



Madbot

Thomas Denecker, Julien Seiler

► **To cite this version:**

Thomas Denecker, Julien Seiler. Madbot. 17ème Atelier de Données bioinformatiques de diversité, Dimitry Khvorostyanov, Mar 2024, Roscoff, France. hal-04502839

HAL Id: hal-04502839

<https://hal.science/hal-04502839>

Submitted on 13 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

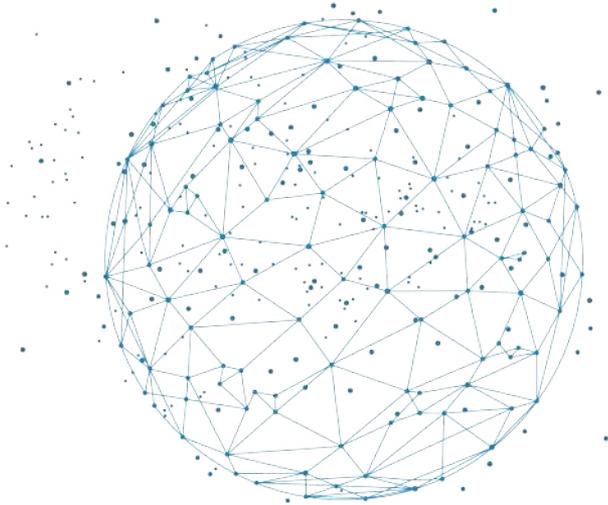
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License



Roscoff- 12 mars 2024



Madbot

Une passerelle vers la science ouverte

Julien Seiler

IGBMC & IFB - <https://orcid.org/0000-0002-4549-5188>

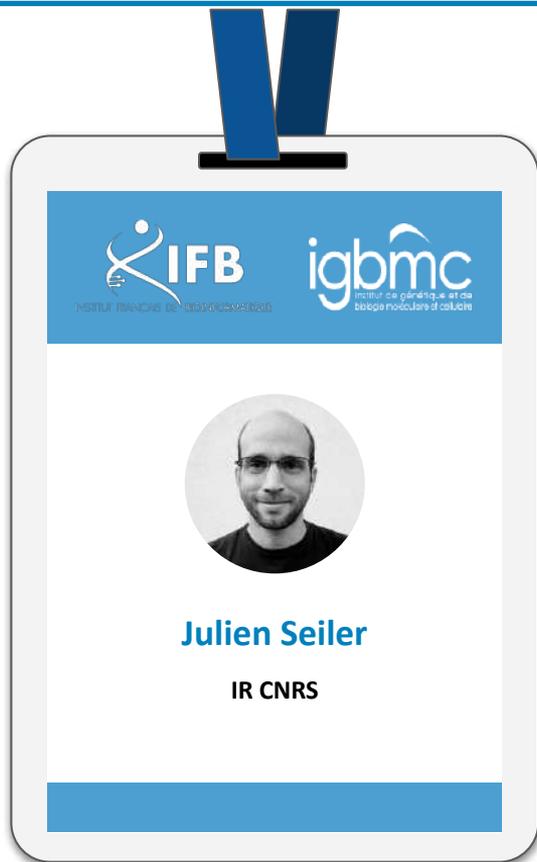
T. Denecker

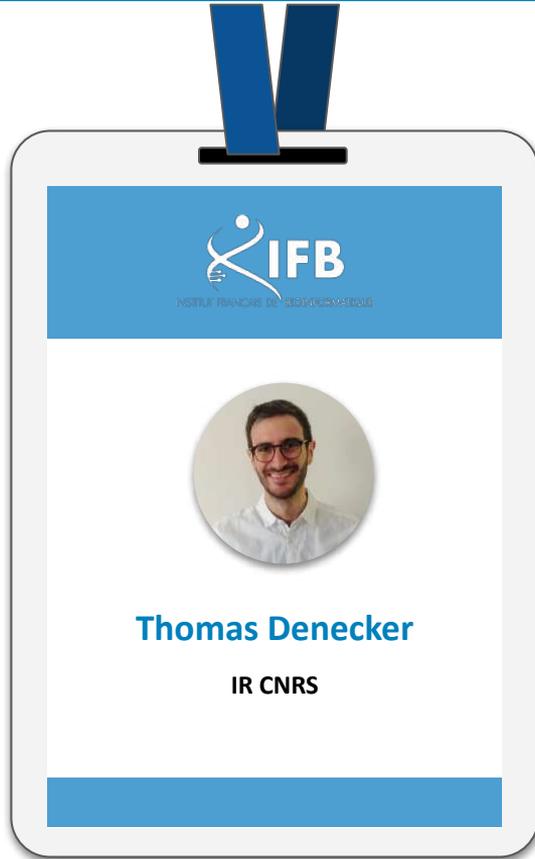
IFB - <https://orcid.org/0000-0003-1421-7641>





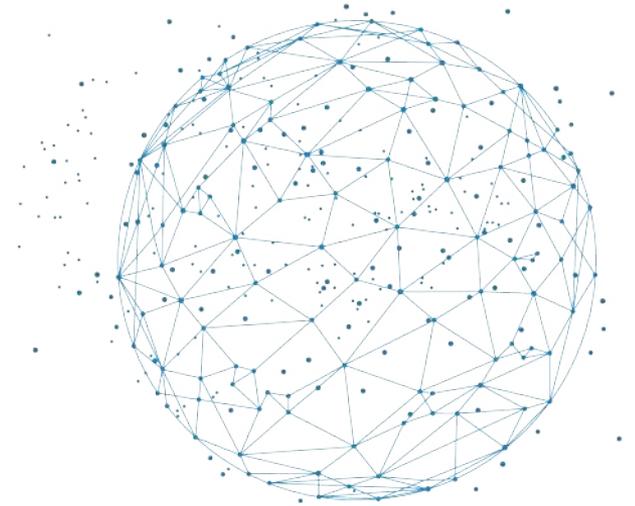
- Contexte de la science ouverte
- Pourquoi nous en sommes arrivés à développer madbot ?
- Qu'est-ce que c'est madbot ?
- La suite du projet

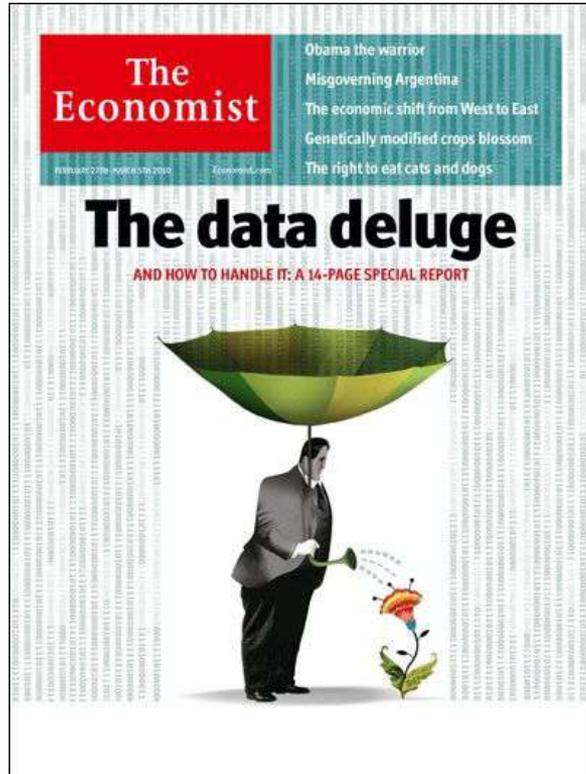




Contexte

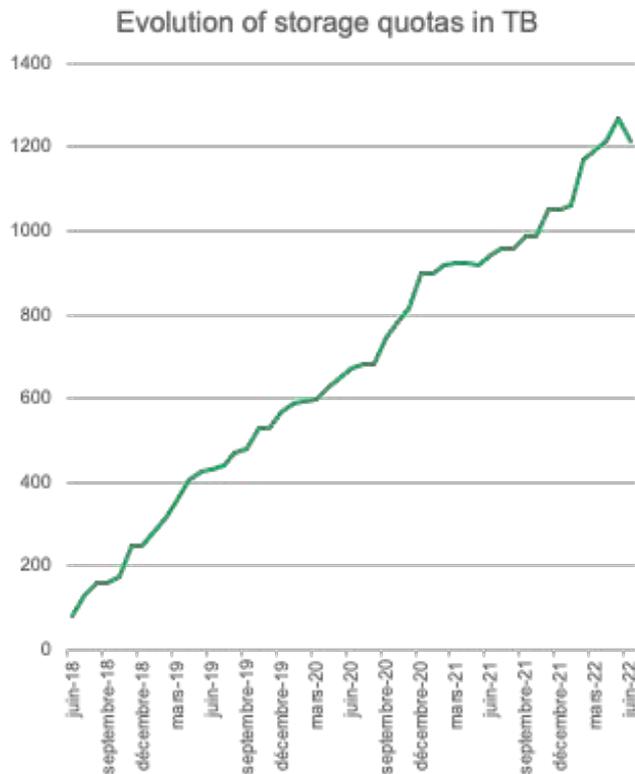
Science ouverte





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

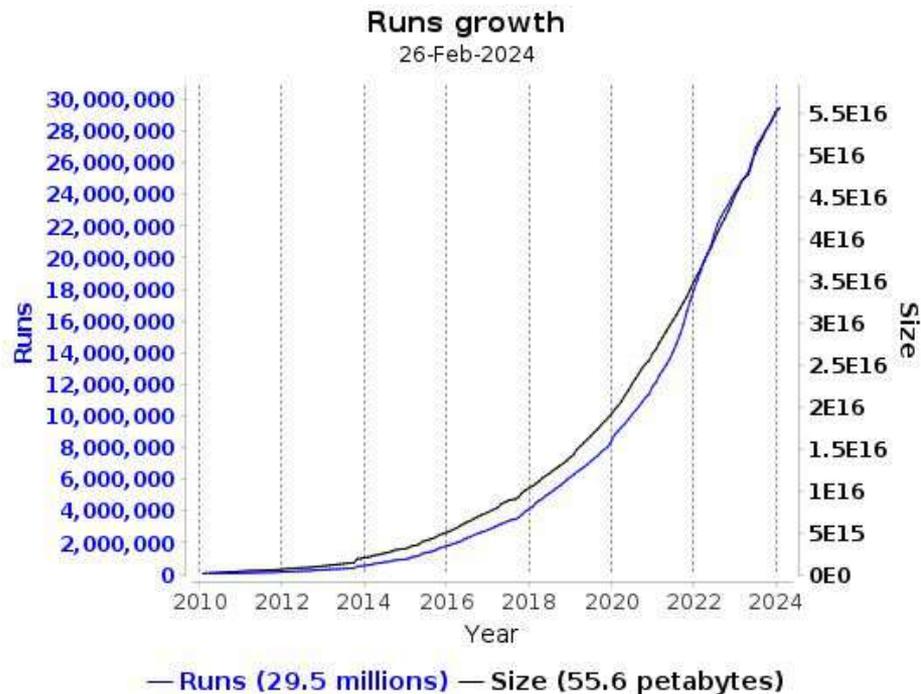
Les chercheurs accumulent les données sans stratégie de conservation à long terme

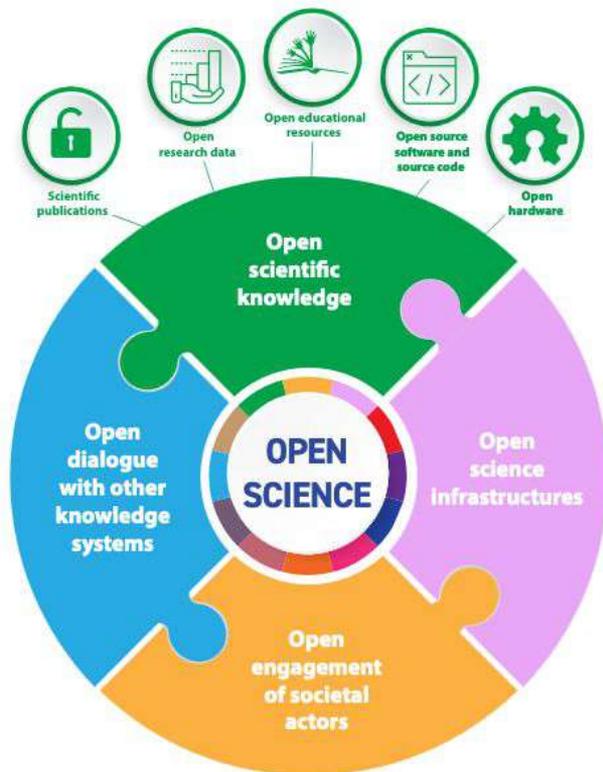
**Estimation basée sur un sondage auprès des équipes de recherche IGBMC en 2020*



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

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





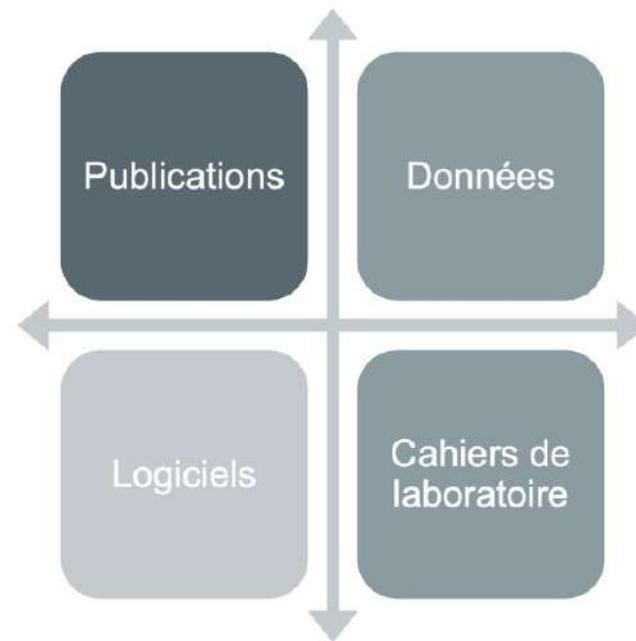
UNESCO Recommendation on Open Science, nov 2021

<https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>

Objectif : rendre la recherche accessible à tous

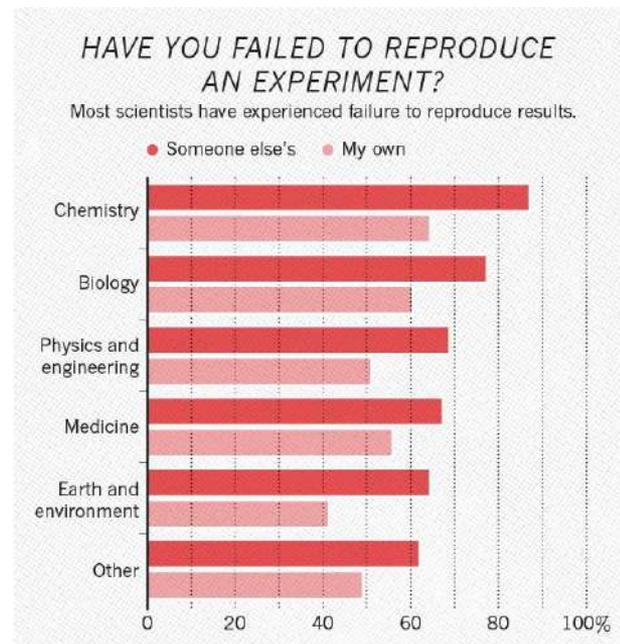
- Pas seulement l'accès à la **connaissance** elle-même
- Tout le processus de **création** et de **dissémination** de celle-ci
- La possibilité de **réutilisation**
- Ouverture au **dialogue** avec tous les acteurs, interdisciplinarité
- Engagement de et vers la **société**

- **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

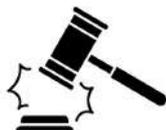


Monya Baker, 2016



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 »**

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

Mesures





RÉPUBLIQUE
FRANÇAISE

recherche.data.gouv.fr

Liberté
Égalité
Fraternité

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

UN ÉCOSYSTÈME AU SERVICE DU PARTAGE ET DE L'OUVERTURE DES DONNÉES DE RECHERCHE

Recherche Data Gov est un écosystème unique, composé de femmes et d'hommes qui ont décidé de mettre en commun leurs expertises, **pour accompagner les équipes de recherche.**

19 ATELIERS DE LA DONNÉE

en proximité géographique des équipes de recherche pour leur apporter une **première expertise** dans la gestion raisonnée des données de recherche.



4 CENTRES DE RÉFÉRENCE ÉTABLISSEMENTS

proposent depuis plusieurs années un **accompagnement propre à leurs orientations de recherche** et à destination de leurs équipes de recherche complètent l'écosystème de Recherche Data Gov.

6 CENTRES DE RÉFÉRENCE THÉMATIQUES

diffusent les bonnes pratiques et les standards internationaux de **gestion, traitement et diffusion des données dans leurs domaines scientifiques respectifs.** Ils complètent ainsi et soutiennent les ateliers de la donnée.



4 CENTRES DE RESSOURCES

complètent l'accompagnement local par des **services en ligne et des formations.**



1 ENTREPÔT PLURIDISCIPLINAIRE

une **solution souveraine de publication pour le partage et l'ouverture des données** aux communautés qui ne disposeraient pas encore d'un entrepôt thématique reconnu.

1 CATALOGUE

pour **repérer et signaler les données** déjà partagées ou ouvertes grâce à un entrepôt thématique français ou international.



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



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"

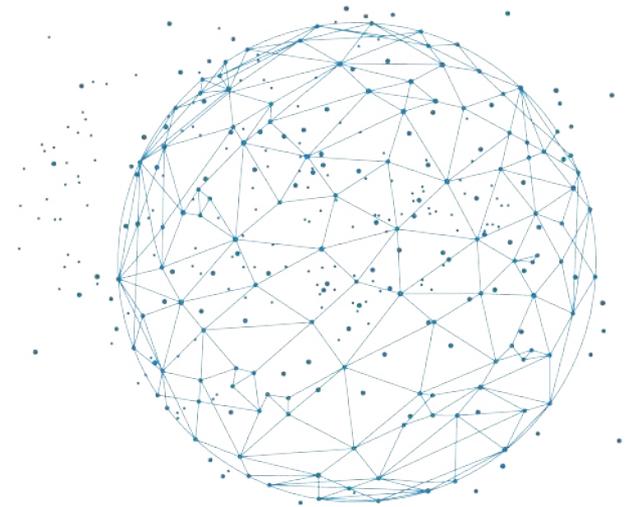
Entrepôts de données

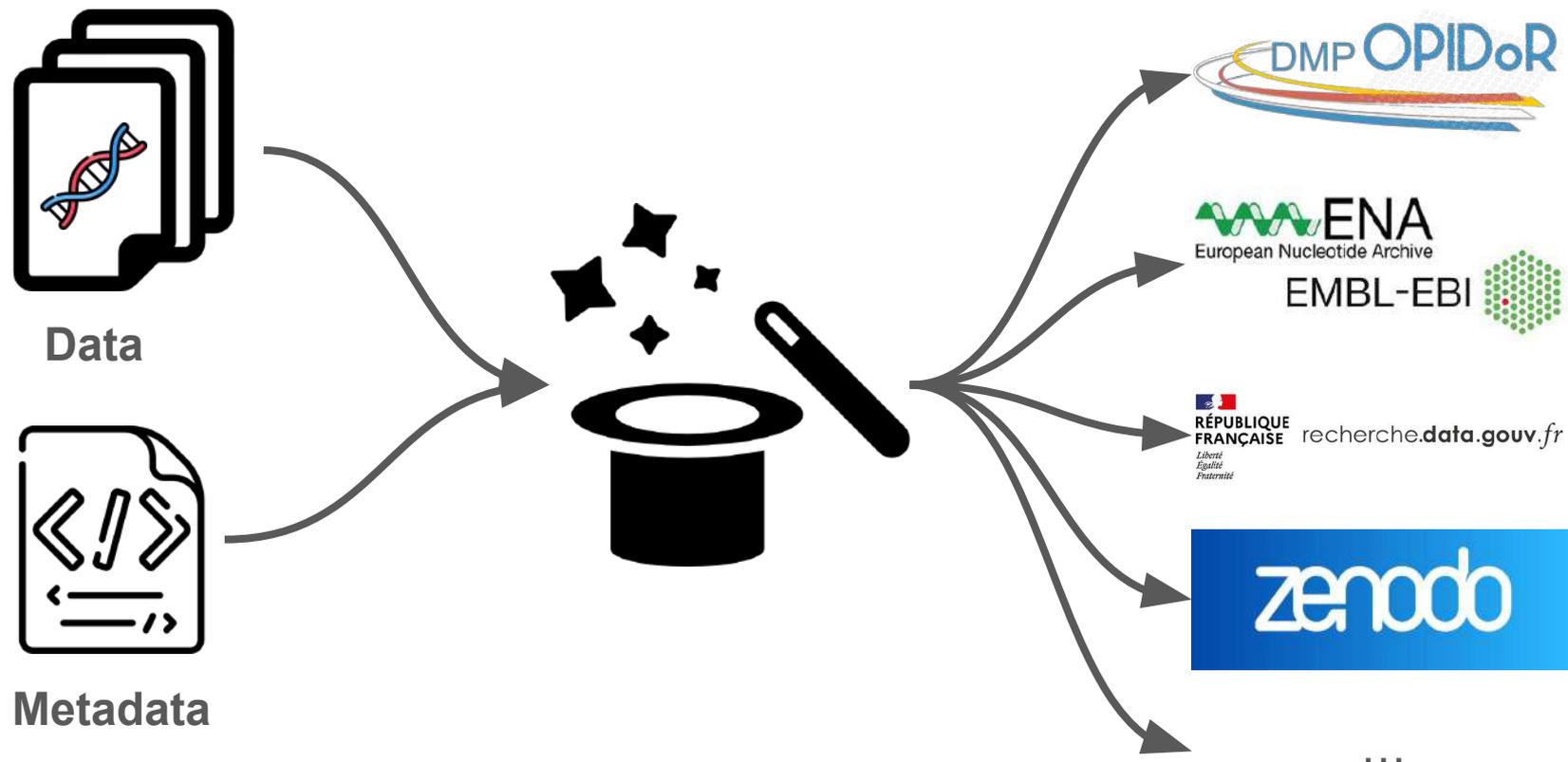
Open Access des publications

MADBOT
Metadata And Data Brokering Online Tool

Pourquoi Madbot ?

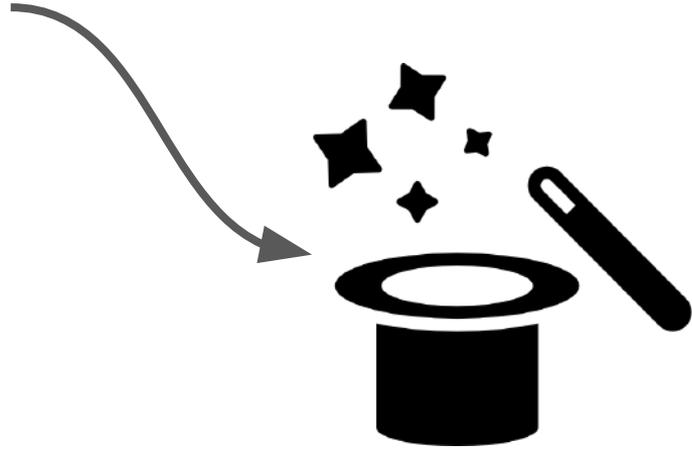
Revenons aux sources

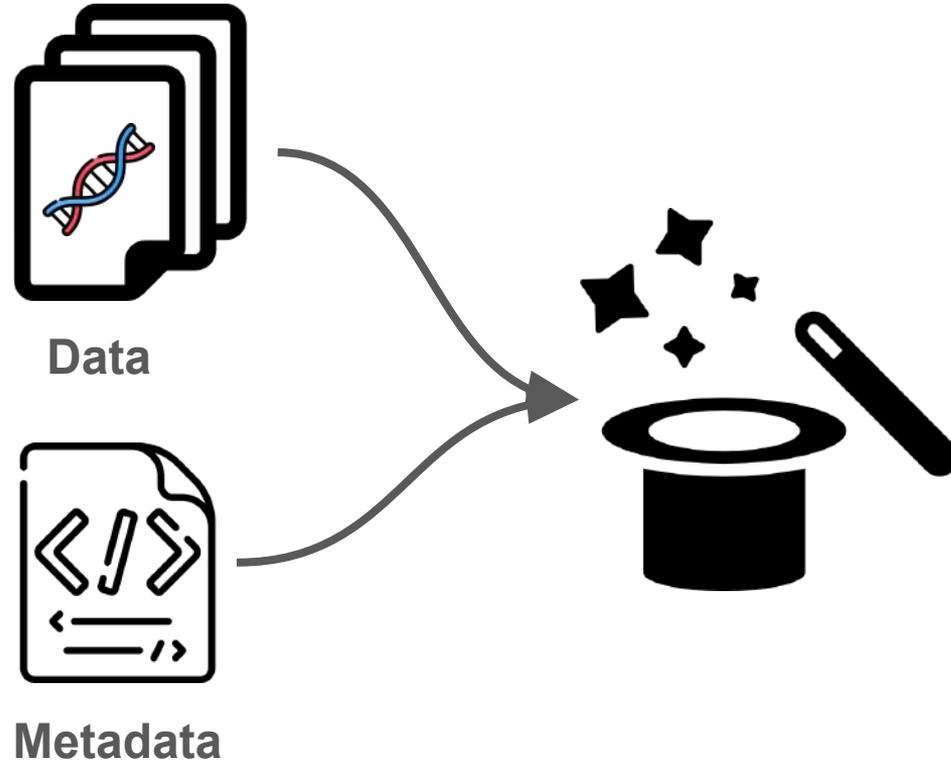






Data







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.

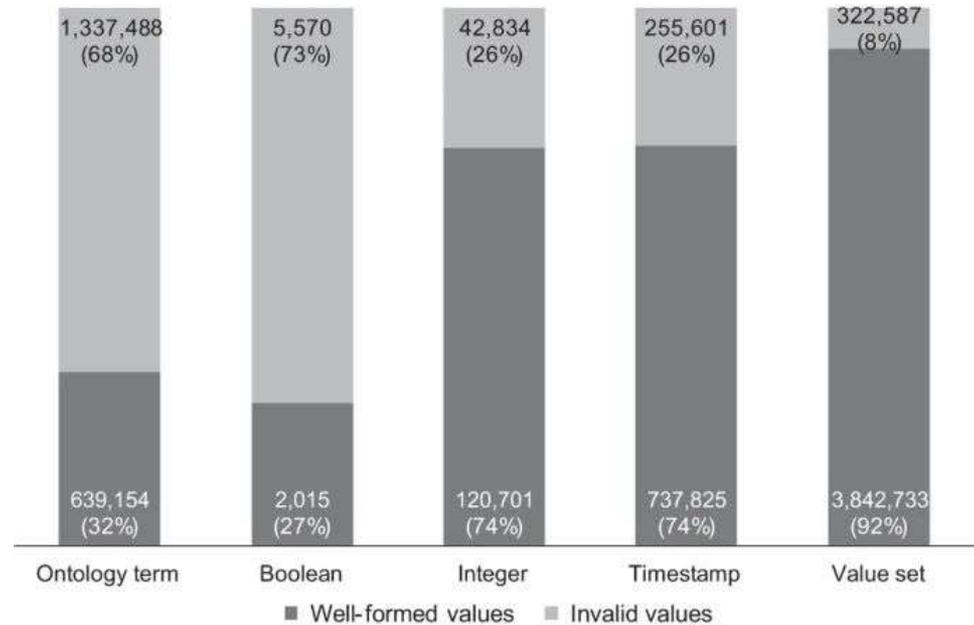
source: <https://fairsharing.org/educational/>



La soumission dans les entrepôts publics est souvent une tâche complexe

Les procédures de soumission sont hétérogènes.

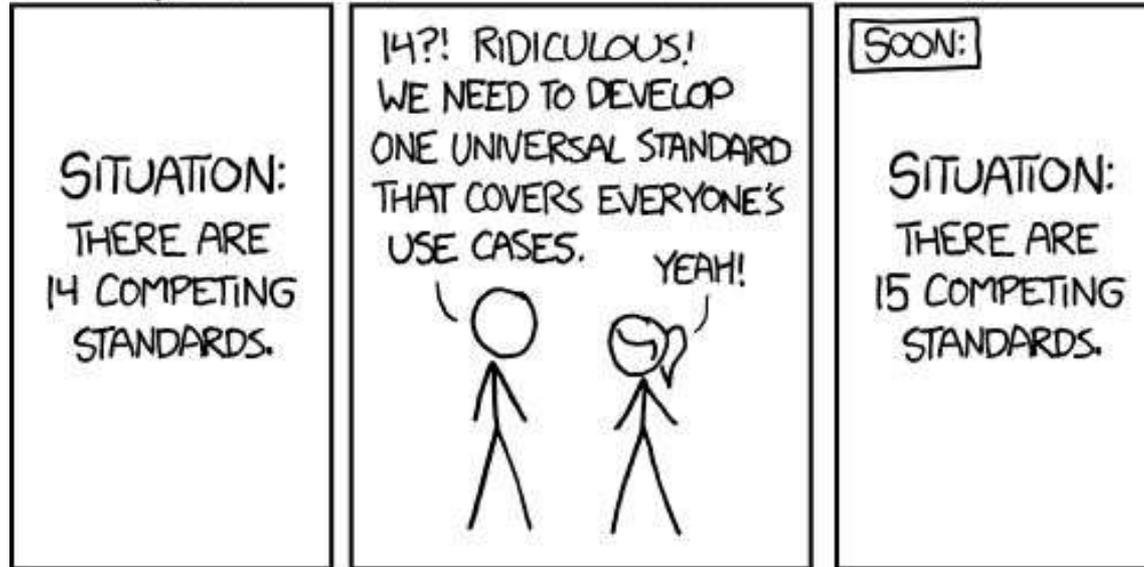
Les métadonnées sont souvent incomplètes, incohérentes, redondantes ou insuffisamment informatives.



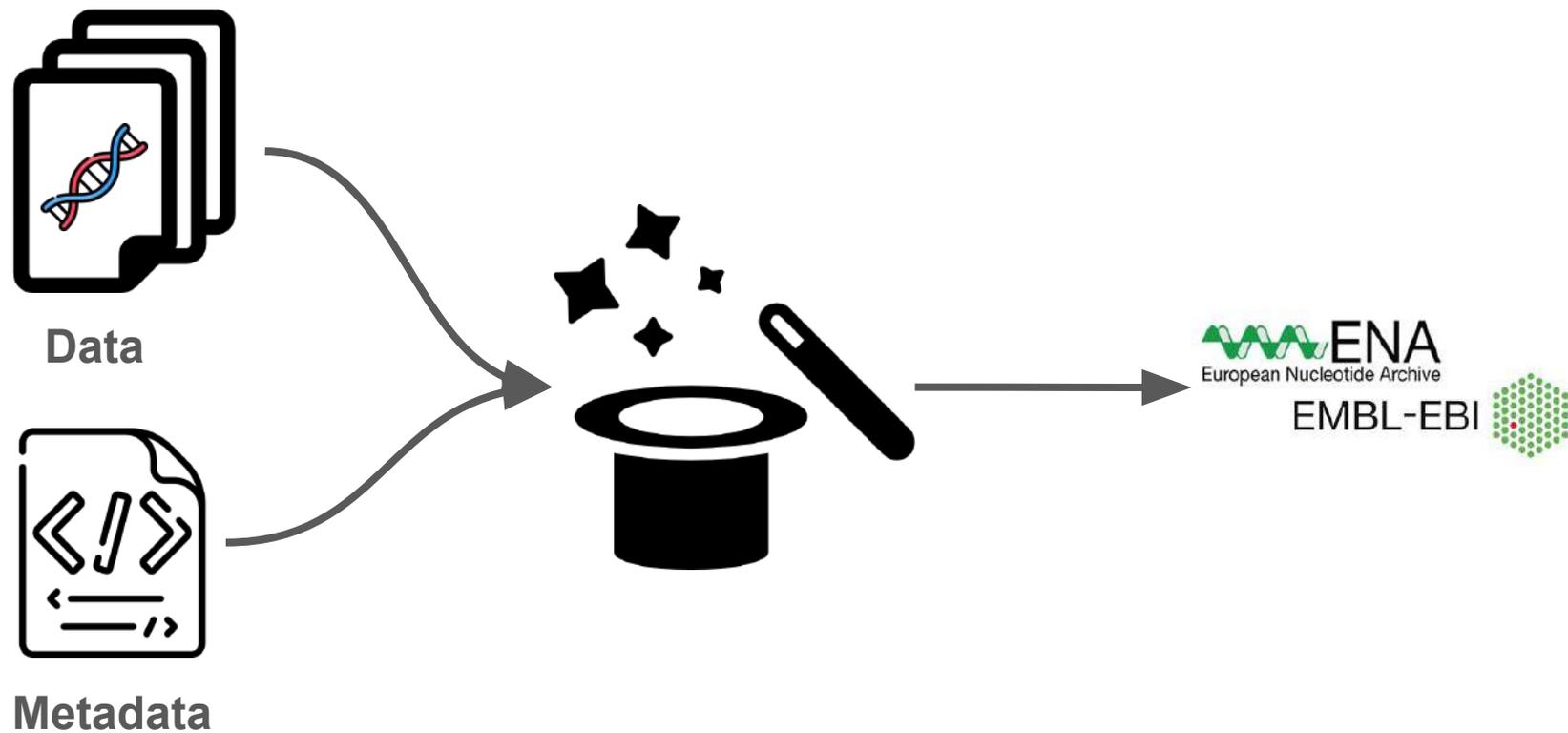
Quality of dictionary attributes in NCBI BioSample according to their type, in [Goncalves et al., 2019](#)



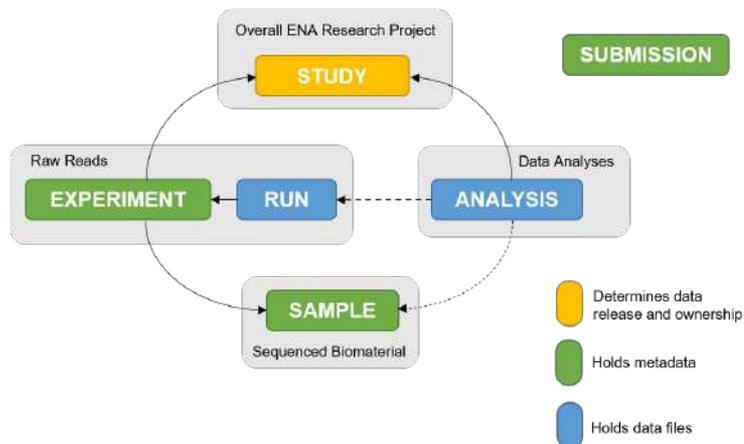
HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)



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



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>

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>

Checklist: ERC000020

Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

Checklist Fields

Filter fields...

Filter by type:

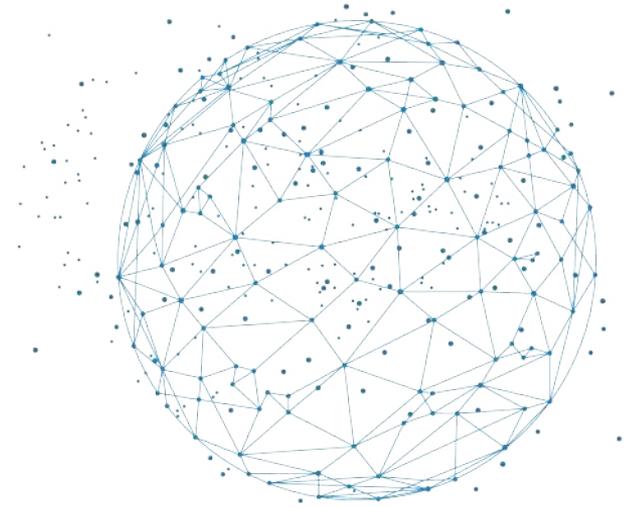
- non-sample terms
- Collection event information
- sample collection
- host identifier
- host description
- local environment conditions
- concentration measurement
- host details
- organism characteristics
- local environment conditions imposed

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
project name	free text		mandatory	
experimental factor	free text		optional	
ploidy	free text		optional	
number of replicons	restricted text	regular expression	optional	
extrachromosomal elements	restricted text	regular expression	optional	
estimated size	restricted text	regular expression	optional	
reference for biomaterial	free text		optional	
annotation source	free text		optional	
sample volume or weight for DNA extraction	restricted text	regular expression	optional	options
nucleic acid extraction	free text		optional	
nucleic acid amplification	free text		optional	
library size	restricted text	regular expression	optional	
library reads sequenced	restricted text	regular expression	optional	



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

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



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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.)





OpenLink



2020

2022

OmicsBroker



Metark



MADBOT

Sept. 2023



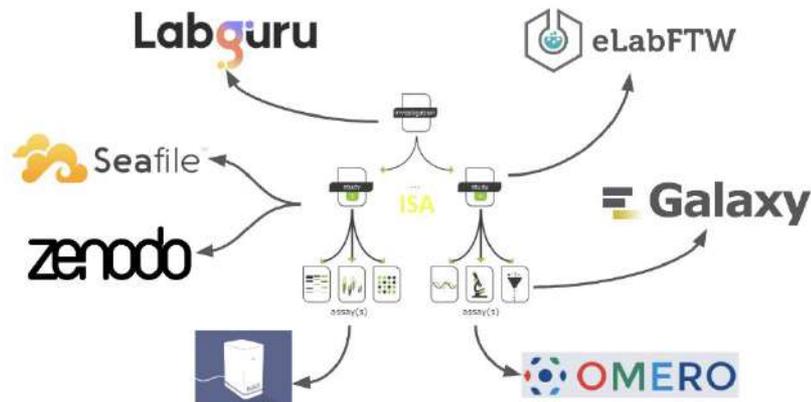
OpenLink

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

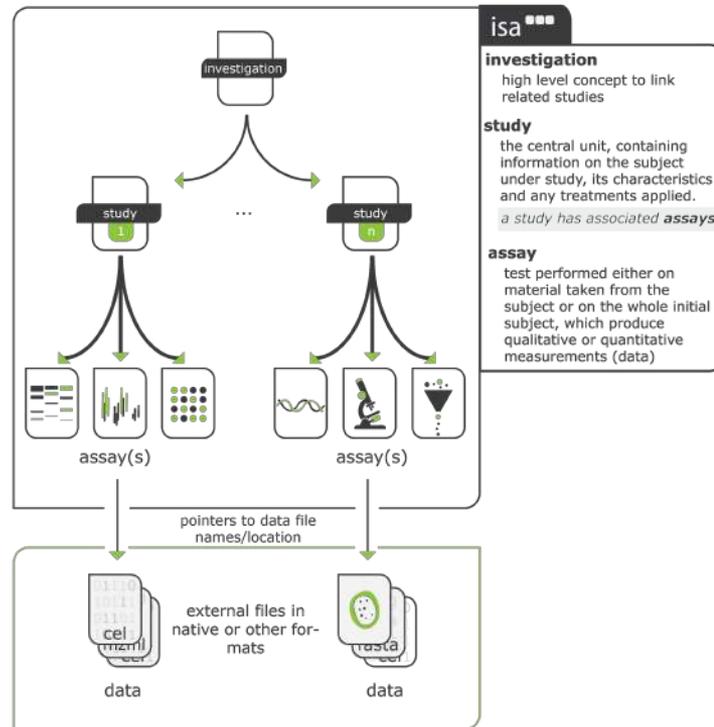
96K€ de financement sur 24 mois



Un standard pour les données des sciences de la vie

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

- **Investigation** : le contexte du projet
- **Study** : une expérimentation en un seul lieu
- **Assay** : 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>

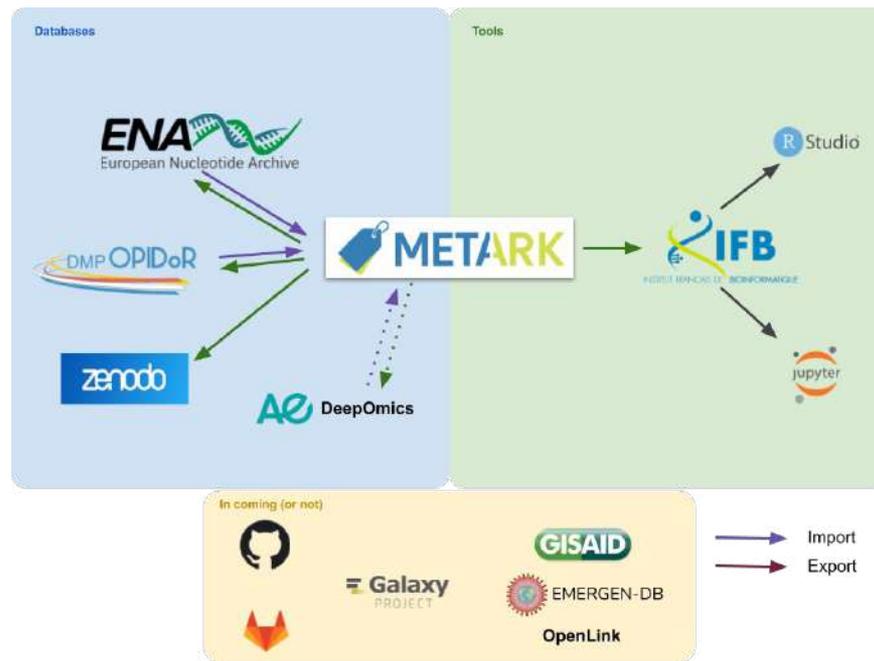


Un projet qui a débuté juste avant la collaboration IFB avec SpF pour la COVID-19

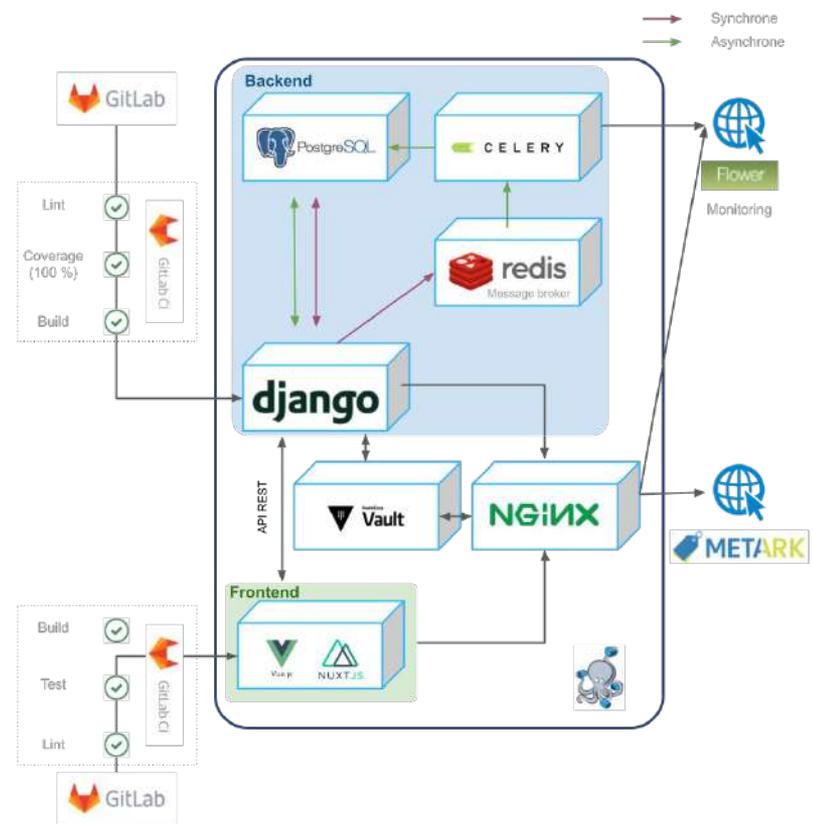
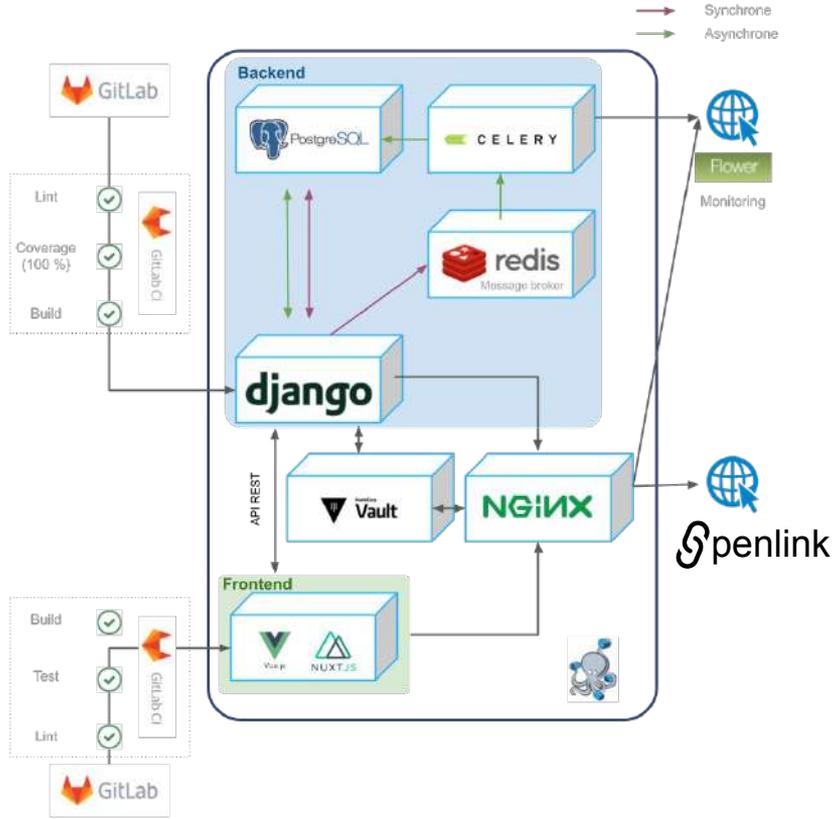
Metark un outil de data brokering pour faciliter la soumission des données et des métadonnées des sciences de la Vie dans des entrepôts internationaux

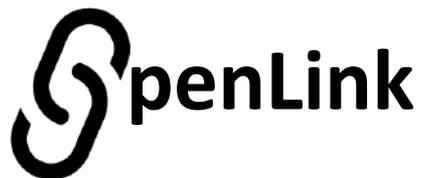
Sur une demande de l'ENA pour que l'IFB deviennent le data broker national.

Basé aussi sur un modèle ISA



Une architecture commune





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



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.

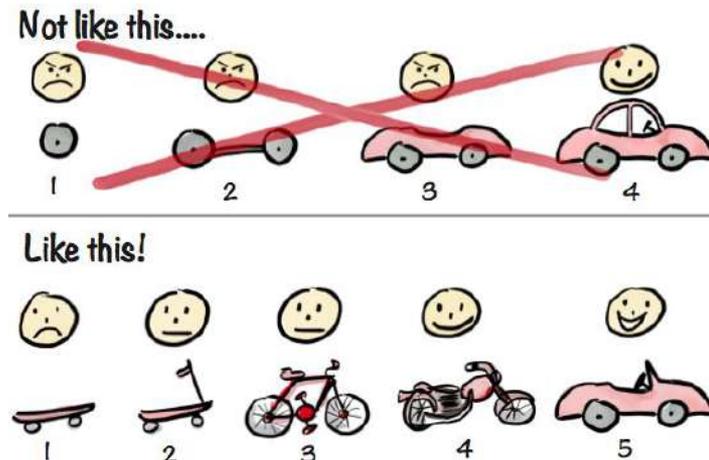
Chaque élément de cet arbre peut être associés à des données, des échantillons biologiques (qui sont également des nodes) et des métadonnées.

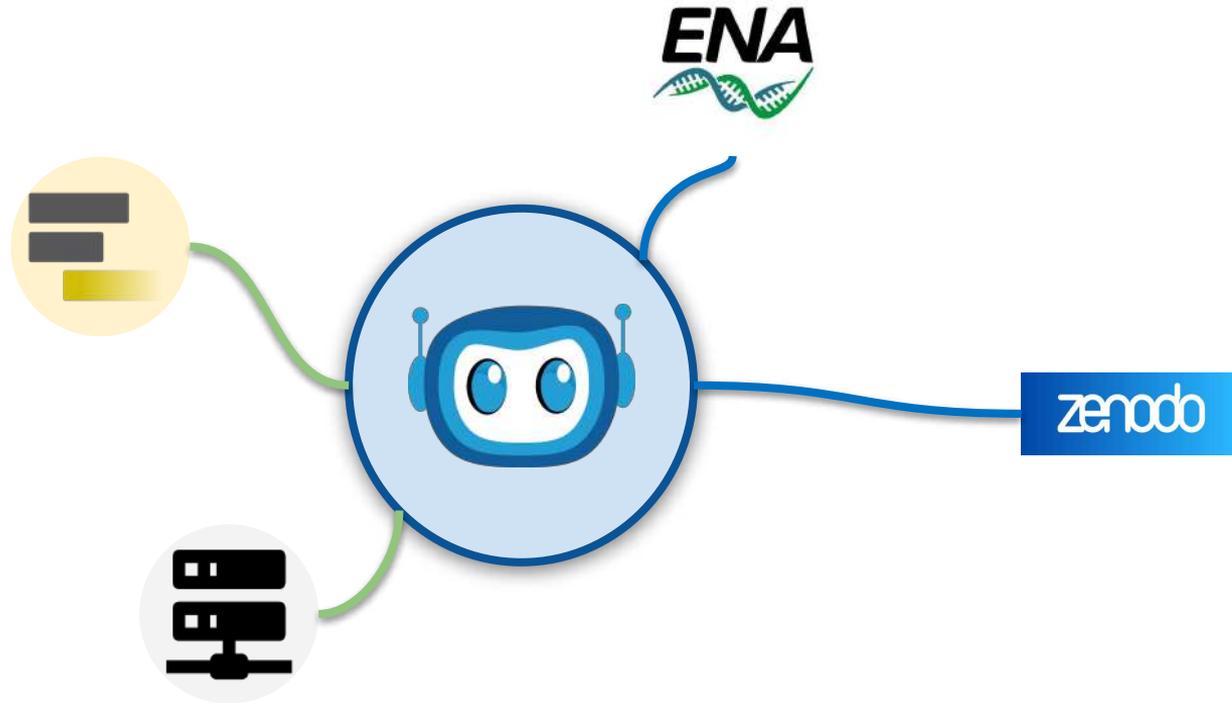
Nous souhaitons produire le plus rapidement possible une application fonctionnelle.

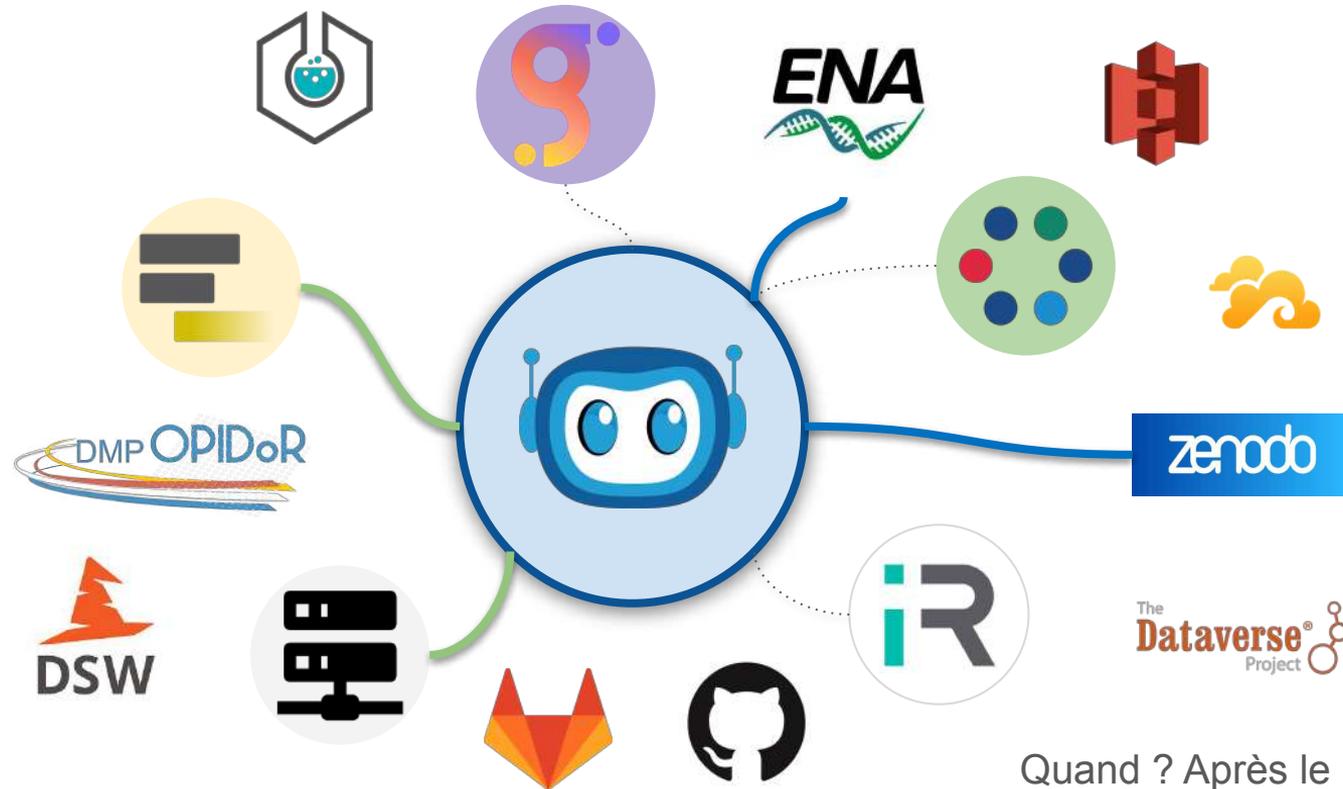
Objectifs principaux

- 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 **Minimal Viable Product (MVP)**







Quand ? Après le MVP ou
par contribution

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



Institut Français de Bioinformatique > Madbot



Madbot

Group ID: 75576438 [Leave group](#)



New project

Subgroups and projects Shared projects Archived projects

Q Search

Name



Madbot API

API of Madbot, the Metadata And Data Brokering Online Tool

★ 3

37 minutes ago



Madbot Client

Web client for Madbot, the Metadata And Data Brokering Online Tool

★ 1

4 days ago



Madbot doc

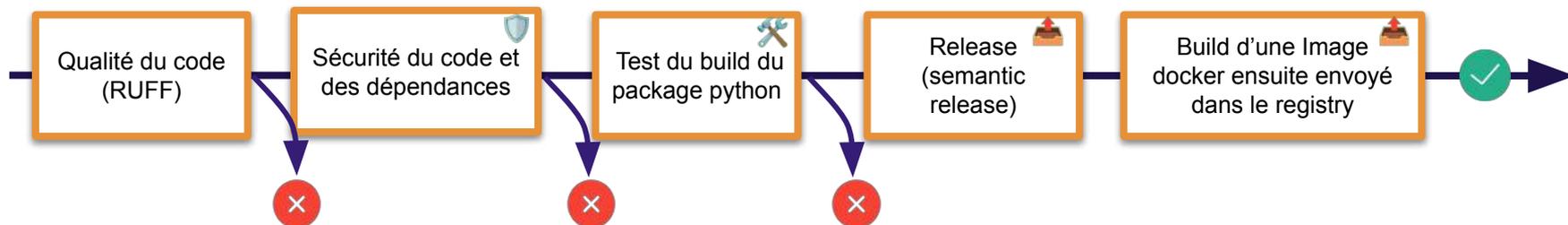
Documentation for Madbot, the Metadata And Data Brokering Online Tool

★ 3

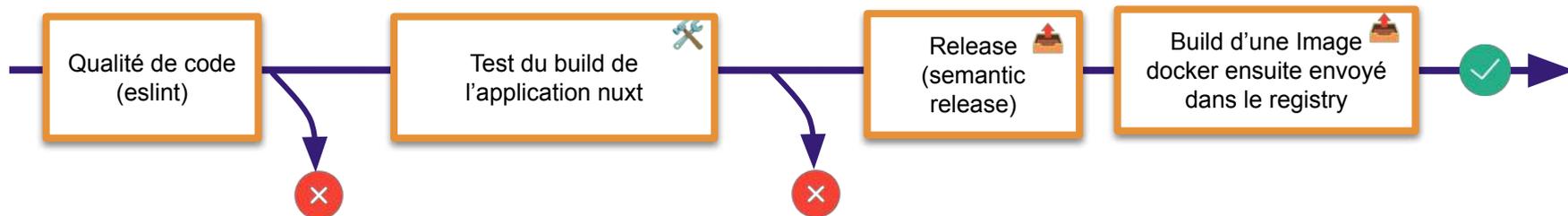
6 days ago



API



Client





Semantic release:

Gestion des **versions** et **publication des packages** entièrement automatisées

v1.0.0-dev.7

▼ Ressources 4

- Code source (zip) ↓
- Code source (tar.gz) ↓
- Code source (tar.bz2) ↓
- Code source (tar) ↓

Collecte de preuves

v1.0.0-dev.7evidences--7010556.json 6f5c21d1

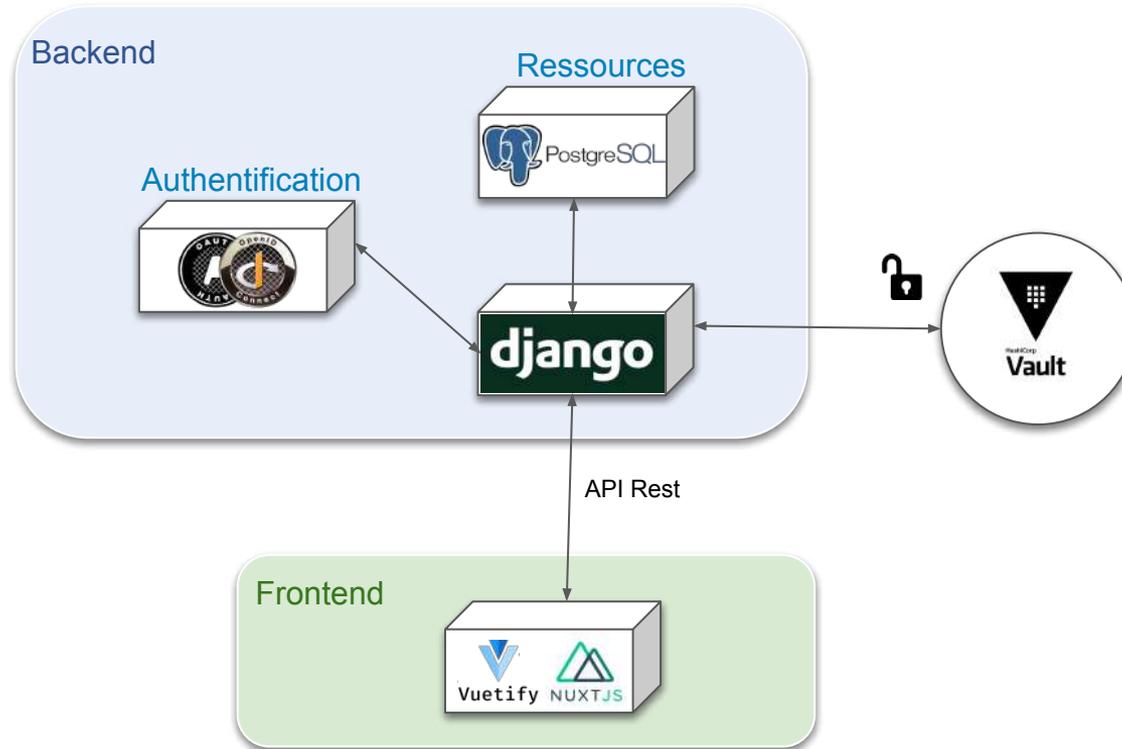
Collecté 1 day ago

[1.0.0-dev.7 \(2023-11-07\)](#)

Features

- Oauth toolkit (7dbe796)

ab503574 v1.0.0-dev.7 Publiée il y a un jour par



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

The screenshot shows the 'Parameters' page in Madbot. On the left is a vertical sidebar with a 'Parameters' header and a user profile for 'Thomas Denecker'. The main content area features a 'Welcome' message: 'Madbot allows you to identify and track the data and metadata associated with the different stages of your research projects. To get started, create your first investigation. An investigation is commonly associated to a general research question or goal.' Below this is a text input field labeled 'Node name' and a blue 'START' button.

2. Choisir comment créer le node



3. Page du node

Parameters
New investigation
Une démo

Une démo Everything saved

Une démo

investigation

No description

DATA LINKS **1** METADATA STUDY SAMPLE

NEW DATALINK

Name	Connections
No data available	

Items per page: 10 00 of 00

Thomas Denecker



4. Ajout d'une description

Une démo Save...

Une démo

Investigation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin posuitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper.

DATA LINKS METADATA STUDY SAMPLE

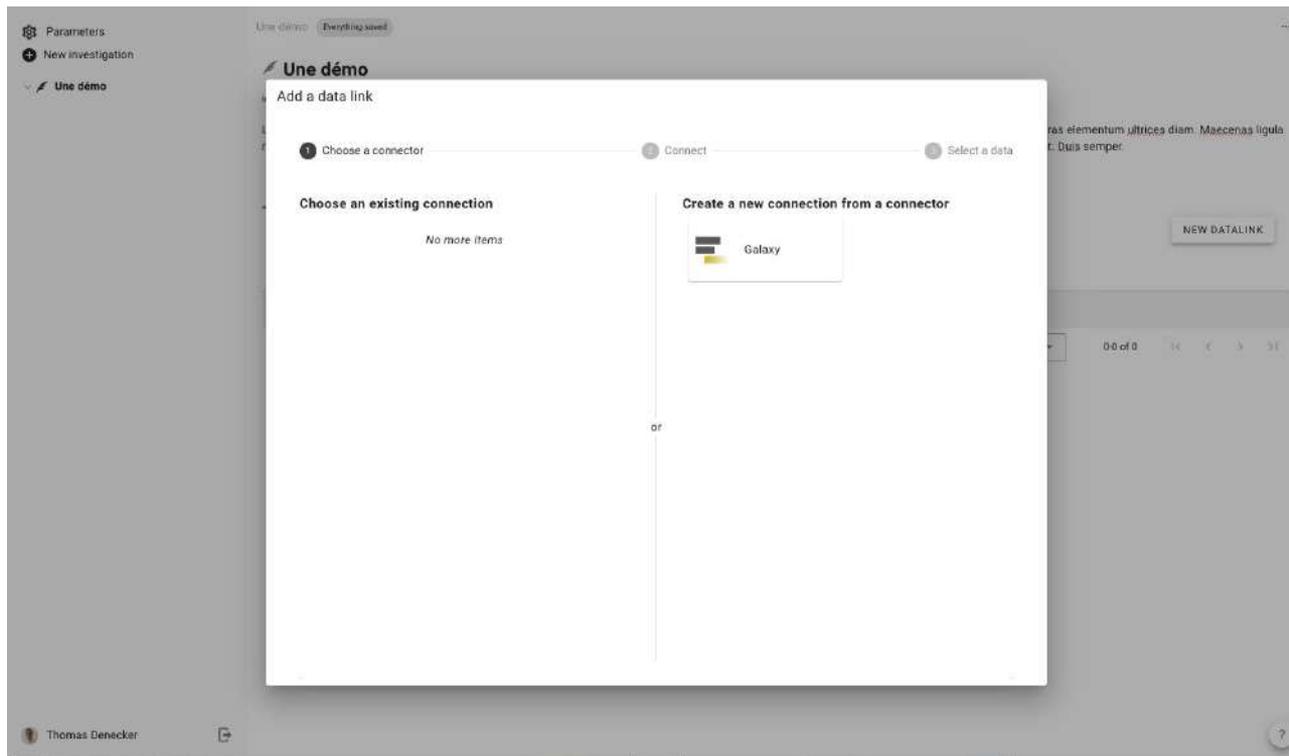
[NEW DATALINK](#)

Name	Connections
No data available	

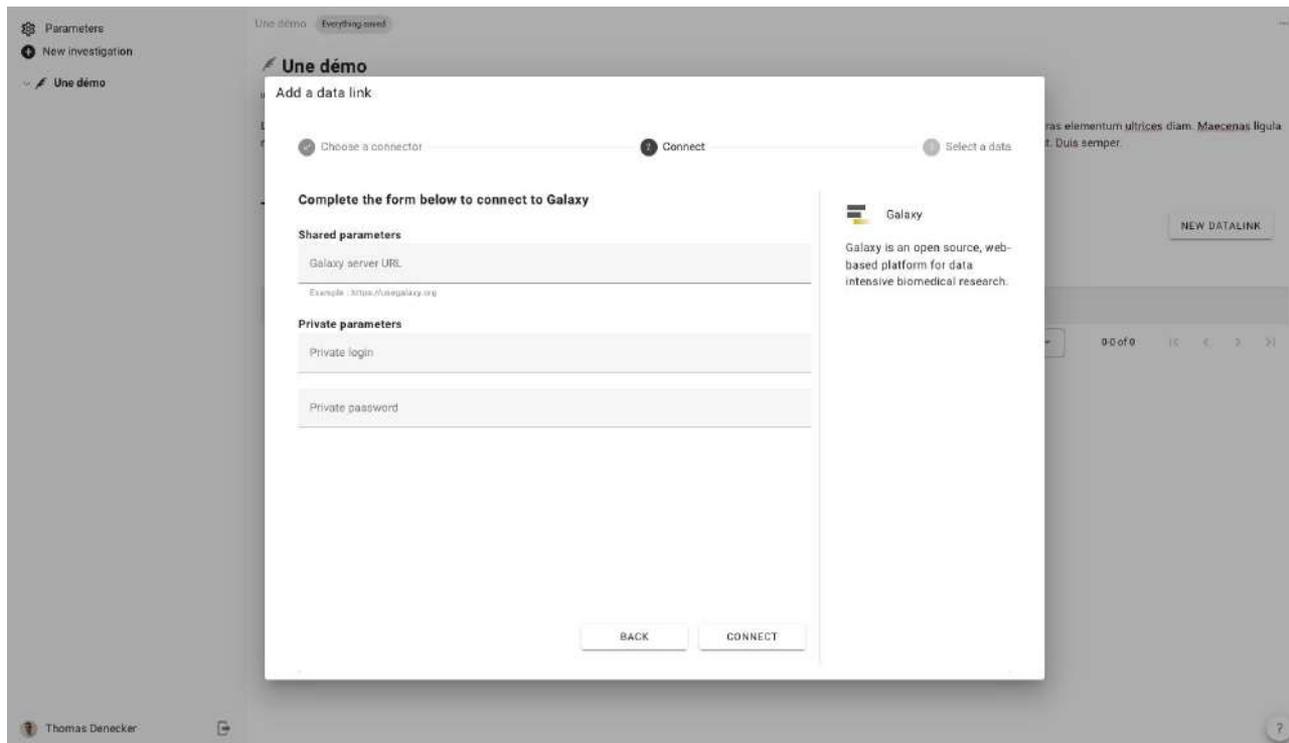
Items per page: 10 0-0 of 0

Thomas Denecker

4. Création d'un datalink



4. Création d'un datalink



The screenshot shows a web application interface with a modal dialog box titled "Une démo" and "Add a data link". The dialog has a progress bar with three steps: "Choose a connector" (checked), "Connect" (active), and "Select a data". Below the progress bar, the text "Complete the form below to connect to Galaxy" is displayed. The form is divided into "Shared parameters" and "Private parameters".

Shared parameters

Galaxy server URL
Example : <https://usegalaxy.org>

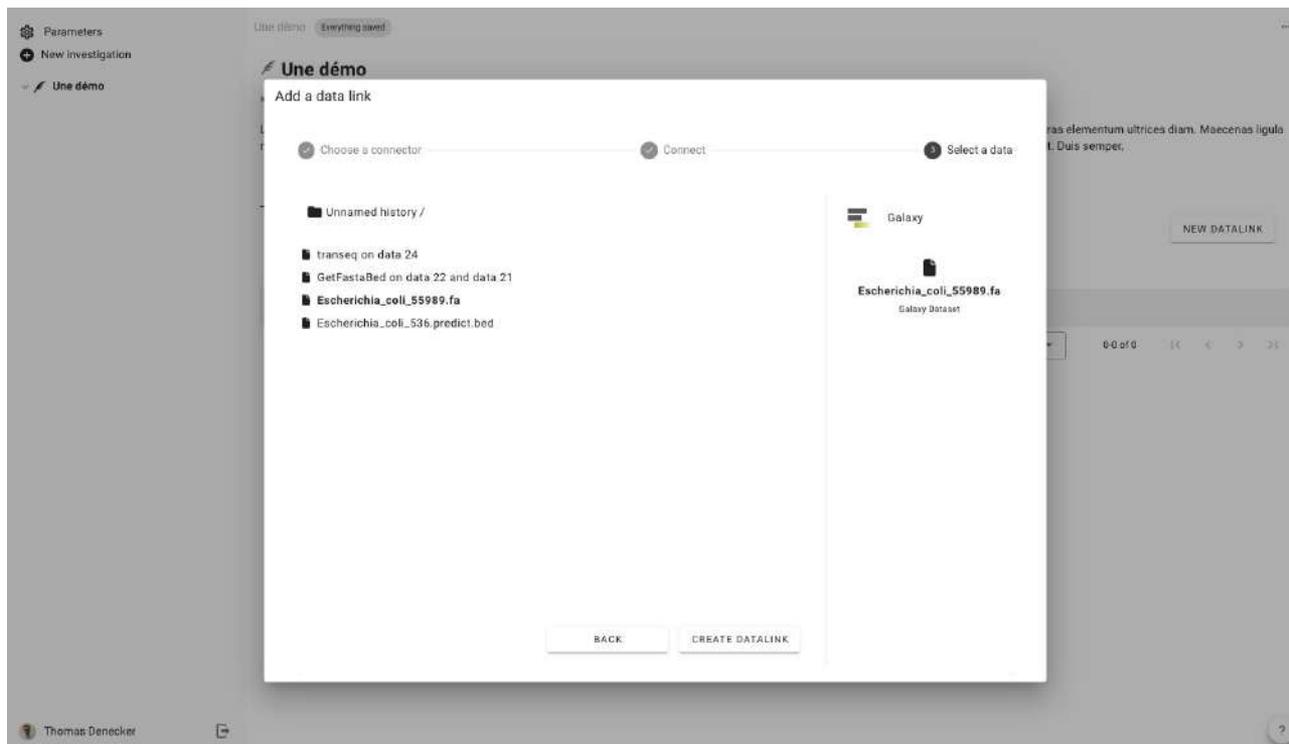
Private parameters

Private login

Private password

At the bottom of the dialog are "BACK" and "CONNECT" buttons. To the right of the form, there is a "Galaxy" logo and a description: "Galaxy is an open source, web-based platform for data intensive biomedical research." A "NEW DATALINK" button is visible in the background.

4. Création d'un datalink





5. Page du node

Une démo Everything saved

Une démo

Investigation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper.

DATA LINKS 1 METADATA STUDY SAMPLE

NEW DATALINK

Name	Connections
Curf/diff on data 181, data 179, and data 297. TSS groups FPKM tracking	

Items per page: 10 1-1 of 1

Thomas Denecker



5Bis. Page du node pour un projet plus riche

Parameters

New investigation

- Une démo
- Growth control of the eukaryote c...
- A time course analysis of trans...
- Study of the impact of change...**
 - metabolite profiling
 - protein expression profil...
 - transcription profiling

Growth control of the eukaryote cell: a systems biology study in yeast

Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast *Saccharomyces cerevisiae* under different nutrient limitations

Everything else

Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast *Saccharomyces cerevisiae* under different nutrient limitations

study

We wished to study the impact of growth rate on the total complement of mRNA molecules, proteins, and metabolites in *S. cerevisiae*, independent of any nutritional or other physiological effects. To achieve this, we carried out our analyses on yeast grown in steady-state chemostat culture under four different nutrient limitations (glucose, ammonium, phosphate, and sulfate) at three different dilution (that is, growth) rates ($D = u = 0.07, 0.1, \text{ and } 0.2/\text{hour}$, equivalent to population doubling times (T_d) of 10 hours, 7 hours, and 3.5 hours, respectively, $u =$ specific growth rate defined as grams of biomass generated per gram of biomass present per unit time).

DATA LINKS (0) METADATA ASSAY (0) SAMPLE

NEW DATALINK

Name	Connections
No data available	

Items per page: 10 0-0 of 0

Thomas Denecker

?

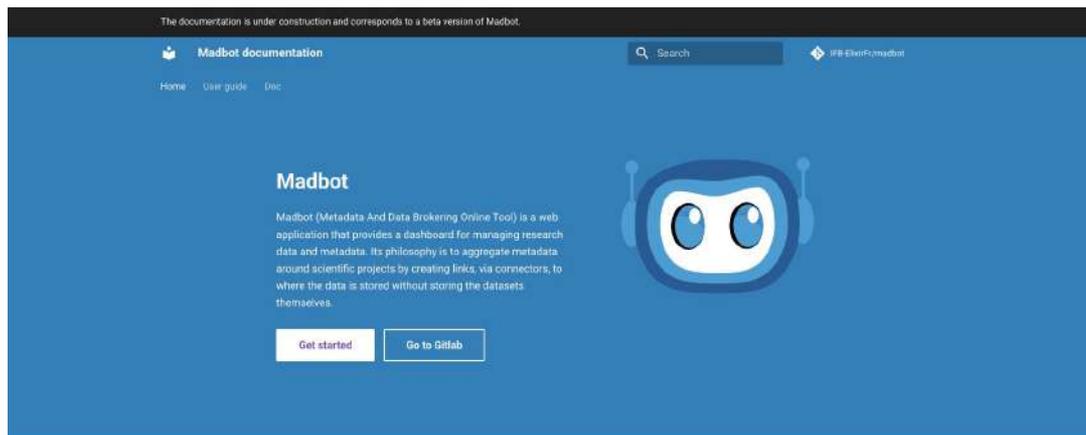
6. Gestion des membres

The screenshot displays the Madbot interface. On the left, a sidebar contains navigation options: 'Parameters', 'New Investigation', and a list of investigations including 'Growth control of the eukaryot...', 'A time course analysis of tran...', 'Study of the impact of change...', and 'Une démo'. The main area shows the details of the selected investigation, 'Growth control of the eukaryote cell: a systems biology study in yeast'. A modal window titled 'Growth control of the eukaryote cell: a systems biology study in ye...' is open, displaying the 'Members' section. This section includes a search bar and a table of members.

Member	Role
Anne Chymous @aonynous	manager
Baptiste Rousseau @brousseau	owner
John Doe @j.doe	owner

Below the table, there is a pagination control showing 'Items per page: 10' and '1-3 of 3'. A 'NEW LINK' button is visible at the bottom right of the modal window.

Pas encore très complète mais elle est en cours de rédaction



Our philosophy ? Connectors !

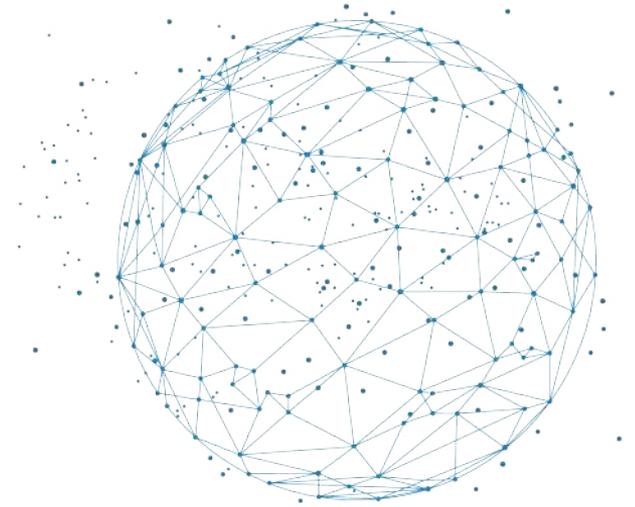
Its philosophy is to aggregate metadata around scientific projects by creating links, via connectors, to where the data is stored without storing the datasets themselves.

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 ?





- ❑ 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



- Bancher d'autres **systemes 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 !

