

# Data management based on Metadata





# Data Management Plan

How does the management of data is it funded, especially in the long term?

## Resources

What does the project consist of?  
Who are the partners?  
What policy on data management?  
Who is responsible for the management of data?

## Responsibilities in the project

What data will be produced/used during the course of the project (type, format, volume and increase...) ?  
How will they be produced?

## Data collection

Who will be the owner of the data produced?  
How will they be used?

## Intellectual Property

How, where, by whom, will be stored, backed up and secured the data?

## Data backup

Who will be able to access the data? The data will they be shared? published? With whom? How? How long does it take? Under which license?

## Data Access and Data sharing

How will the data be identified, described? What metadata standards will be used? How will the metadata be generated?

## Data Documentation

What is the plan for long-term archiving and preservation?

## Data Archiving

**Planning must be followed by implementation and therefore concrete actions**



# Data management



## Being able to easily describe your data (descriptive metadata)

- with its professional vocabulary
- without tedious entries
- without having to re-enter the same information each time
- by associating external resources (links)



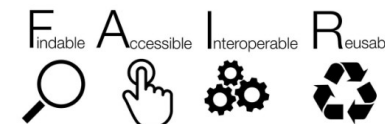
## Being able to easily manage your data

- by limiting data loss (after the departure of temporary staff)
- by sharing only metadata
- be able to easily find data (from metadata)
- be able to provide access to data if necessary
- be able to distribute them without having to re-enter everything



## Ensuring metadata follows FAIR principles

- Respect a standard (metadata schema)
- Use controlled vocabulary consistent with your domain (thesaurus, ontologies)
- Be at least “Findable, Accessible & Interoperable”





# Data management

## The Ariadne's thread ...

### Organize

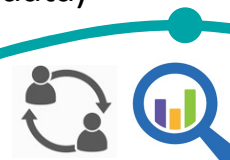
Your workspace

(storage, backup, naming, ...)



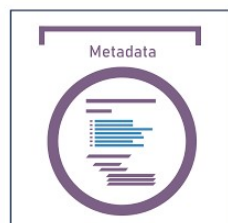
Share & Search data/metadata

(metadata)



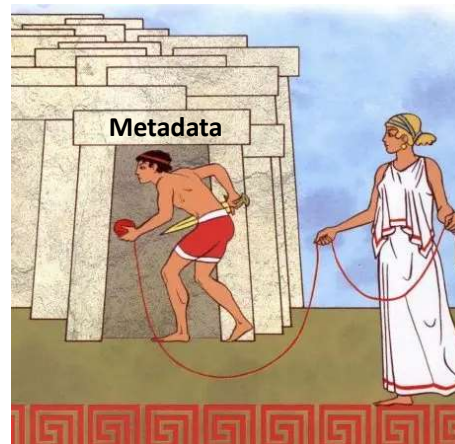
Publish your data

(with metadata)



Describe your data

(metadata)





# Data management

## The Ariadne's thread ...

### Organize

Your workspace

(storage, backup, naming, ...)

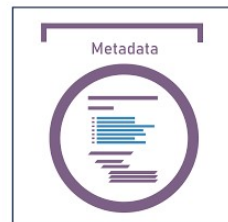
Share & Search data/metadata

(metadata)

Publish your data

(with metadata)

Standards & Terminologies  
4 FAIR pillars



Describe your data  
(metadata)



Metadata schema



(what to describe)



Controlled Vocabulary



SKOSMOS

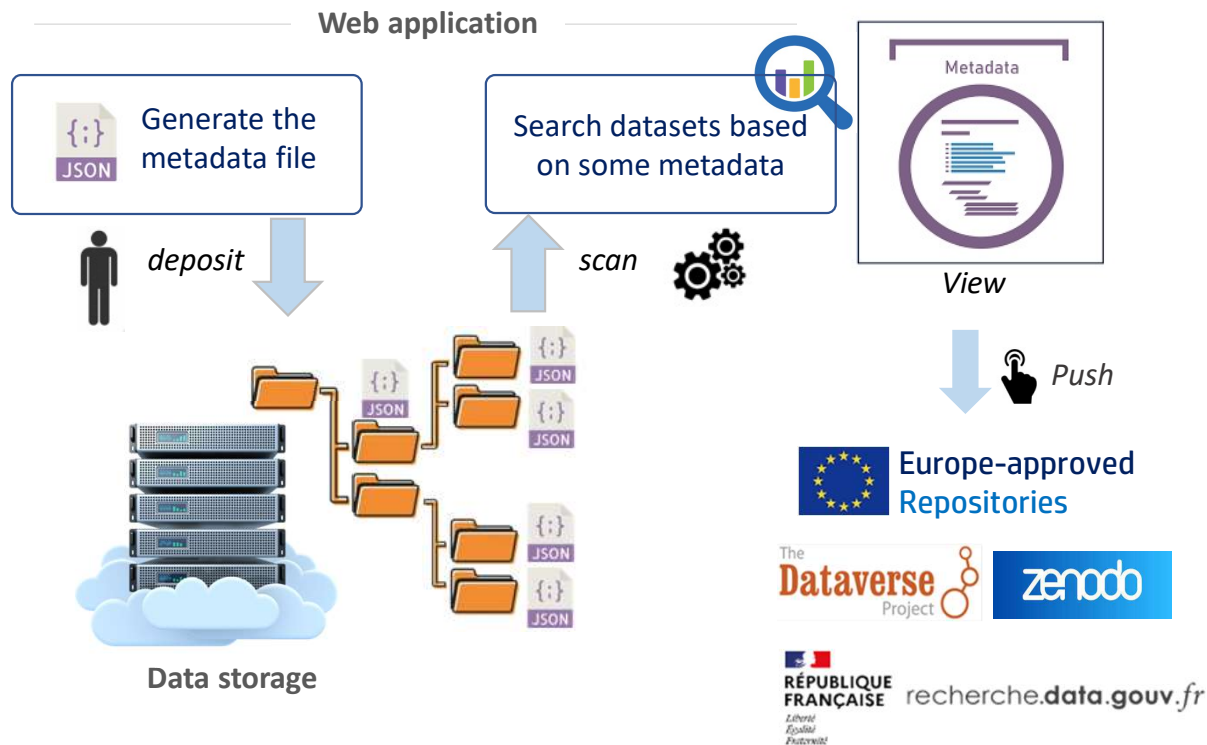
Dictionaries

AgroPortal, BioPortal,  
VOINRAE, LOTERRE, ONTOSTACK,  
...

(How to describe)



# An ecosystem for metadata management

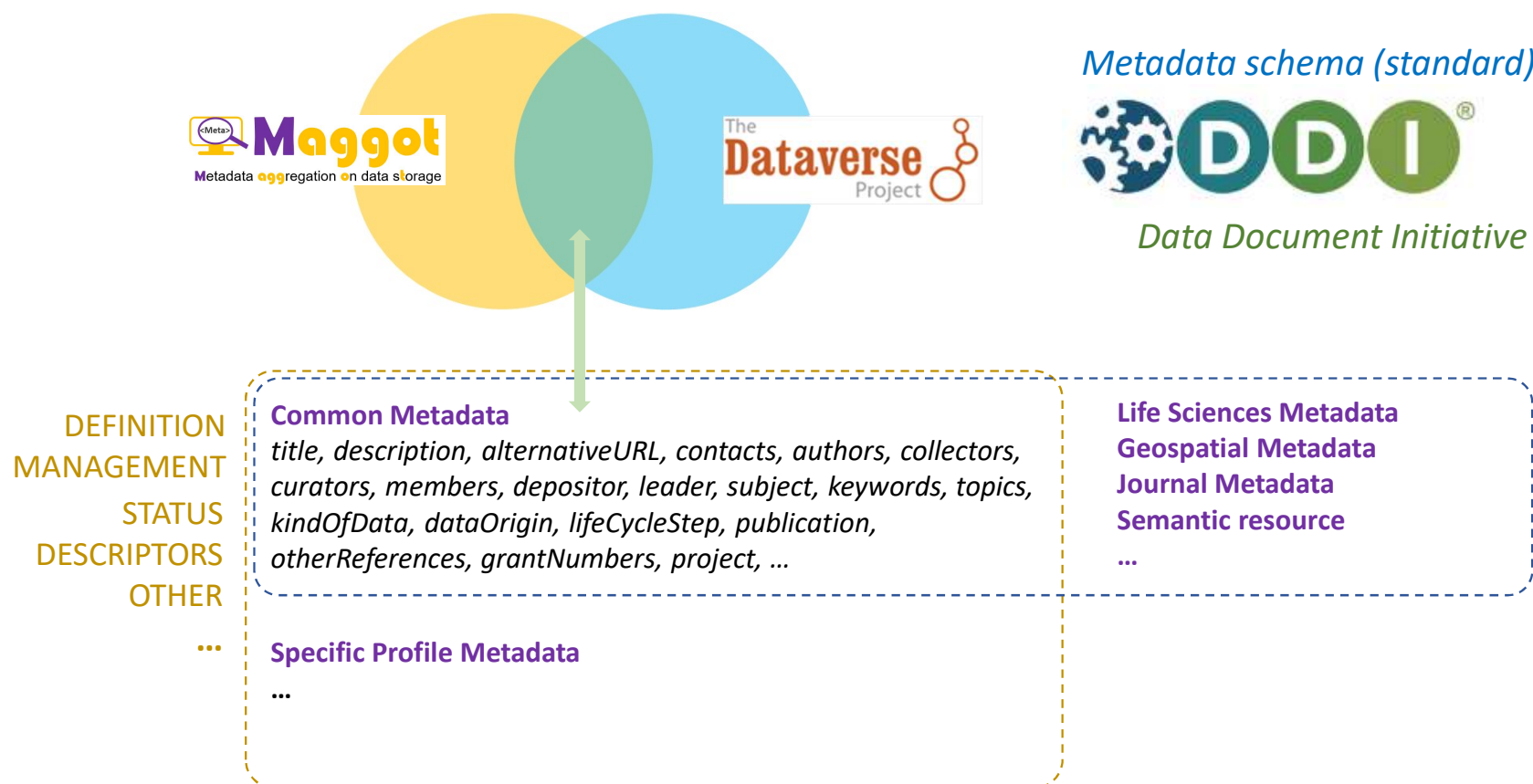


**Describe metadata**  
with controlled vocabulary

**Share & Search**  
metadata / data

**Publish metadata**  
... with data

Adoption of the *Common Metadata* of the schema implemented within the *Harvard Dataverse software*



## Common Metadata of « Dataverse »

### Common Metadata

*title, description, alternativeURL, contacts, authors, collectors, curators, members, depositor, leader, subject, keywords, topics, kindOfData, dataOrigin, lifeCycleStep, publication, otherReferences, grantNumbers, project, ...*

*Mandatory fields*

*Recommended fields*

*Desirable fields*

DEFINITION \*  
STATUS  
MANAGEMENT \*  
DESCRIPTORS \*  
OTHER  
RESOURCES

Short name \* ⓘ  
  
Full title \* ⓘ  
  
Subject \* ⓘ  
☐ Agricultural Sciences ☐ Arts and Humanities ☐ Astronomy and Astrophysics ☐ Business and Management ☐ Chemistry  
☐ Computer and Information Science ☐ Earth and Environmental Sciences ☐ Engineering ☐ Law ☐ Mathematical Sciences  
☐ Medicine Health and Life Sciences ☐ Other ☐ Physics ☐ Social Sciences  
Description of the dataset \* ⓘ

\* mandatory fields

DEFINITION \*  
STATUS  
MANAGEMENT \*  
DESCRIPTORS \*  
OTHER  
RESOURCES

Kind of Data \* ⓘ  
☐ Audiovisual ☐ Collection ☐ Dataset ☐ Event ☐ Image ☐ Interactive Resource ☐ Model ☐ Other ☐ Physical Object  
☐ Service ☐ Software ☐ Sound ☐ Text ☐ Workflow  
Keywords ⓘ  
  
Search a value:  enter the first letters  
Topic Classification ⓘ  
  
Search a value:  enter the first letters  
Data origin ⓘ  
☐ Other ☐ aggregate data ☐ analysis data ☐ audiovisual corpus ☐ computer code ☐ experimental data ☐ observational data  
☐ simulation data ☐ survey data ☐ text corpus

## Specific Profile Metadata

Experimental Factor ⓘ  
  
Search a value:  enter the first letters  
Measurement type ⓘ  
  
Search a value:  enter the first letters  
Technology type ⓘ  
  
Search a value:  enter the first letters



*List of well-chosen and limited CVs (according to a reference e.g. Data Document Initiative)*

**Kind of Data** \* ?

☐ Audiovisual ☐ Collection ☐ Dataset ☐ Event ☐ Image ☐ Interactive Resource ☐ Model ☐ Other ☐ Physical Object

☐ Service ☐ Software ☐ Sound ☐ Text ☐ Workflow

**Keywords** ?

Search a value:

*List of ontologies to choose according to your domain*

**Topic Classification**

Search a value:

- Experimental** Model of Disease (NCBITAXON)
- experimentally** modified cell in vitro (OBI)
- experimental**\_feature (OBI)
- experimental** infection of cell culture (OBI)
- experimental** disease induction (OBI)
- Experimental** measurement (EDAM)

*Use of dictionaries to target the CV by mixing thesaurus and ontologies*

**Thesaurus SKOSMOS**

(VOINRAE, LOTERRE, ONTOSTACK, ...)



**Building a vocabulary**

<https://vocabulaires-ouverts.inrae.fr/construire/>

NAME (*)	ONTOLOGY	URL	Add new	
NMR spectroscopy assay	OBI	http://purl.obolibrary.org/obo/OBI_0000623	Edit	Del
agricultural science	EDAM	http://edamontology.org/topic_3810	Edit	Del
amino acid	IOBC	http://purl.jp/bio/4/id/200906089657456524	Edit	Del
analyte assay	MS	http://purl.obolibrary.org/obo/OBI_0000443	Edit	Del
biochemical analysis	IOBC	http://purl.jp/bio/4/id/200906072808564316	Edit	Del
biochemical characterization	IOBC	http://purl.jp/bio/4/id/201306093820876862	Edit	Del
biochemical composition	IOBC	http://purl.jp/bio/4/id/201106016579695836	Edit	Del
biochemistrv	EDAM	http://edamontology.org/topic_3292	Edit	Del

The use of **dictionaries within Maggot** has no other purpose to **facilitate the entry of metadata**, entry which can be long and repetitive in generalist data warehouses (such as repository based on Dataverse).

LAST NAME (*)	FIRST NAME (*)	INSTITUTE (*)	ORCID	EMAIL	<input type="button" value="Add new"/>	
			5828	bordeaux.fr	<input type="button" value="Edit"/>	<input type="button" value="Del"/>
Dai	Zhanwu	UMR 1287 EGFV, INRAE			<input type="button" value="Edit"/>	<input type="button" value="Del"/>
Deborde	Catherine	UMR 1332 BFP INRAE	0000-0001-5687-9059	catherine.deborde@inrae.fr	<input type="button" value="Edit"/>	<input type="button" value="Del"/>
<input type="text" value="Dussarrat"/>	<input type="text" value="Thomas"/>	<input type="text" value="UMR 1332 BFP INRAE"/>	<input type="text" value="0000-0001-6245-365"/>	<input type="text" value="thomas.dussarrat@inrae.fr"/>	<input type="button" value="Save"/>	<input type="button" value="Cancel"/>
Eveillard	Sandrine	Biologie du Fruit et Pathologie Facility, France <i>BFP</i>	002-8078-	sandrine.eveillard@inrae.fr	<input type="button" value="Edit"/>	<input type="button" value="Del"/>
Fouillen	Laetitia				<input type="button" value="Edit"/>	<input type="button" value="Del"/>
Gautier	Roselyne	National Research Institute for Agriculture, Food and Environment			<input type="button" value="Edit"/>	<input type="button" value="Del"/>
Giauffret	Catherine	Government, France	002-1469-		<input type="button" value="Edit"/>	<input type="button" value="Del"/>

**Dictionaries** allow you to record **multiple information** necessary to **define an entity**, such as the names of people, or even the funders.

Its information, once entered and saved in a file called a dictionary, can be **subsequently associated with the corresponding entity**.

## Example : people

**Contacts**

Canlet Cécile, Deborde Catherine

Add a value: enter the first three letters

**Authors**

Canlet Cécile

Add a value: de

Giraudeau Patrick

**Project Lead** Deborde Catherine

Cahoreau Edern

Add a value: enter the first three letters

**Data Collector**

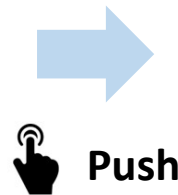
Add a value: enter the first three letters

**Data Curator**

Add a value: enter the first three letters

**Data Member**

Add a value: enter the first three letters



**Contact** ? Use email button above to contact.

Canlet, Cécile (INRAE)  
Deborde, Catherine (INRAE)

**Author** ? Canlet, Cécile (INRAE) - ORCID: 0000-0002-6389-0712  
Deborde, Catherine (INRAE) - ORCID: 0000-0001-5687-9059

**Contributor** ? Project Leader : Giraudeau, Patrick (Univ. Nantes) - ORCID: 0000-0001-9346-9147  
Data Collector : Deborde, Catherine (INRAE) - ORCID: 0000-0001-5687-9059  
Data Collector : Canlet, Cécile (INRAE) - ORCID: 0000-0002-6389-0712  
Data Collector : Gautier, Roselyne (INRAE)  
Data Collector : Jousse, Cyril (Univ. Clermont Auvergne) - ORCID: 0000-0002-5899-8243  
Data Collector : Lacaze, Méliia (INRAE)  
Data Collector : Martineau, Estelle (Univ. Nantes) - ORCID: 0000-0001-5093-2138  
Data Collector : Peyriga, Lindsay (INRAE) - ORCID: 0000-0002-6138-7961  
Data Collector : Richard, Tristan (Univ. Bordeaux) - ORCID: 0000-0002-5308-8697  
Data Collector : Silvestre, Virginie (Univ. Nantes)  
Data Collector : Traïkia, Mounir (Univ. Clermont Auvergne) - ORCID: 0000-0002-4595-0400  
Data Curator : Deborde, Catherine (INRAE) - ORCID: 0000-0001-5687-9059  
Data Curator : Canlet Cécile (INRAE) - ORCID: 0000-0002-6389-0712  
Data Curator : Moing, Annick (INRAE) - ORCID: 0000-0003-1144-3600  
Data Curator : Jacob, Daniel (INRAE) - ORCID: 0000-0002-6687-7169  
Project Member : Cahoreau, Edern (INRAE) - ORCID: 0000-0001-8637-0448  
Project Member : Da Costa, Grégory (Univ. Bordeaux) - ORCID: 0000-0002-0336-5828  
Project Member : le Mao, Inès (Univ. Bordeaux)



Thus, entering (by autocompletion) just the name of a person will allow the ORCID number, email address and institutional assignment to be associated when **distributing metadata in Dataverse** for example.

## “Data Fragmentation”



Data Hub

Links to Resources

Resource Type: ---

Media Type: ---

Description: ---

Location: ---

Add a resource ?

Dataset

Dissertation

Event

Image

InteractiveResource

Journal

JournalArticle

Model

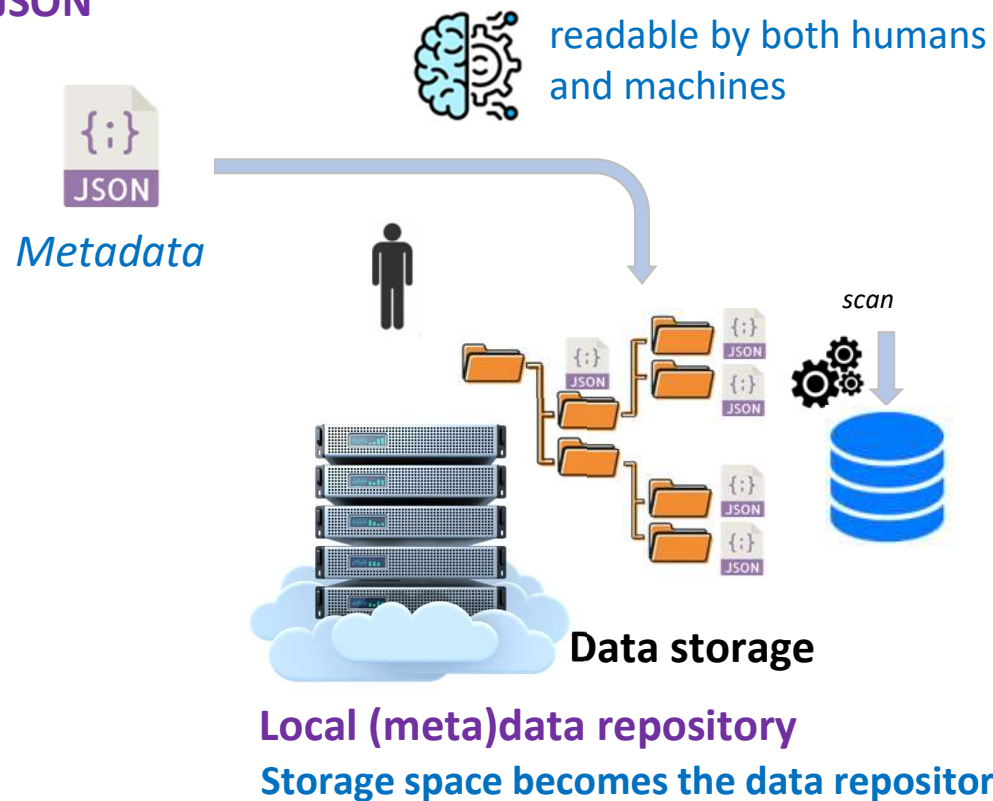


## RESOURCES

Type	Media	Description	Location
JournalArticle		Journal of Experimental Botany, Oxford University Press, 2020	<a href="http://doi.org/10.1093/jxb/eraa302">http://doi.org/10.1093/jxb/eraa302</a>
Collection		ODAM Experimental data tables	<a href="https://pmb-bordeaux.fr/dataexplorer/?dc=Frimouss">https://pmb-bordeaux.fr/dataexplorer/?dc=Frimouss</a>
Report	<a href="#">application/pdf</a>	Fruit Growth Modelling	<a href="https://pmb-bordeaux.fr/getdata/pdf/Frimouss/FruitGrowthModelling.pdf">https://pmb-bordeaux.fr/getdata/pdf/Frimouss/FruitGrowthModelling.pdf</a>
Software		Growth modeling applied to several fruit species	<a href="https://github.com/djacob65/growthmodel">https://github.com/djacob65/growthmodel</a>

We can also define **external resources** (URL links) relating to documents, publications or other related data. Maggot thus becomes a hub for your datasets connecting different resources, local and external.

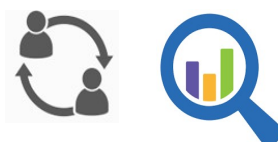
As output we produce  
a file in the format  
**JSON**



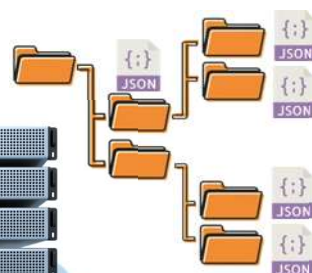
**Infrastructure**  
Local, Remote or Mixte



## Share & Search (meta)data



Metadata



Data storage

Local (meta)data repository

Storage space becomes the data repository

### ▼ DESCRIPTORS

Kind of Data

☐ Audiovisual ☐ Collection

☒ Dataset ☐ Event ☐ Image

☐ Interactive Resource ☐ Model

☐ Other ☐ Physical Object

☐ Service ☐ Software ☐ Sound

☐ Text ☐ Workflow

### Keywords

Search a value:

enter the first letters

### Topic Classification

Search a value:

enter the first letters

### Data origin

enter the first letters

Search

Empty the form

( KindOfData => Dataset )

Short name	Full title	Status of the dataset	Access rights to data	Metadata
AmaizingEnzymes	Leaf enzyme activities and total proteins of maize hybrids cultivated in the field	Processed	Private	
AmaizingNMR	NMR metabolomic and starch data of young leaf of maize hybrids cultivated in the field with normal sowing in 2013	Processed	Private	
Atacama	Atacama	Processed	Public	
Frimouss	FRUIT Interactive MOdelling for a Unified Selection System	Processed	Public	

--- Metadata export ---

### DESCRIPTION

Predictive metabolomics performed on 24 extremophile plant species in 19 different sites along an elevation transect in the Atacama Desert. (2021-06-01)

### DEFINITION

Full title	Subject
Atacama	Earth and Environmental Sciences

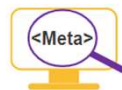
### STATUS

Status of the dataset	Access rights to data	Language	Life cycle step
Processed	Public	English	<ul style="list-style-type: none"> <li>Original release</li> <li>Deposit</li> </ul>

### MANAGEMENT


Contacts	Authors	Data curators	Project members
Dussarrat Thomas	<ul style="list-style-type: none"><li>Dussarrat Thomas</li><li>Gibon Yves</li><li>Gutierrez Rodrigo</li><li>Petriaq Pierre</li></ul>	Jacob Daniel	Cassan Cédric

Project leader	WP leader	Depositor	Producer	Grant Information
Gutierrez Rodrigo	Gibon Yves	Jacob Daniel	<ul style="list-style-type: none"><li>Bordeaux Metabolome</li><li>Plant Systems Biology Lab</li></ul>	<ul style="list-style-type: none"><li>MetaboHub</li><li>Phenome</li></ul>

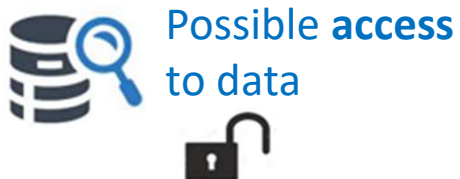


## Maggot Data Browser

Username  
Password

☐ Je ne suis pas un robot   
reCAPTCHA  
Confidentialité - Conditions

Login



Possible access  
to data



JSON

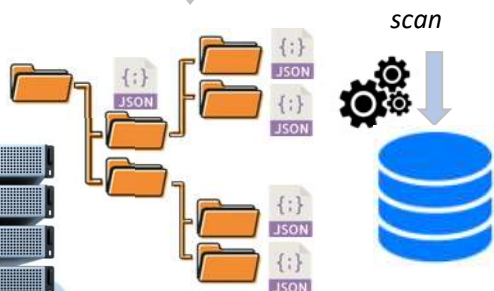
Metadata



Data storage

Local (meta)data repository

Storage space becomes the data repository



Search...



My files

Atacama

New folder

New file

Settings

Logout

18.9 GiB of 44 GiB used

File Browser 2.27.0  
Help

Name ↑

Size

Last modified...

images

—

3 years ago

pdf

—

2 years ago

a\_attributes.tsv

686.08 KiB

2 years ago

Biochemical.txt

20.43 KiB

2 years ago

Chemical.txt

60.46 KiB

3 years ago

Experiment.txt

36 B

3 years ago

infos.md

3.77 KiB

2 years ago

META\_Atacama.json

2.01 KiB

5 months ago

Metabolic\_data\_Neg.txt

5.62 MiB

3 years ago

Metabolic\_data\_Pos.txt

3.39 MiB

3 years ago

Metabolic\_data\_QI.txt

220.22 KiB

3 years ago

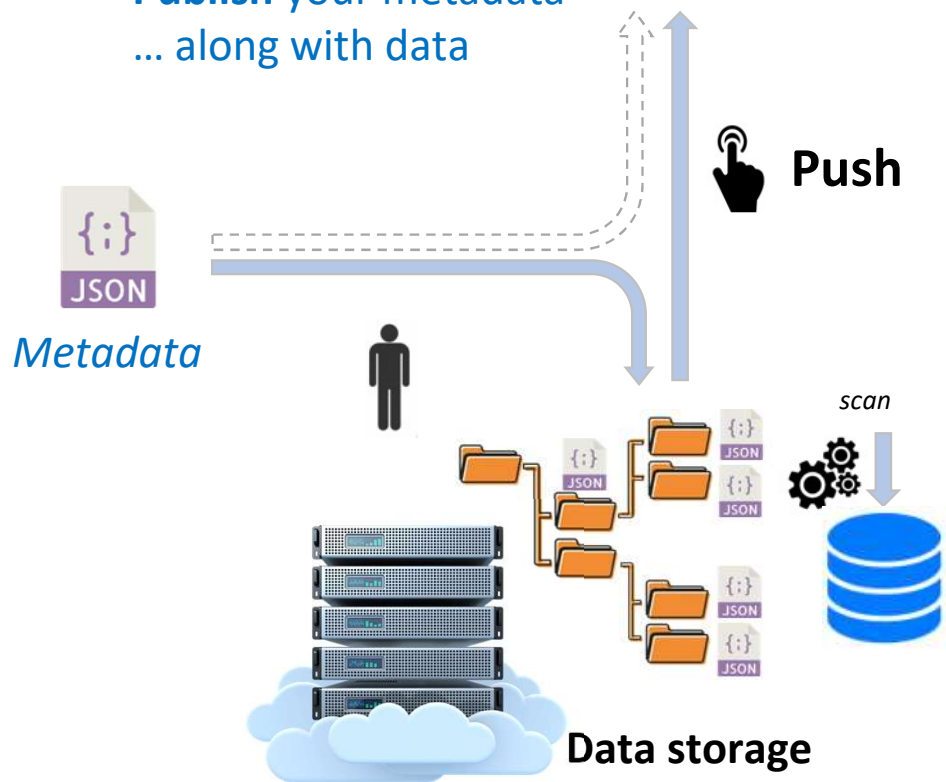
## Institutional data repositories



⇒ Have the privileges to do so  
(creation/modification rights).



**Publish your metadata  
... along with data**

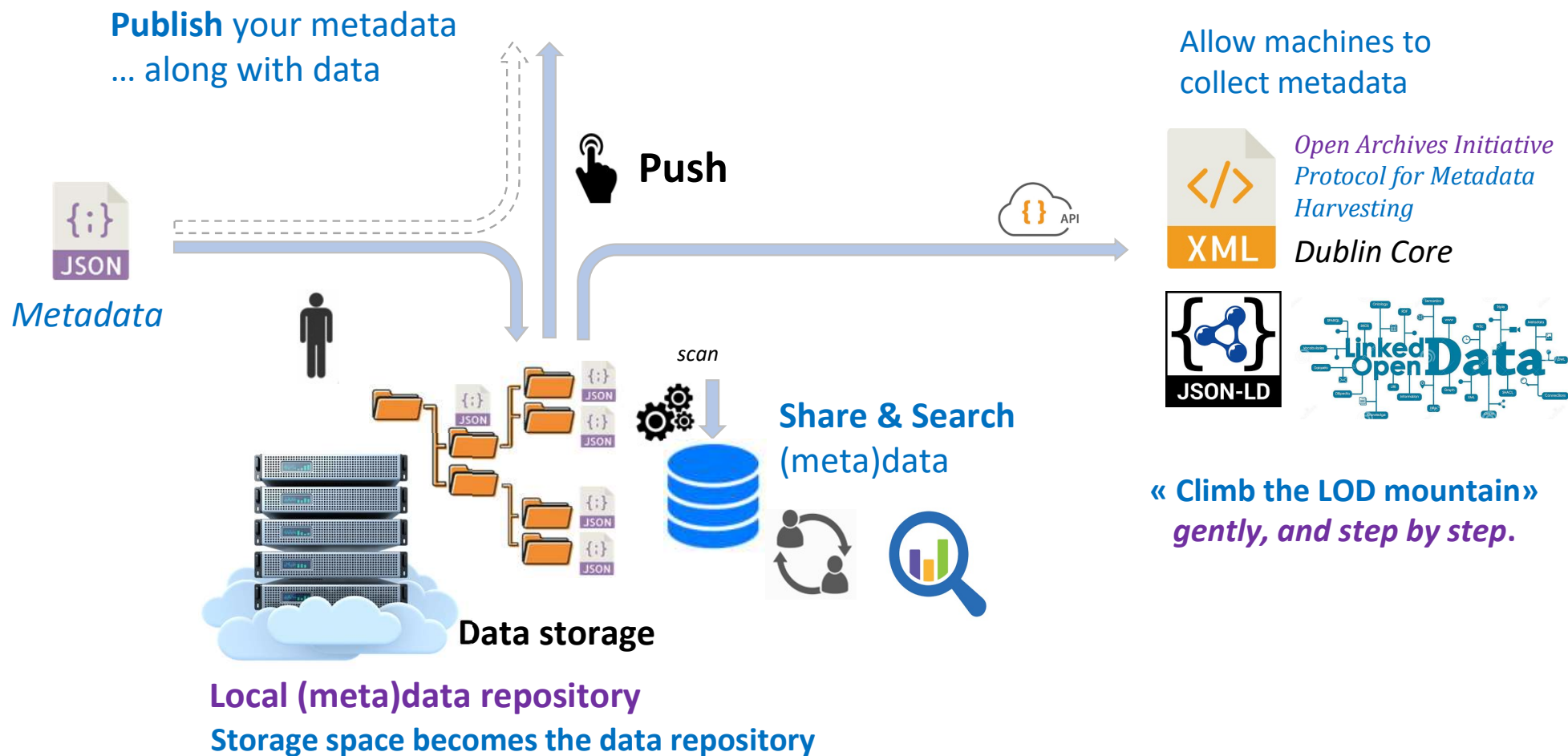


**Local (meta)data repository**

**Storage space becomes the data repository**

Citation Metadata	
Dataset Persistent ID	doi:10.82233/FK2/3MHX6
Title	FRIM - Fruit Integrative Modelling
Other ID	Maggot: frim1
Contact	Use email button above to contact.
Author	Gibon Yves (INRAE)
Contributor	Bénard Camille (INRAE) Biais Benoit (INRAE) Beauvoit Bertrand (Univ. Bordeaux) - ORCID: 0000-0002-7666-6429 Colombié Sophie (INRAE)
Producer	Bordeaux Metabolome (INRAE) <a href="https://metabolome.cgib.u-bordeaux.fr/">https://metabolome.cgib.u-bordeaux.fr/</a>
Language	English
Subject	Computer and Information Science, Medicine, Health and Life Sciences
Keyword	tomato <a href="http://purl.obolibrary.org/obo/NCBITaxon_4081">http://purl.obolibrary.org/obo/NCBITaxon_4081</a> (EFO) fruit growth <a href="http://purl.obolibrary.org/obo/PO_0009001">http://purl.obolibrary.org/obo/PO_0009001</a> (EFO) experimental measurement <a href="http://www.ebi.ac.uk/efo/EFO_0001444">http://www.ebi.ac.uk/efo/EFO_0001444</a> (EFO)
Topic Classification	fruit growth (thesaurus-inrae) <a href="http://opendata.inrae.fr/thesaurusINRAE/c_7655">http://opendata.inrae.fr/thesaurusINRAE/c_7655</a> plant health (thesaurus-inrae) <a href="http://opendata.inrae.fr/thesaurusINRAE/c_11833">http://opendata.inrae.fr/thesaurusINRAE/c_11833</a> omics (thesaurus-inrae) <a href="http://opendata.inrae.fr/thesaurusINRAE/c_e3728dc6">http://opendata.inrae.fr/thesaurusINRAE/c_e3728dc6</a> computer analysis (thesaurus-inrae) <a href="http://opendata.inrae.fr/thesaurusINRAE/c_16182">http://opendata.inrae.fr/thesaurusINRAE/c_16182</a>
Data Type	Dataset
Data Origin	experimental data
Life cycle step	Study design; Data collection
Related Publication	Biais B, Bénard C, Beauvoit B, Colombié S, Prodhomme D, Ménard G, Bernillon S, Gehl B, Gautier H, Ballias P, Mazat J-P, Sweetlove L, Génard M, Gibon Y. 2014. Remarkable reproducibility of enzyme activity profiles in tomato fruits grown under contrasting environments provides a roadmap for studies of fruit metabolism. Plant Physiology 164, 1204-1221 doi: 10.1104/pp.113.231241 <a href="https://doi.org/10.1104/pp.113.231241">https://doi.org/10.1104/pp.113.231241</a>
Other Reference	Experimental data tables: ODAM dataexplorer, <a href="https://pmb-bordeaux.fr/dataexplorer/?ds=frim1">https://pmb-bordeaux.fr/dataexplorer/?ds=frim1</a> , Article: Beauvoit et al (2014) Plant Cell 26: 3224-3242, <a href="https://doi.org/10.1105/tpc.114.127761">https://doi.org/10.1105/tpc.114.127761</a> ; Article: Bénard et al (2015) Journal of Experimental Botany Vol. 66, No. 11 pp. 3391-3404, <a href="https://doi.org/10.1093/jxb/erv151">https://doi.org/10.1093/jxb/erv151</a> ; Article: Colombié et al (2015) Plant Physiology 180, 1709-1724, <a href="https://doi.org/10.1104/pp.19.00086">https://doi.org/10.1104/pp.19.00086</a>
Funding Information	ANR: ANR-11-INBS-0010 ANR: ANR-11-INBS-0012
Depositor	Jacob Daniel
Deposit Date	2023-04-04





## Metadata management (local repository)



## Metadata crosswalks (remote repository and/or metadata harvesting)



## What to describe & how to describe it



Metadata Schema 



Controlled Vocabulary

AgroPortal, BioPortal,  
VOINRAE, LOTERRE, ONTOSTACK, ...

## Repositories & export formats



...

## The Maggot tool allows a collective to :

- **Have visibility** of what is produced within the collective
  - datasets, software, databases, images, sounds, videos, analyses, codes, ...
  - $\Rightarrow$  **share metadata**
- **Raise awareness** among newcomers and students about a better description of what they produce
  - Limit data loss (after temporary staff leave)
- **Promote FAIR** within the collective
  - particularly as part of a quality approach



<https://pmb-bordeaux.fr/maggot/>



<https://inrae.github.io/pgd-mmdt/>



## Meet Open Data requirements

This is not necessarily making data to open access without conditions, but rather

1. **provide access to metadata** defining the conditions of access and use of data,
2. **open the data beyond itself**, i.e. that they can be interoperable.

So the data must be **as FAIR as possible**, ...  
...even if it is not possible to make them open.