Data management
based on Metadata

<Meta> Maggot

Metadata aggregation on data storage

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

Daniel Jacob - Maggot – INRAE

Data management

Being able to easily describe your data (descriptive metadata)

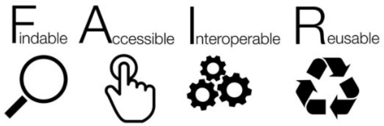
- with its professional vocabulary
- without tedious entries
- without having to re-enter the same information each time
- by associating external resources (links)

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

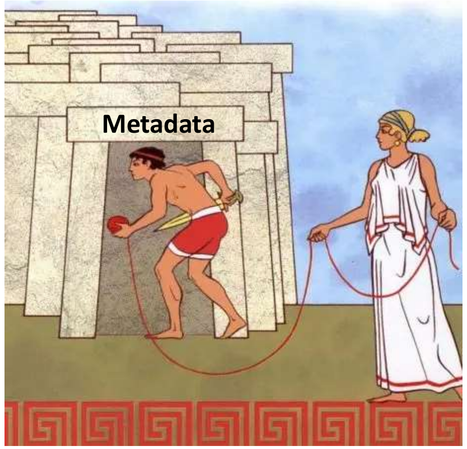
The Ariadne's thread …

Organize
Your workspace
(storage, backup, naming, …)

Share & Search data/metadata
(metadata)

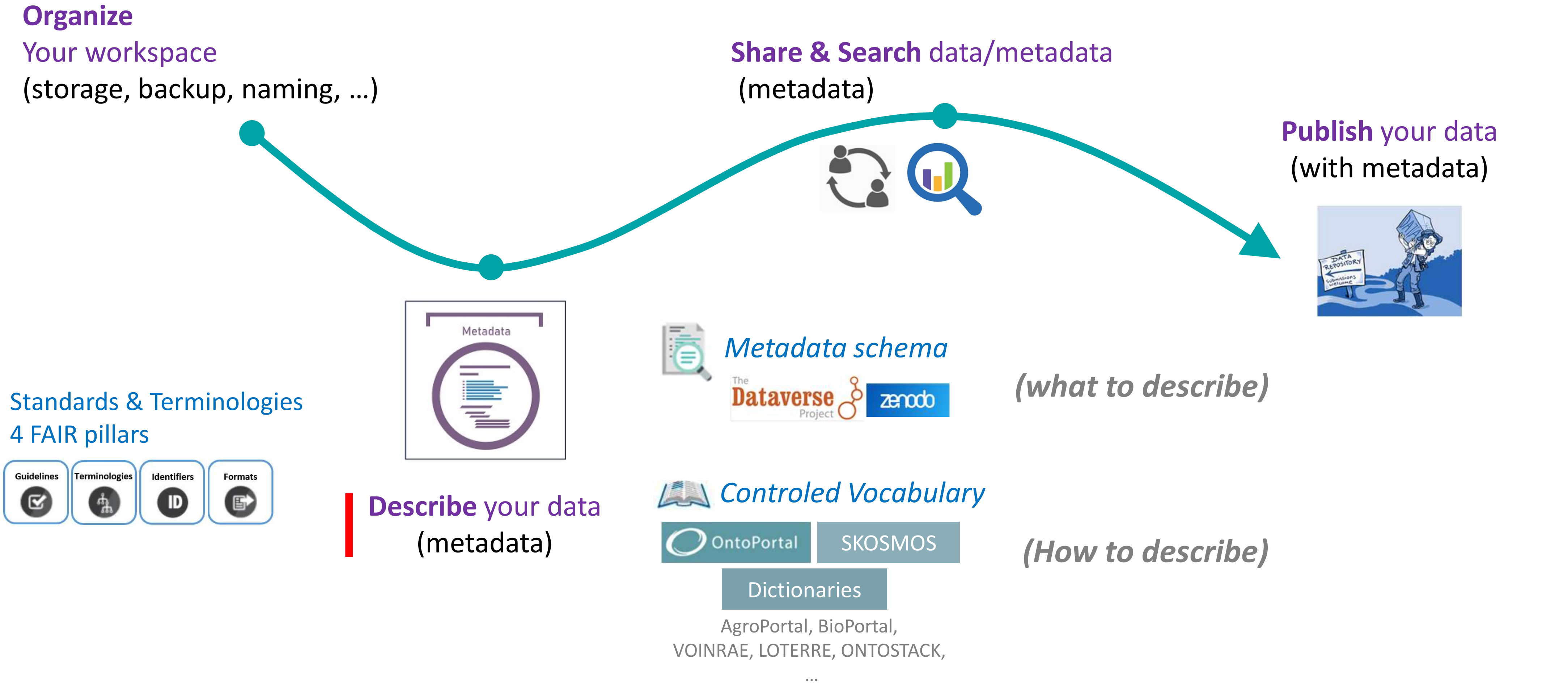
Publish your data
(with metadata)

Describe your data
(metadata)



Data management

The Ariadne's thread …



Organize
Your workspace
(storage, backup, naming, …)

Describe your data
(metadata)

Share & Search data/metadata
(metadata)

Publish your data
(with metadata)

Standards & Terminologies
4 FAIR pillars

Metadata schema

(what to describe)

Controled Vocabulary

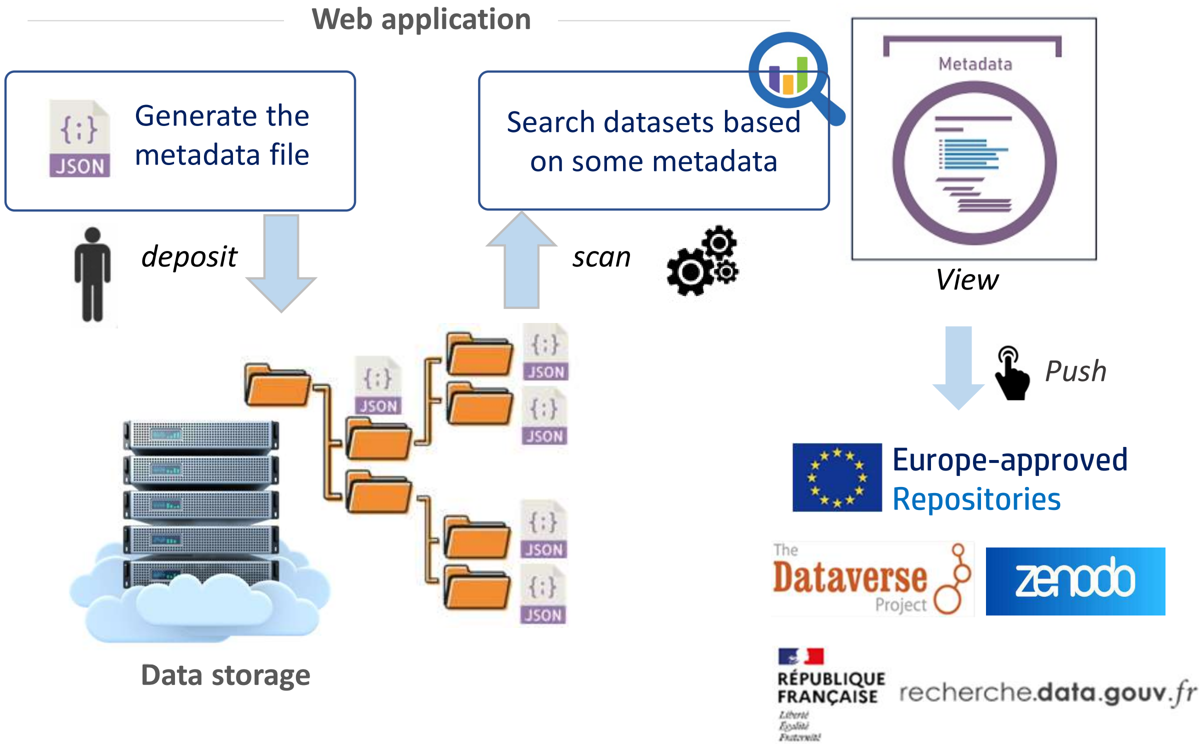
AgroPortal, BioPortal,
VOINRAE, LOTERRE, ONTOSTACK,
…

(How to describe)

Daniel Jacob - Maggot – INRAE

Maggot
Metadata aggregation on data storage

An ecosystem for metadata management


Web application

Generate the metadata file

Search datasets based on some metadata

deposit

scan

View

Push

Data storage

Europe-approved Repositories

Describe metadata
with controled vocabulary

Share & Search
metadata / data

Publish metadata
… with data

Maggot

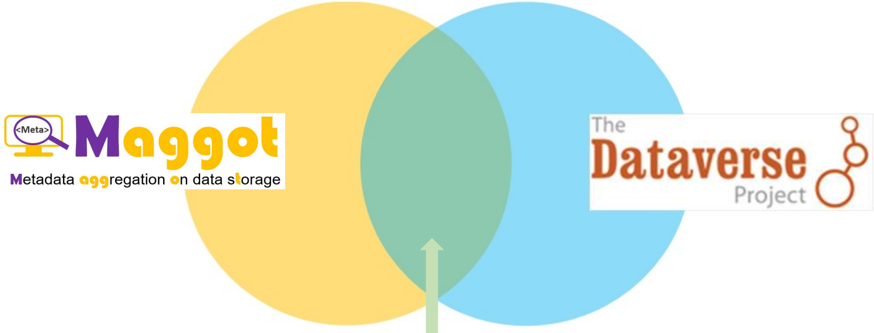
Metadata aggregation on data storage

Describe metadata

Guidelines

Metadata Schema

Choose a schema common to « Dataverse »



Metadata schema (standard)

DDI®

Data Document Initiative

DEFINITION
MANAGEMENT
STATUS
DESCRIPTORS
OTHER
…

Common Metadata

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

Life Sciences Metadata
Geospatial Metadata
Journal Metadata
Semantic resource
…

Specific Profile Metadata

…

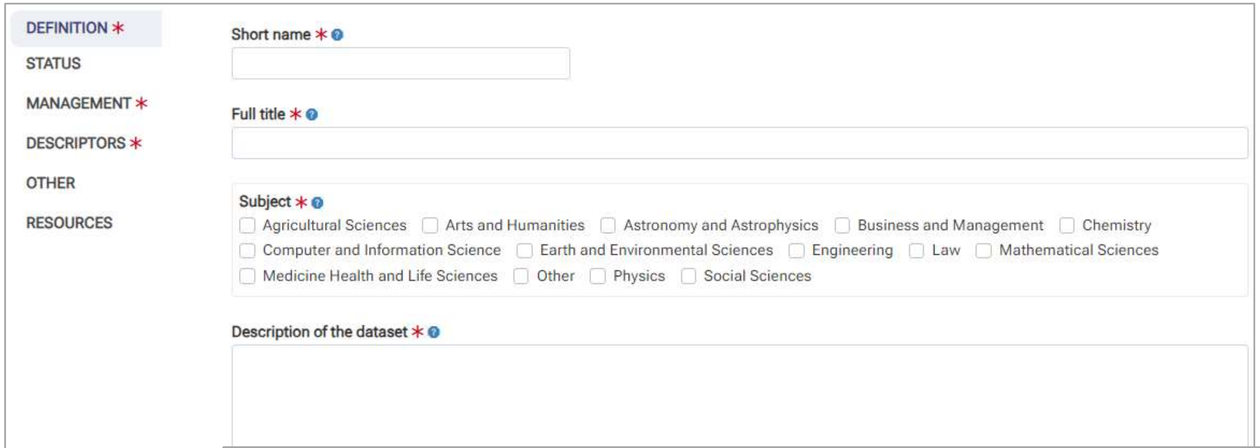
Maggot

Metadata aggregation on data storage

Describe metadata

Guidelines

Metadata Schema



DEFINITION *

STATUS

MANAGEMENT *

DESCRIPTORS *

OTHER

RESOURCES

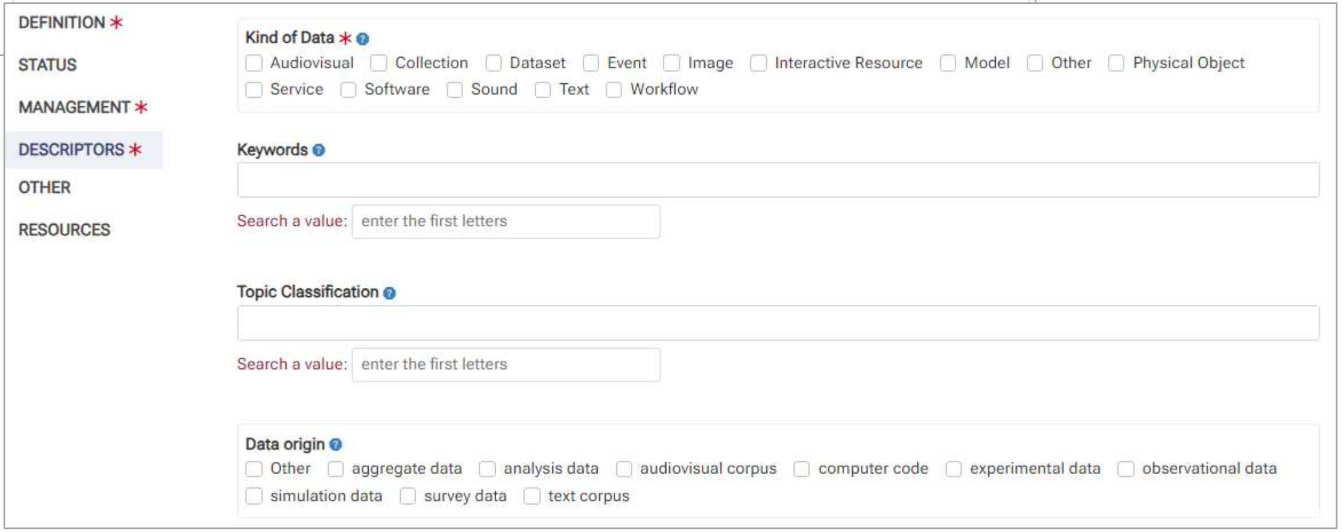
Short name *

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 *



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

* *mandatory fields*

Schema common to « Dataverse »

Common Metadata

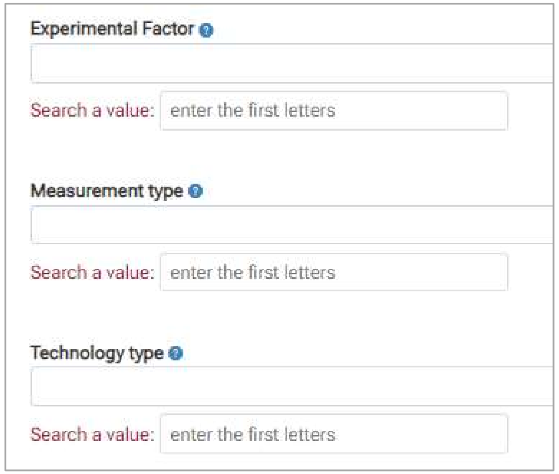
title, description, alternativeURL, contacts, authors, collectors, curators, members, depositor, leader, subject, keywords, topics, kindOfData, dataOrigin, lifeCycleStep, publication, otherReferences, grantNumbers, project, …

Mandatory fields

Recommended fields

Desirable fields

Specific Profile Metadata



Experimental Factor

Search a value: enter the first letters

Measurement type

Search a value: enter the first letters

Technology type

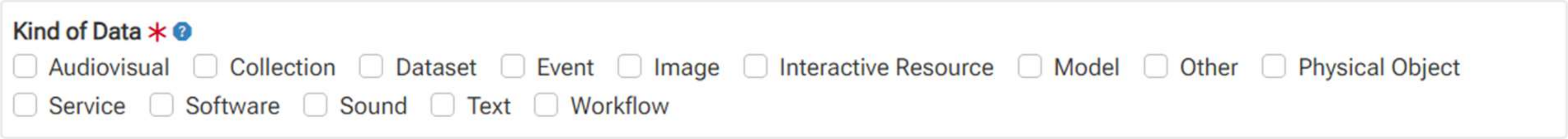
Search a value: enter the first letters

Maggot
Metadata aggregation on data storage

Controlled vocabulary

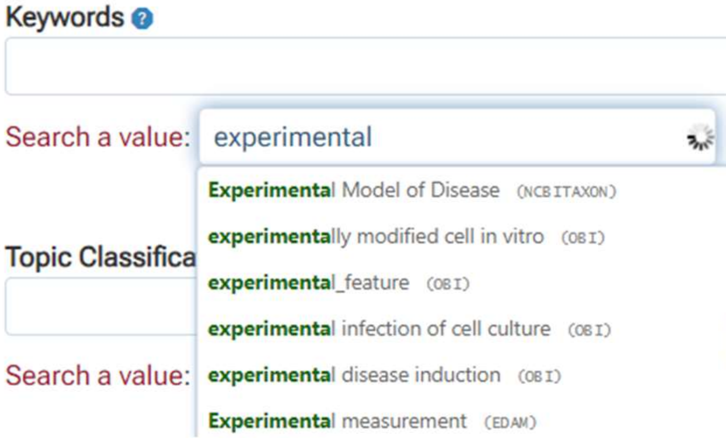
Terminologies
Controled Vocabulary

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: experimental

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

Topic Classifica

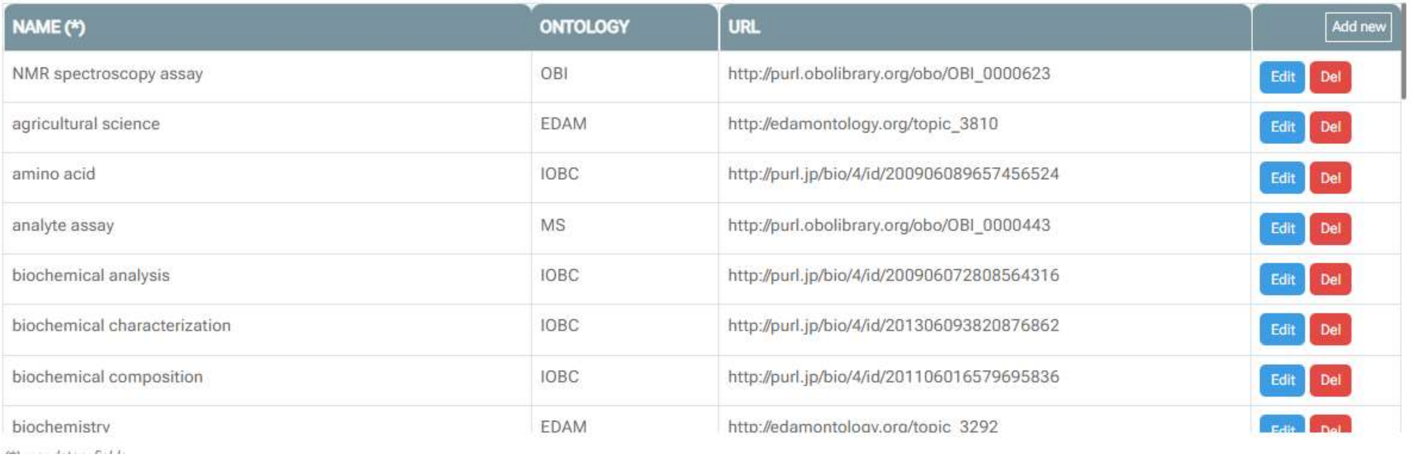
Search a value:

AgroPortal

BioPortal

List of ontologies to choose according to your domain

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



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
biochemistry	EDAM	http://edamontology.org/topic_3292	Edit Del

Thesaurus SKOSMOS

(VOINRAE, LOTERRE, ONTOSTACK, …)

V.O.INRAE
VOCABULAIRES OUVERTS

Building a vocabulary

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

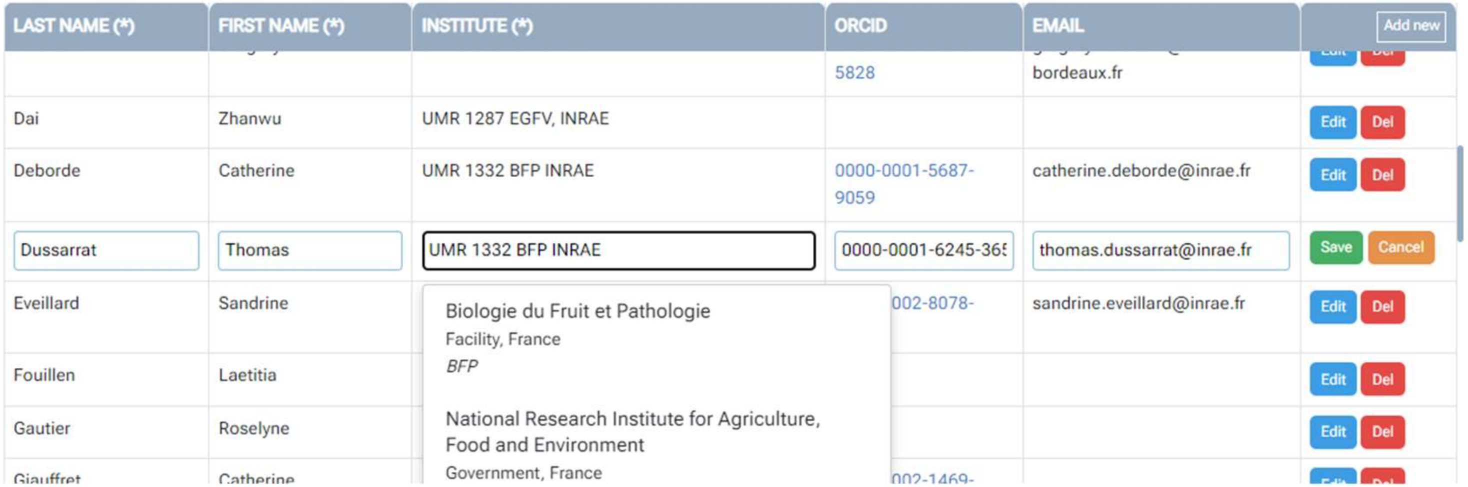
Maggot

Metadata aggregation on data storage

Dictionaries

people, grant, producer, vocabulary, …

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	Add new
			5828	bordeaux.fr	Edit Del
Dai	Zhanwu	UMR 1287 EGFV, INRAE			Edit Del
Deborde	Catherine	UMR 1332 BFP INRAE	0000-0001-5687-9059	catherine.deborde@inrae.fr	Edit Del
Dussarrat	Thomas	UMR 1332 BFP INRAE	0000-0001-6245-365	thomas.dussarrat@inrae.fr	Save Cancel
Eveillard	Sandrine		002-8078-	sandrine.eveillard@inrae.fr	Edit Del
Fouillen	Laetitia				Edit Del
Gautier	Roselyne				Edit Del
Giauffret	Catherine		002-1469-		Edit Del

Biologie du Fruit et Pathologie
Facility, France
BFP

National Research Institute for Agriculture, Food and Environment
Government, France

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

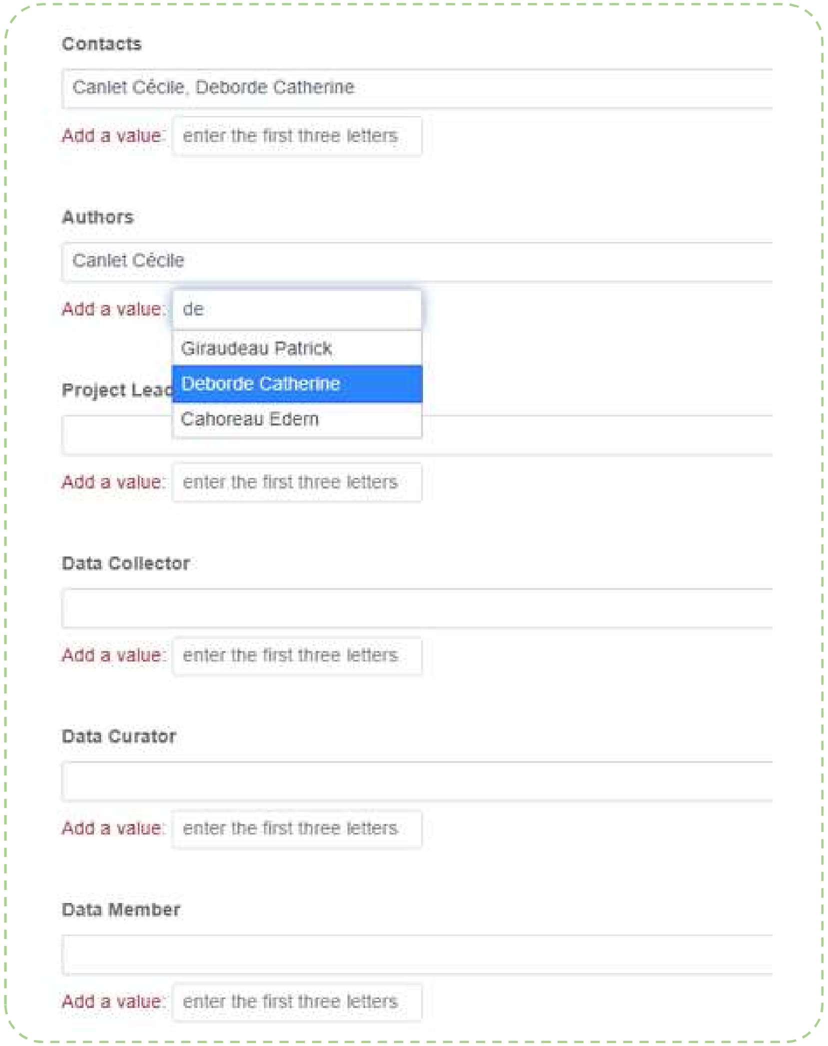
Its information, once entered and saved in a file called a dictionary, can be subsequently associated with the corresponding entity.

Maggot

Metadata aggregation on data storage

Dictionaries

Example : people



Contacts

Canlet Cécile, Deborde Catherine

Add a value: enter the first three letters

Authors

Canlet Cécile

Add a value: de

Giraudeau Patrick

Deborde Catherine

Cahoreau Edern

Project Lead

Add a value: enter the first three letters

Data Collector

Add a value: enter the first three letters

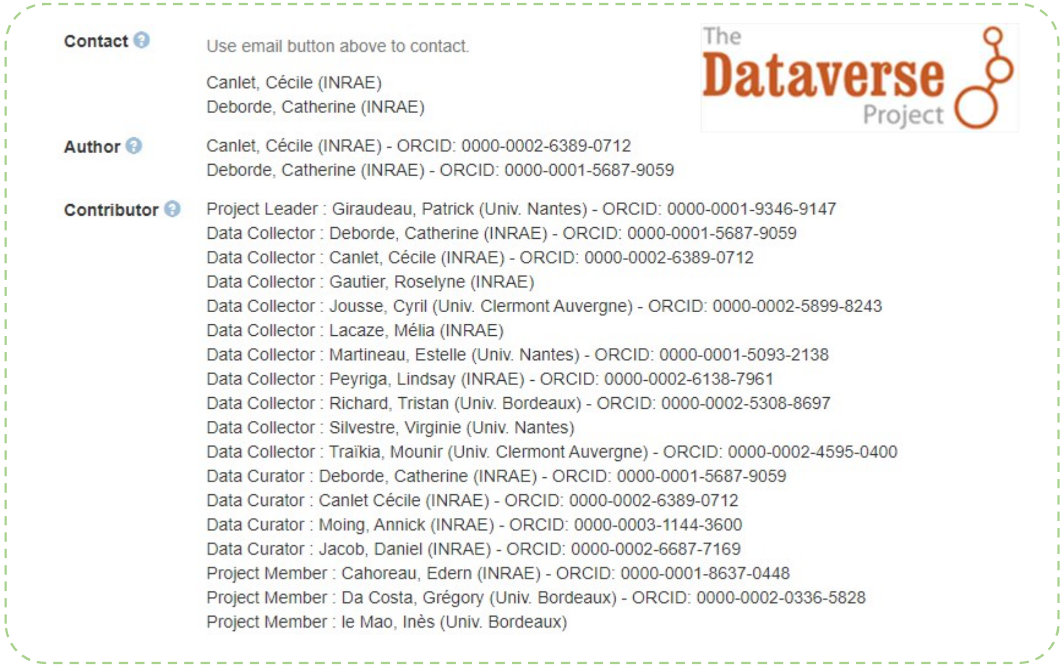
Data Curator

Add a value: enter the first three letters

Data Member

Add a value: enter the first three letters

Push



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, Mélia (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)

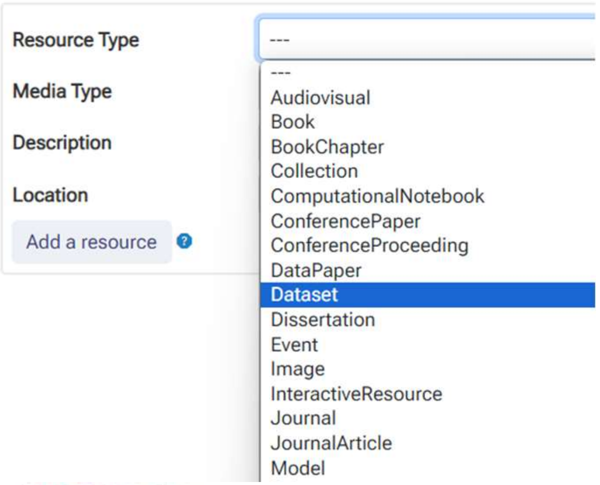
Thus, entering (by autocompletion) just the name of a person will allow the ORCID number, email address and institutional assignment to be associated when distributing metadata in Dataverse for example.

Maggot
Metadata aggregation on data storage

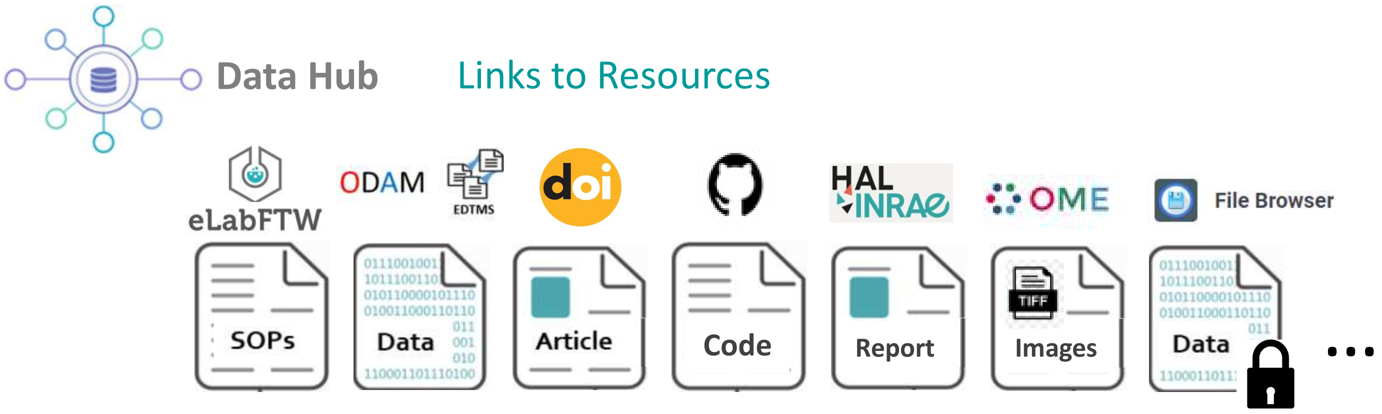
External Resources

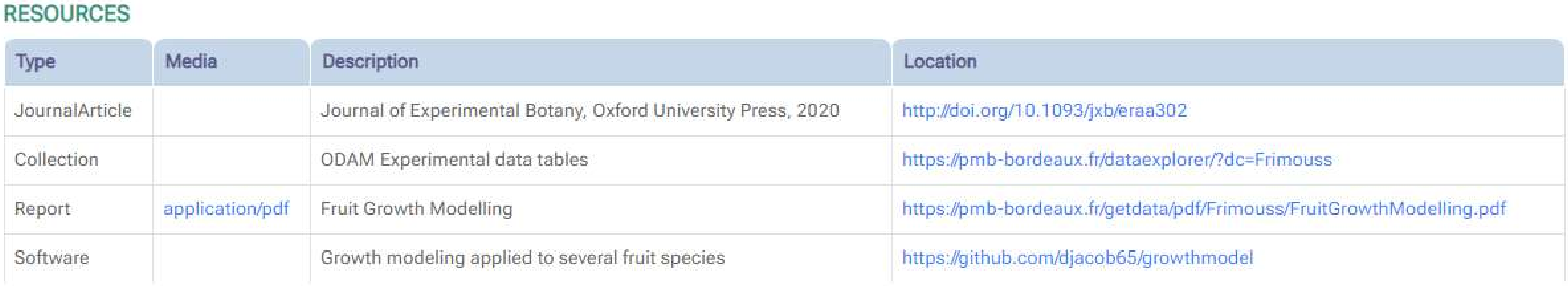
"Data Fragmentation"

Identifiers
Persistent and unique identifiers



Resource Type: ---
Media Type
Description
Location
Add a resource





RESOURCES

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

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.

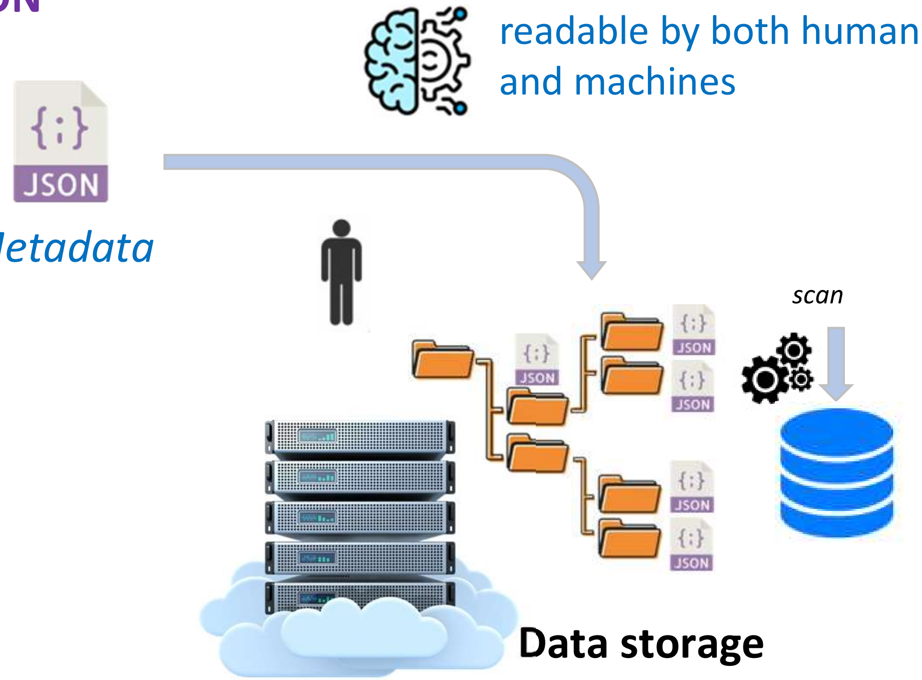
Maggot

Metadata aggregation on data storage

Formats

Format for interoperability purposes.

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



readable by both humans and machines

Local (meta)data repository

Storage space becomes the data repository

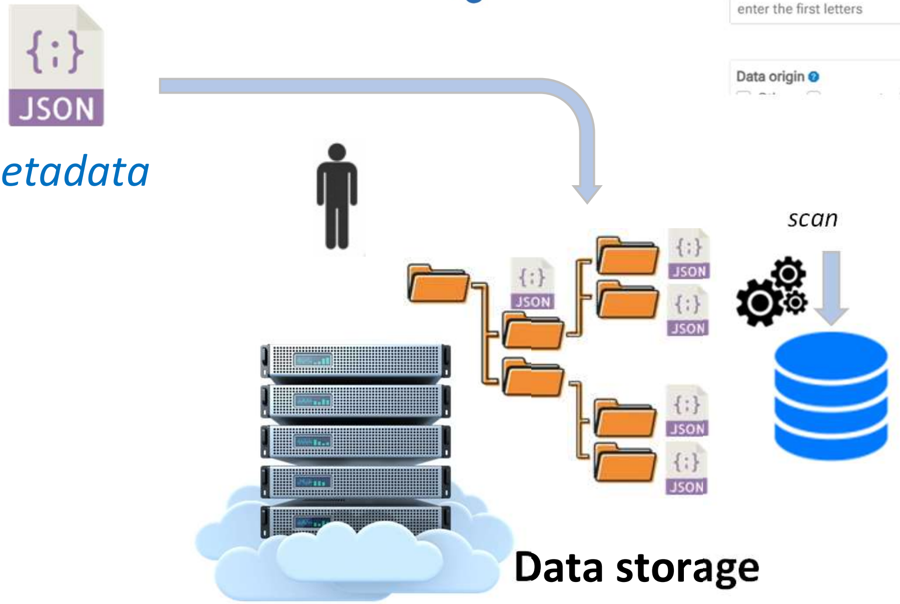
Infrastructure

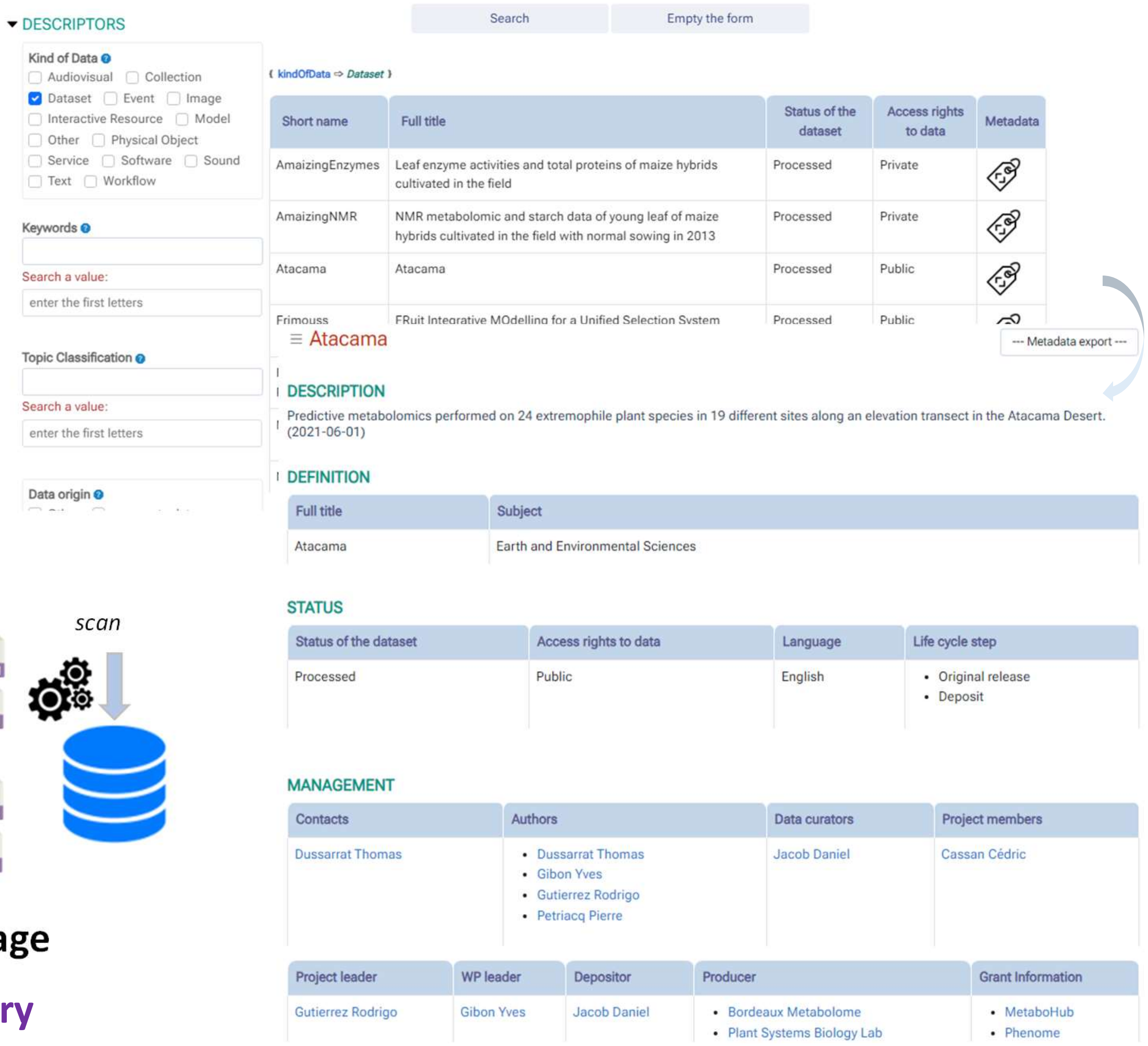
Local, Remote or Mixte

Maggot

Metadata aggregation on data storage

Share & Search (meta)data





▾ DESCRIPTORS

Search | Empty the form

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

(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 Integrative MOdelling for a Unified Selection System	Processed	Public	

≡ Atacama

--- 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	• Original release • Deposit

MANAGEMENT

Contacts	Authors	Data curators	Project members
Dussarrat Thomas	• Dussarrat Thomas • Gibon Yves • Gutierrez Rodrigo • Petriacq Pierre	Jacob Daniel	Cassan Cédric

Project leader	WP leader	Depositor	Producer	Grant Information
Gutierrez Rodrigo	Gibon Yves	Jacob Daniel	• Bordeaux Metabolome • Plant Systems Biology Lab	• MetaboHub • Phenome

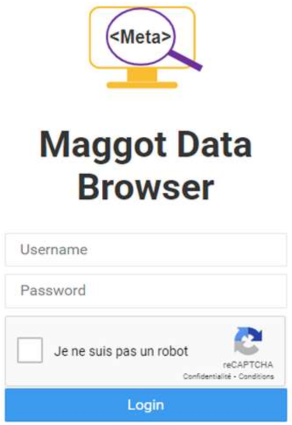
Local (meta)data repository

Storage space becomes the data repository

Maggot

Metadata aggregation on data storage

Possible access to data



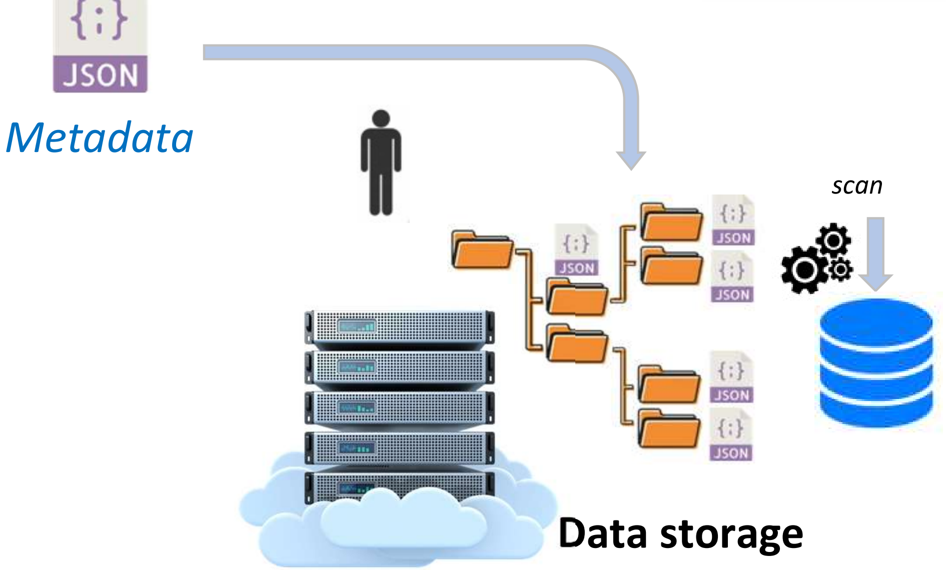
Maggot Data Browser

Username

Password

□ Je ne suis pas un robot

Login

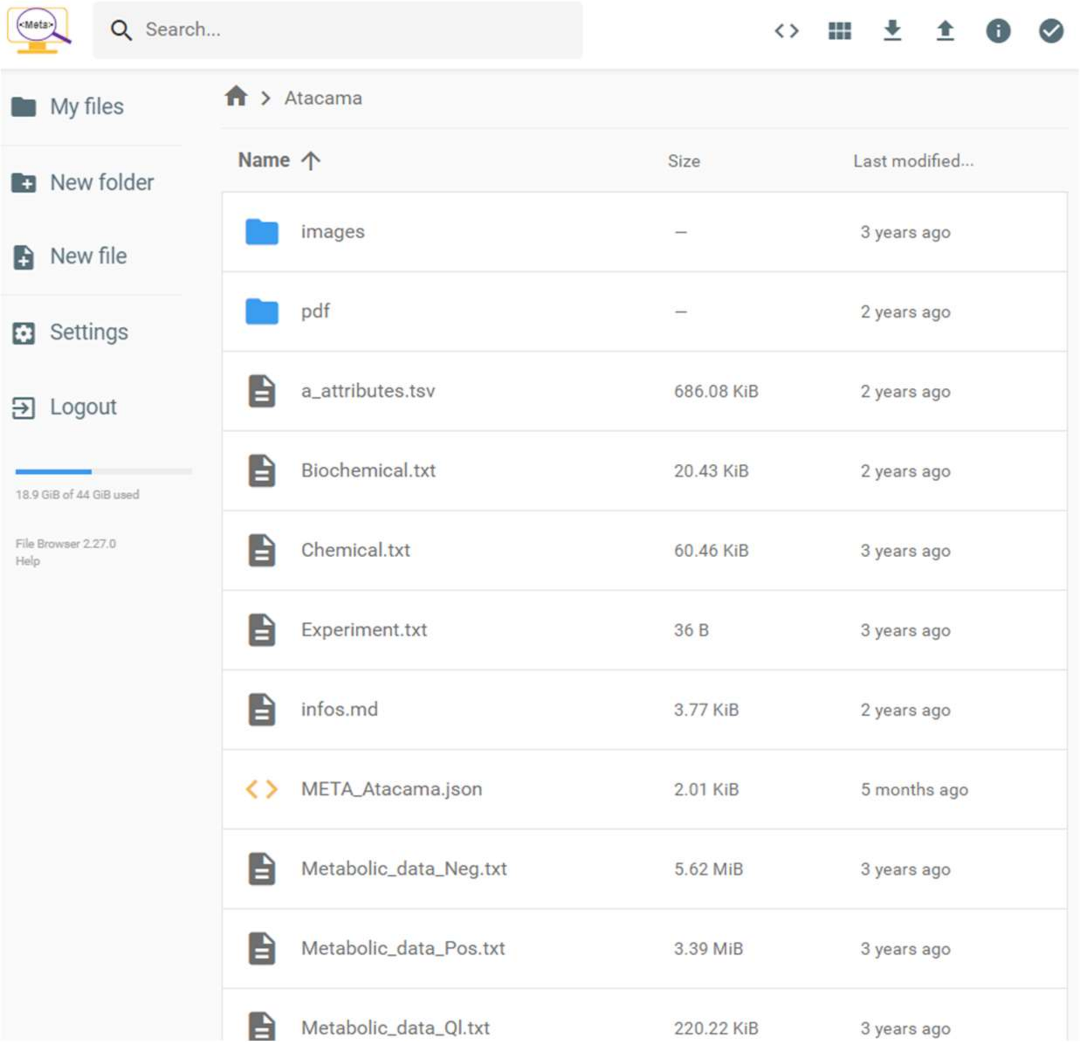


Metadata

Data storage

Local (meta)data repository

Storage space becomes the data repository



Search...

- My files
- New folder
- New file
- Settings
- Logout

18.9 GiB of 44 GiB used

File Browser 2.27.0
Help

Atacama

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

Maggot

Metadata aggregation on data storage

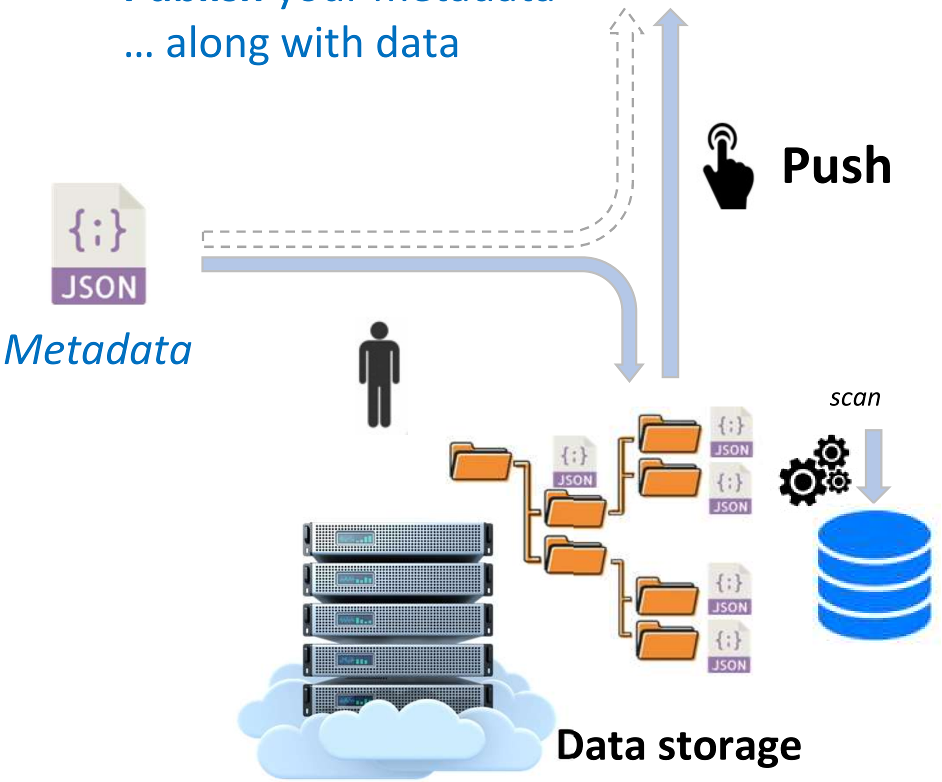
Institutional data repositories

The Dataverse Project / zenodo

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

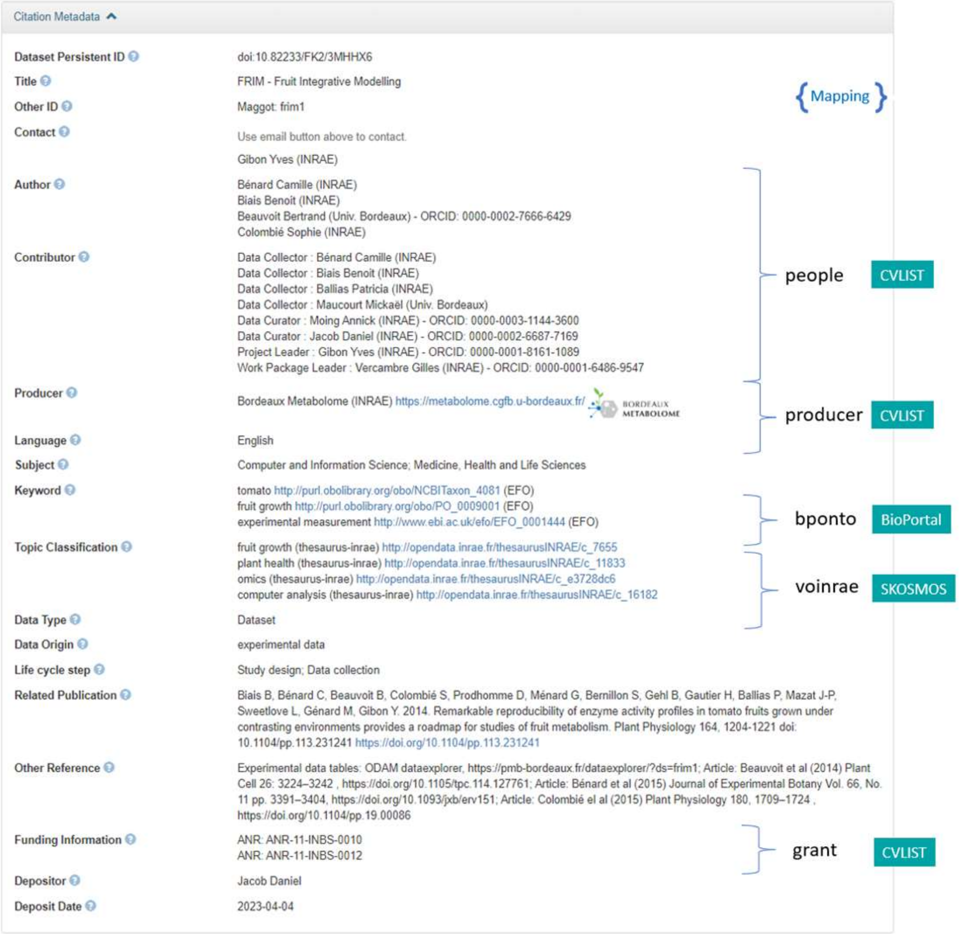
Publish your metadata

… along with data



Local (meta)data repository

Storage space becomes the data repository

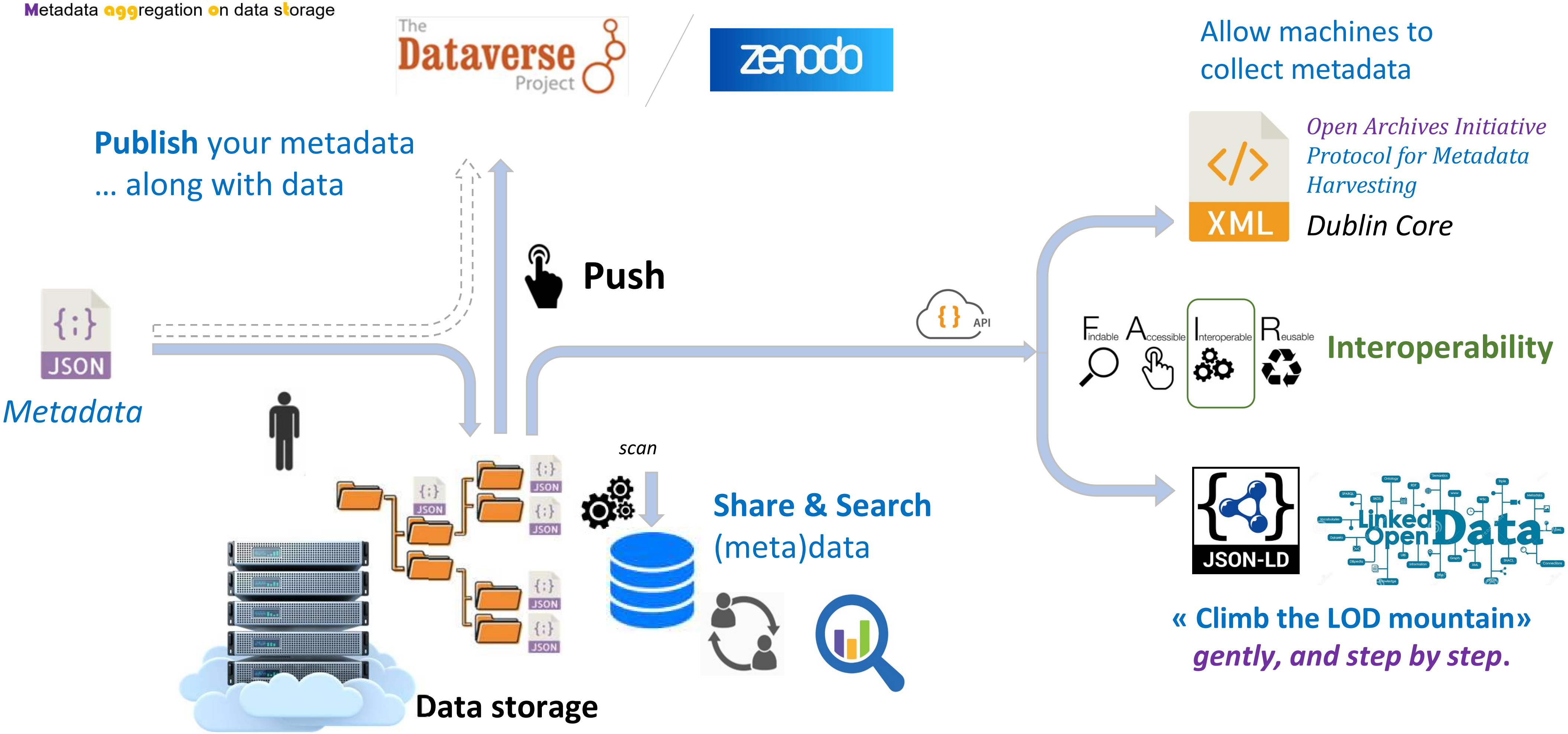

Citation Metadata

Dataset Persistent ID	doi:10.82233/FK2/3MHHX6
Title	FRIM - Fruit Integrative Modelling
Other ID	Maggot: frim1
Contact	Use email button above to contact. Gibon Yves (INRAE)
Author	Bénard Camille (INRAE) Biais Benoit (INRAE) Beauvoit Bertrand (Univ. Bordeaux) - ORCID: 0000-0002-7666-6429 Colombié Sophie (INRAE)
Contributor	Data Collector : Bénard Camille (INRAE) Data Collector : Biais Benoit (INRAE) Data Collector : Ballias Patricia (INRAE) Data Collector : Maucourt Mickaël (Univ. Bordeaux) Data Curator : Moing Annick (INRAE) - ORCID: 0000-0003-1144-3600 Data Curator : Jacob Daniel (INRAE) - ORCID: 0000-0002-6687-7169 Project Leader : Gibon Yves (INRAE) - ORCID: 0000-0001-8161-1089 Work Package Leader : Vercambre Gilles (INRAE) - ORCID: 0000-0001-6486-9547
Producer	Bordeaux Metabolome (INRAE) https://metabolome.cgfb.u-bordeaux.fr/ BORDEAUX METABOLOME
Language	English
Subject	Computer and Information Science; Medicine, Health and Life Sciences
Keyword	tomato http://purl.obolibrary.org/obo/NCBITaxon_4081 (EFO) fruit growth http://purl.obolibrary.org/obo/PO_0009001 (EFO) experimental measurement http://www.ebi.ac.uk/efo/EFO_0001444 (EFO)
Topic Classification	fruit growth (thesaurus-inrae) http://opendata.inrae.fr/thesaurusINRAE/c_7655 plant health (thesaurus-inrae) http://opendata.inrae.fr/thesaurusINRAE/c_11833 omics (thesaurus-inrae) http://opendata.inrae.fr/thesaurusINRAE/c_e3728dc6 computer analysis (thesaurus-inrae) http://opendata.inrae.fr/thesaurusINRAE/c_16182
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 https://doi.org/10.1104/pp.113.231241
Other Reference	Experimental data tables: ODAM dataexplorer, https://pmb-bordeaux.fr/dataexplorer/?ds=frim1; Article: Beauvoit et al (2014) Plant Cell 26: 3224–3242 , https://doi.org/10.1105/tpc.114.127761; Article: Bénard et al (2015) Journal of Experimental Botany Vol. 66, No. 11 pp. 3391–3404, https://doi.org/10.1093/jxb/erv151; Article: Colombié el al (2015) Plant Physiology 180, 1709–1724 , https://doi.org/10.1104/pp.19.00086
Funding Information	ANR: ANR-11-INBS-0010 ANR: ANR-11-INBS-0012
Depositor	Jacob Daniel
Deposit Date	2023-04-04

Maggot

Metadata aggregation on data storage

Institutional data repositories



Publish your metadata
… along with data

Push

Metadata

Allow machines to
collect metadata

Open Archives Initiative Protocol for Metadata Harvesting

Dublin Core

Interoperability

Share & Search
(meta)data

Data storage

« Climb the LOD mountain»
gently, and step by step.

Local (meta)data repository

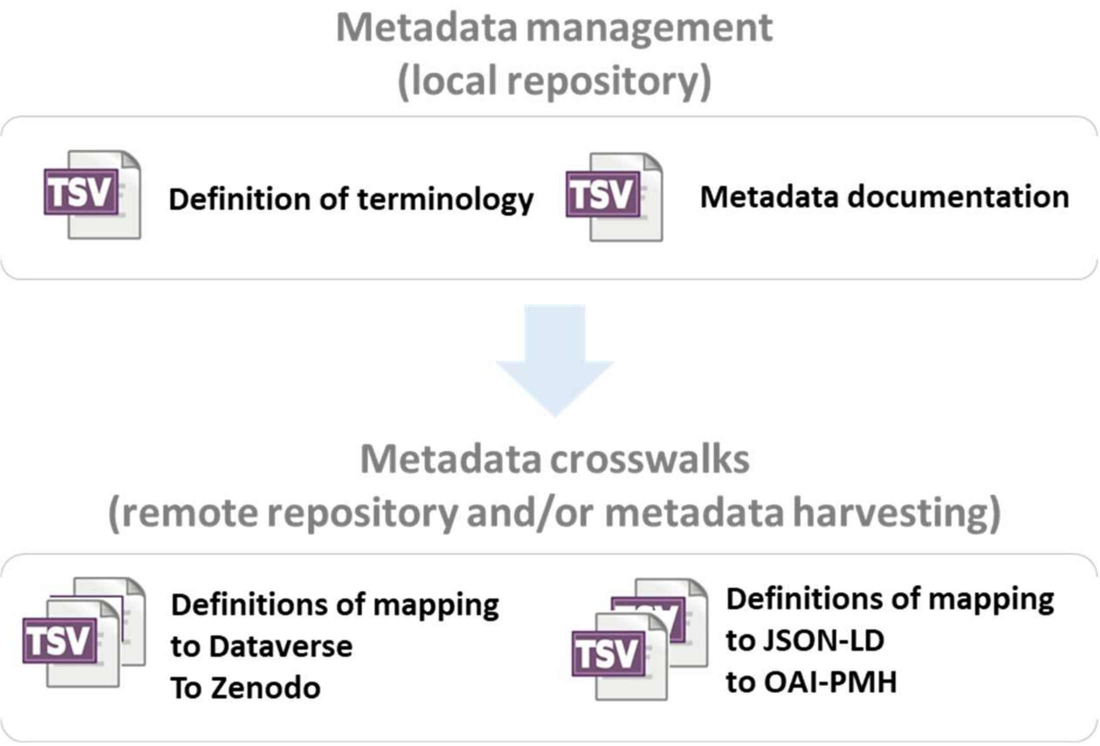
Storage space becomes the data repository

Daniel Jacob - Maggot – INRAE

Maggot

Metadata aggregation on data storage

Configuration flexibility



Metadata management
(local repository)

Definition of terminology

Metadata documentation

Metadata crosswalks
(remote repository and/or metadata harvesting)

Definitions of mapping
to Dataverse
To Zenodo

Definitions of mapping
to JSON-LD
to OAI-PMH

What to describe & how to describe it

Metadata Schema

Controled Vocabulary

AgroPortal, BioPortal,
VOINRAE, LOTERRE, ONTOSTACK, …

Repositories & export formats

…

Maggot

Metadata aggregation on data storage

The Maggot tool allows a collective to :

- **Have visibility** of what is produced within the collective
  - datasets, software, databases, images, sounds, videos, analyses, codes, ...
  - ⇒ **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.