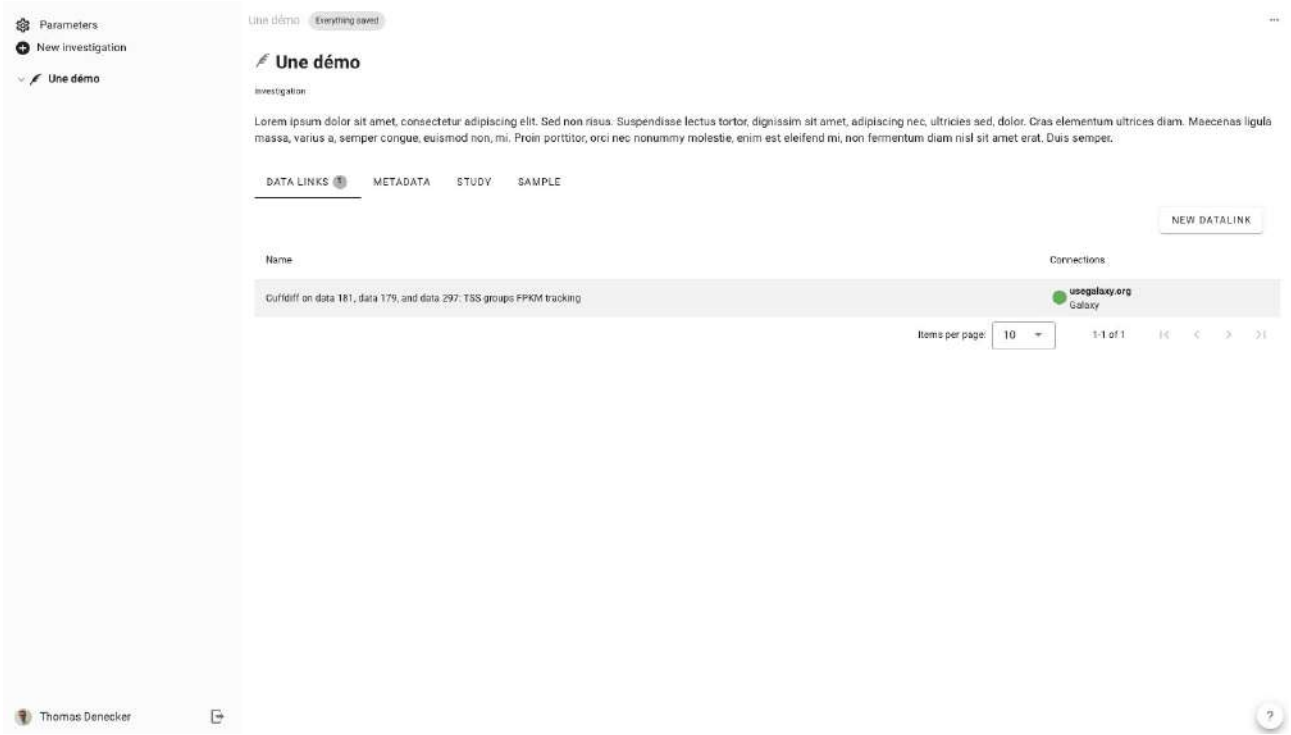


# 4. Création d'un datalink



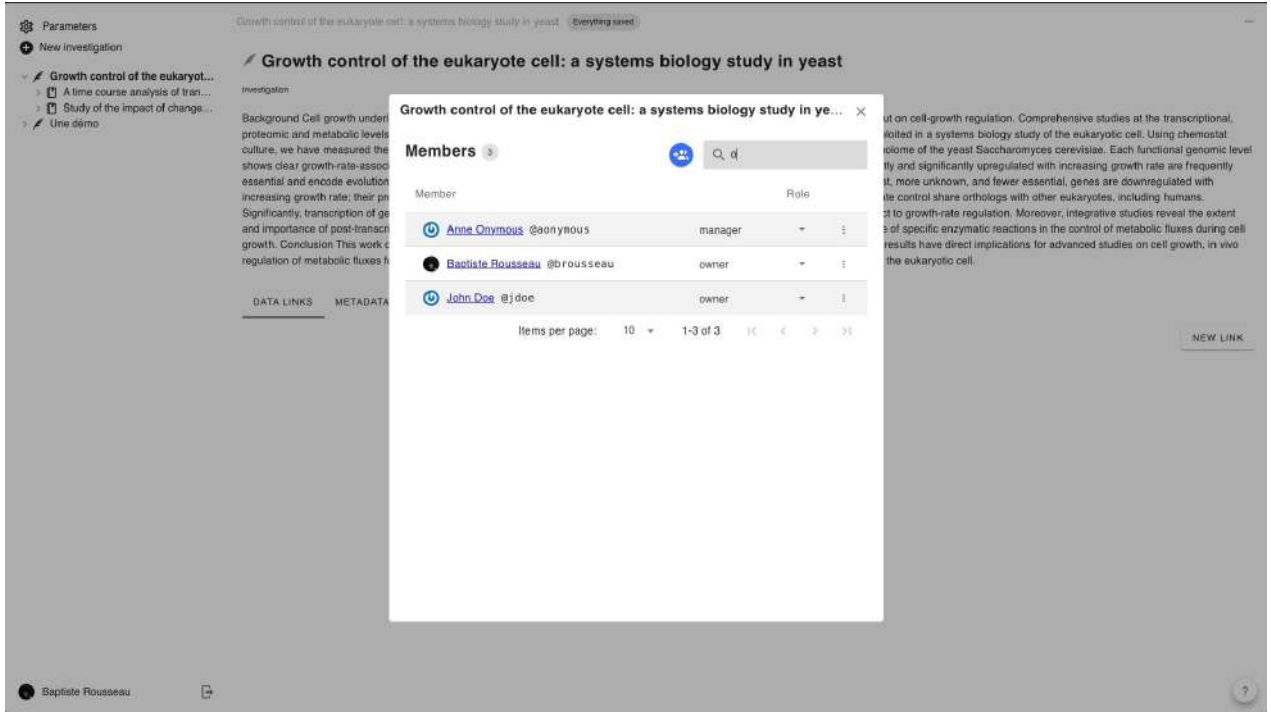
# 4. Création d'un datalink

## 5. Page du node



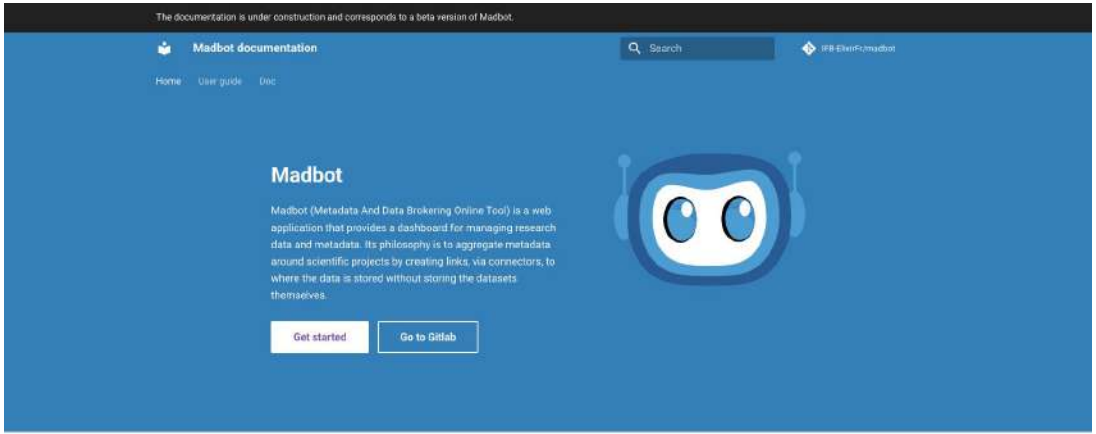
## 5Bis. Page du node pour un projet plus riche

## 6. Gestion des membres



# Nous avons une doc !

Pas encore très complète mais elle est en cours de rédaction




## Our philosophy ? Connectors !

Its philosophy is to aggregate metadata around scientific projects by creating links, via connectors, to where the data is stored without storing the datasets themselves.

These connectors make Madbot highly flexible and adaptable to a wide range of research environments. It already interfaces with many data management and analysis tools used daily by researchers, such as electronic laboratory notebooks (Labguru, eLabFTW), file storage systems (NAS, Seafile), research data warehouses (Zenodo, very soon ENA), web-based analysis platforms (Omero, Galaxy) and cluster infrastructures accessible via SSH. The application's architecture is based on the ISA (Investigation, Study, Assay) standard, enabling connectors to be created with virtually all the tools.

Madbot offers a cross-disciplinary view of the data and metadata for each research project. It enables researchers and their teams to gradually adopt a virtuous approach to managing their data, while helping them to achieve FAIR publication.

All the developments and results of this project will be made available to the entire scientific community on the GitLab collaborative platform under an open licence.

# Et après ?

# La suite pour le MVP

- Créer une connexion à un connecteur SSHFS
- Créer un lien secondaire entre un datalink et un Sample
- Importer une collection de métadonnées (checklist) dans un node
- Saisir la valeur d'une métadonnée pour des types simples (Text, Choice list)
- Sélectionner un ensemble de datalinks pour démarrer une publication
- Créer une publication au travers d'un connecteur Zenodo
- Suivre l'état d'une tâche de publication (tâche asynchrone)
- Consulter les identifiants de publication associés à un datalink
- Créer une publication au travers d'un connecteur ENA

# Les perspectives après le MVP ?

- Bancher d'autres **systèmes d'authentification**
  - LifeScience AAI (Elixir)
  - DOI
  - GitLab (déjà testé)
  - …
- Ajout des **nouveaux connecteurs**
  - Recherche data gouv
  - PGD (opidor et DSW)
  - iRODs
  - …
- Proposer de la **formation** pour le déploiement et l'utilisation de Madbot
- Proposer un **hackathon** pour le développement de nouveaux connecteurs

# Merci !