Data management
based on Metadata

Daniel Jacob - Maggot - INRAE - January 2024

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 - January 2024

Data management

Being able to easily describe your data (descriptive metadata)

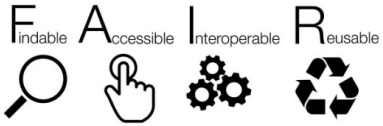
- with its professional vocabulary
- without tedious entries
- without having to re-enter the same information each time
- by associating external resources (links)

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

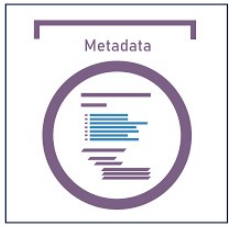
- Respect a standard (metadata schema)
- Use controlled vocabulary consistent with your domain (thesaurus, ontologies)
- Be at least “Findable, Accessible & Interoperable”



Data management

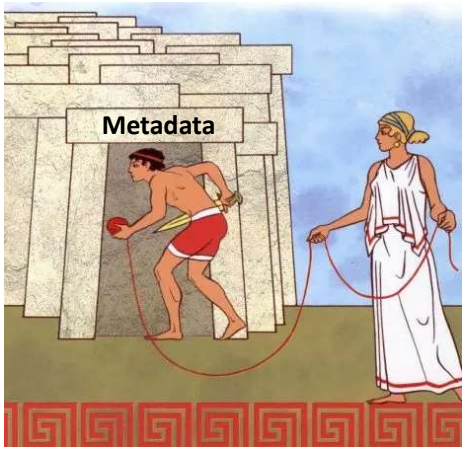
The Ariadne's thread …

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



Describe your data
(metadata)

Share & Search data/metadata
(metadata)



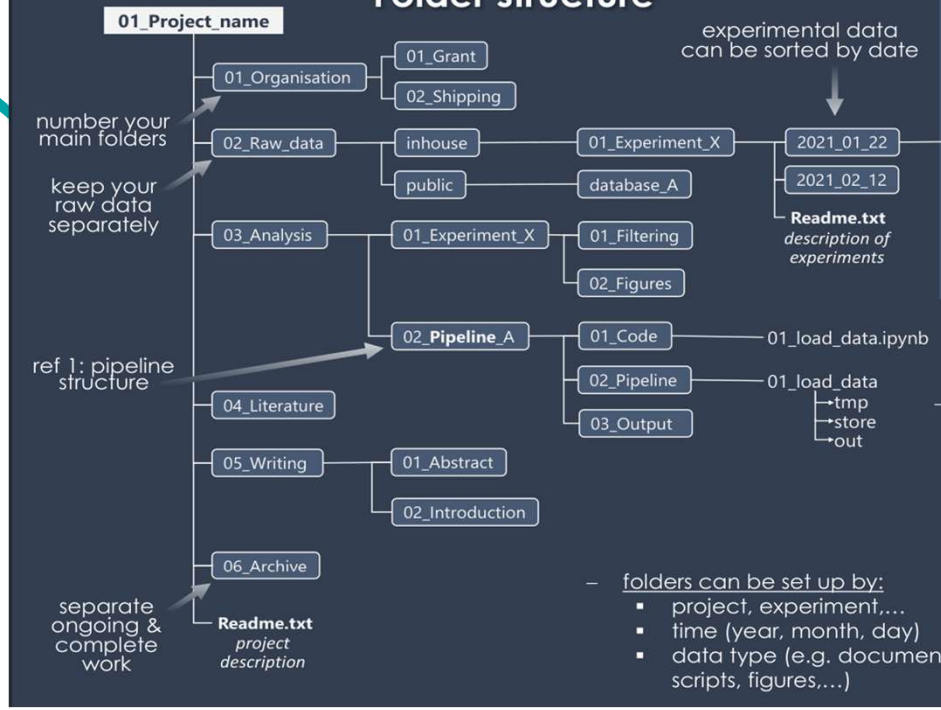
Publish your data
(with metadata)

Data management

The Ariadne's thread …

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

Folder structure



- folders can be set up by:
  - project, experiment,…
  - time (year, month, day)
  - data type (e.g. documents, scripts, figures,…)

- find the balance between a shallow & deep folder hierarchy
  - too deep → too many clicks might be needed to get to the right file
  - too shallow → too many files might end up in one folder (organise them in subfolders)

File naming

General naming tips for folders & files

- use unique, meaningful names
- not too long (not >30-40 characters)
- no spaces, dots, or special characters (£"$%!"&*^()+=[]:{}~@)
- hyphens (-) & underscores (_) to separate elements

Data management

The Ariadne's thread …

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

Describe your data
(metadata)

readme.txt files
can be used to describe projects, folders, and files

Share & Search (meta)data
(metadata)

Publish your data
(with metadata)

What to put in it ?
How to use them effectively when we want to find information ?
With what vocabulary ?

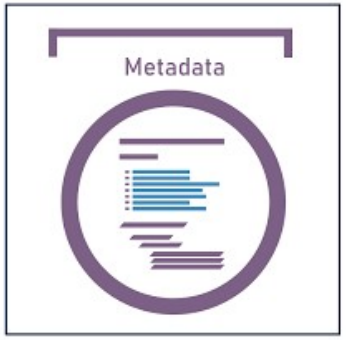
Metadata : descriptive data about the data, which provide the essential contextual information for interpreting and reusing the data.

Data management

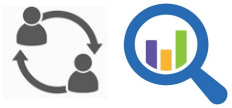
The Ariadne's thread …

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

Describe your data
(metadata)



Share & Search (meta)data
(metadata)



Publish your data
(with metadata)

Metadata Schema ***(what)***

+

Controlled Vocabulary ***(with what)***

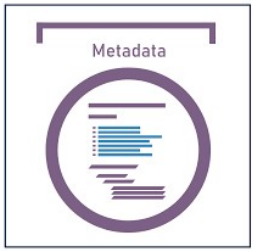
How ?

Metadata creation *(1/4)*

1 *Metadata schema (standard)*

(container)

⇔ Template / Metadata model

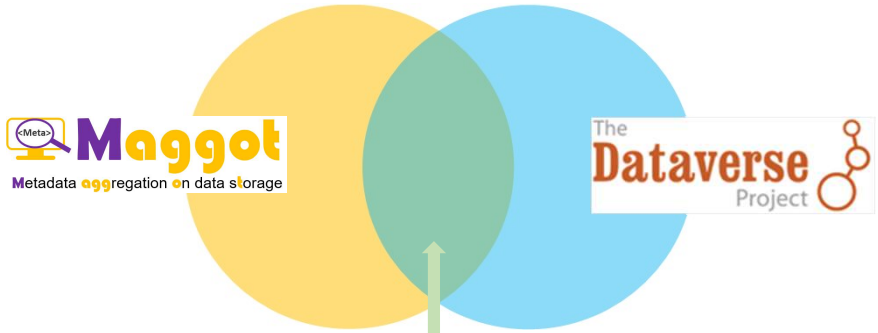


Describe your data
(metadata)

What metadata ?

What information ?

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



Metadata schema (standard)

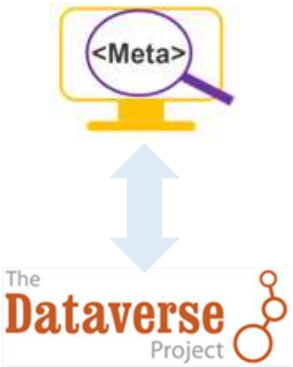
Data Document Initiative

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, …

Specific Profile Metadata
…

Life Sciences Metadata
Geospatial Metadata
Journal Metadata
Semantic resource
…



Common Metadata

Citation Metadata	
Dataset Persistent ID	doi:10.82233/FK2/1TMOX3
Title *	Test API sur un jeu de Métadonnées complètes
Link to data	http://pmb-bordeaux.fr/maggot/metadata/frim1
Other ID	Maggot: frim1
Contact *	Use email button above to contact. Bourafai-Aziez Asma (EVEAR-Extraction)
Author *	Bourafai-Aziez Asma (EVEAR-Extraction) - ORCID: 0000-0001-5974-7641 Deborde Catherine (INRAE) - ORCID: 0000-0001-8189-9238 Jacob Daniel (INRAE) - ORCID: 0000-0002-6687-7169
Producer *	Rousselot Guillaume (EVEAR-Extraction) INRAE
Description *	Development, validation, and use of 1H NMR spectroscopy for evaluating the quality of acerola-based food supplements and quantifying ascorbic acid English (2022-08-05)
Subject *	Medicine, Health and Life Sciences
Keyword *	Acerola ascorbic acid Metabolomic profiling vitamin C NMR
Topic Classification *	metabolic process (Plant Trait Ontology) http://purl.obolibrary.org/obo/GO_0008152 fruit (Plant Trait Ontology) http://purl.obolibrary.org/obo/PO_0009001
Kind of Data *	Dataset; Other
Other Kind of Data *	1H NMR spectra
Data Origin *	experimental data
Life cycle step *	Original release
Related Publication	Development, validation, and use of 1H NMR spectroscopy for evaluating the quality of acerola-based food supplements and quantifying ascorbic acid, Asma Bourafai-Aziez , Daniel Jacob, Gwladys Charpentier, Emmanuel Cassin, Guillaume Rousselot, Annick Moing and Catherine Deborde. Molecules. 2022; 27(17):5614 doi: 10.3390/molecules27175614 https://doi.org/10.3390/molecules27175614
Other References *	Experimental data tables: https://pmb-bordeaux.fr/dataexplorer/?ds=frim1; Simulation/Modelisation: https://hal.science/hal-02611223; Article: https://doi.org/10.1104/pp.113.231241
Grant Information *	ANR: ANR-11-INBS-0010
Project Information *	(INRAE-EVEAR-Extraction Research Project "Quantification de l'acide ascorbique dans les extraits d'acérola par 1D 1H-RMN" (2019-2020).)
Depositor	Jacob, Daniel
Deposit Date	2023-02-13

What metadata ?

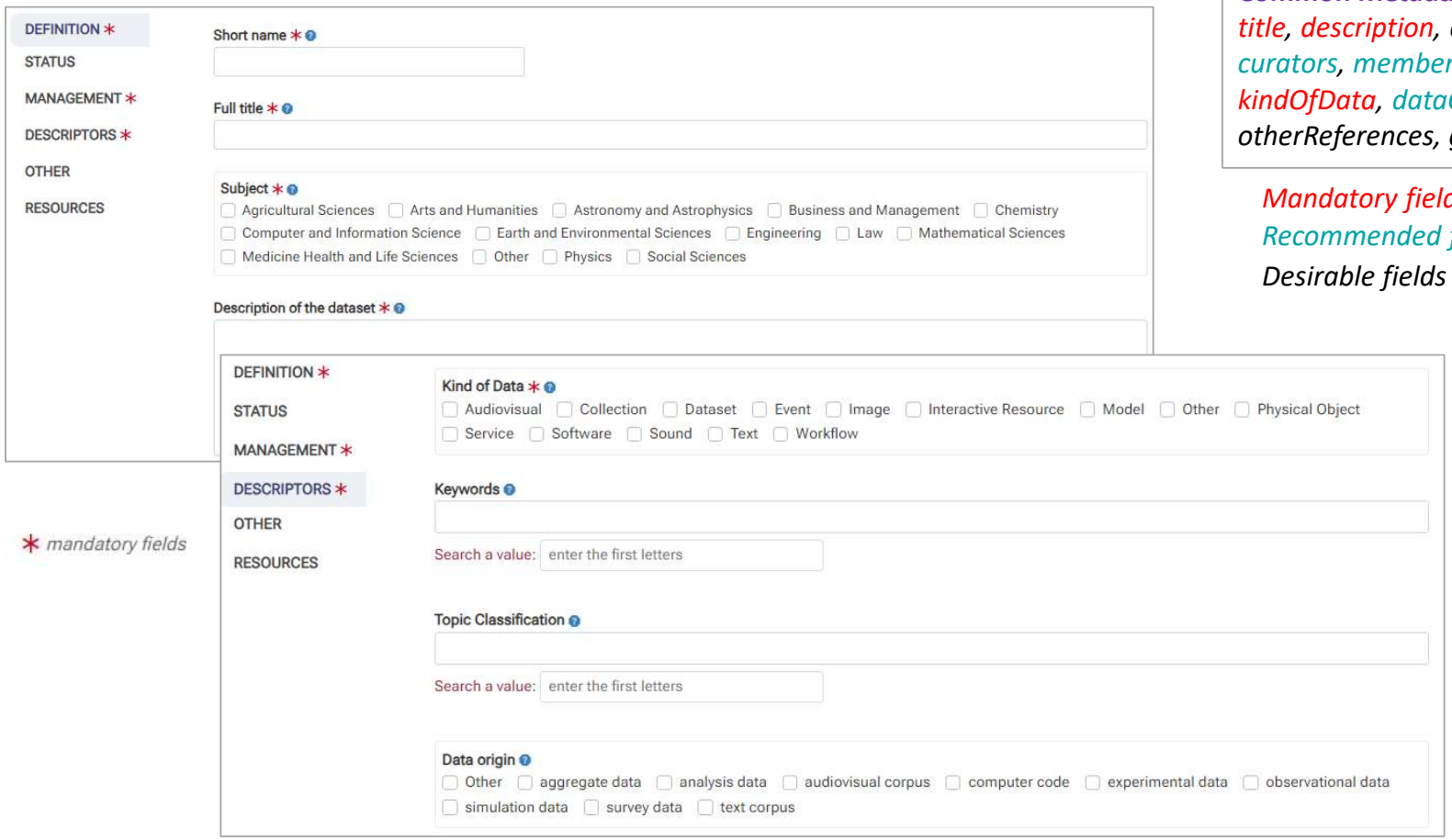
* Mandatory fields

* Recommended fields
that can be linked to an ontology
(or to a C.V)

* Desirable fields

Maggot — Metadata creation

Metadata aggregation on data storage



* *mandatory fields*

Common Metadata of « Dataverse »

Common Metadata

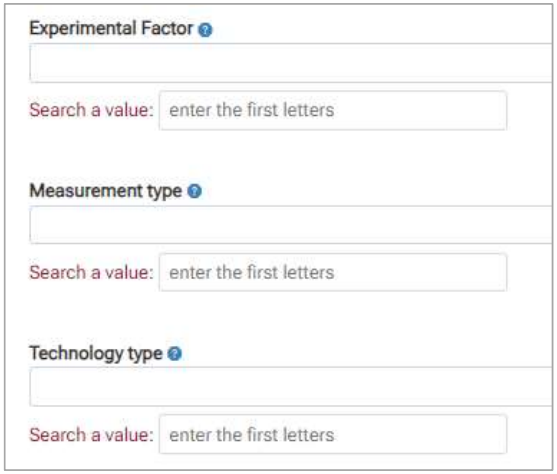
title, *description*, *alternativeURL*, *contacts*, *authors*, *collectors*, *curators*, *members*, *depositor*, *leader*, *subject*, *keywords*, *topics*, *kindOfData*, *dataOrigin*, *lifeCycleStep*, *publication*, *otherReferences*, *grantNumbers*, *project*, …

Mandatory fields

Recommended fields

Desirable fields

Specific Profile Metadata



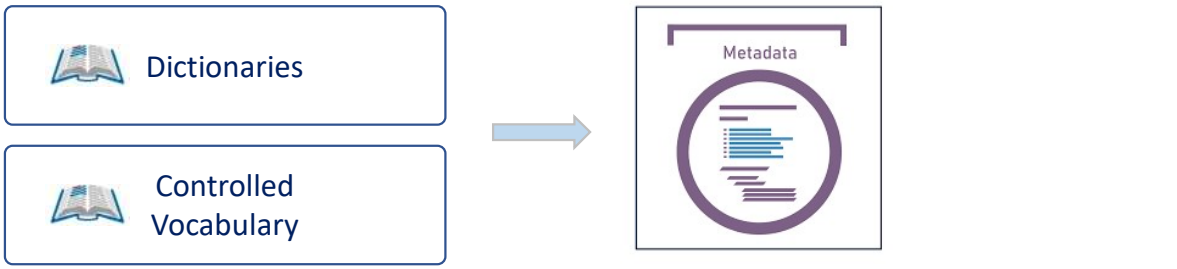
Metadata creation *(2/4)*

1 *Metadata schema (standard)*

(container)

Data Document Initiative

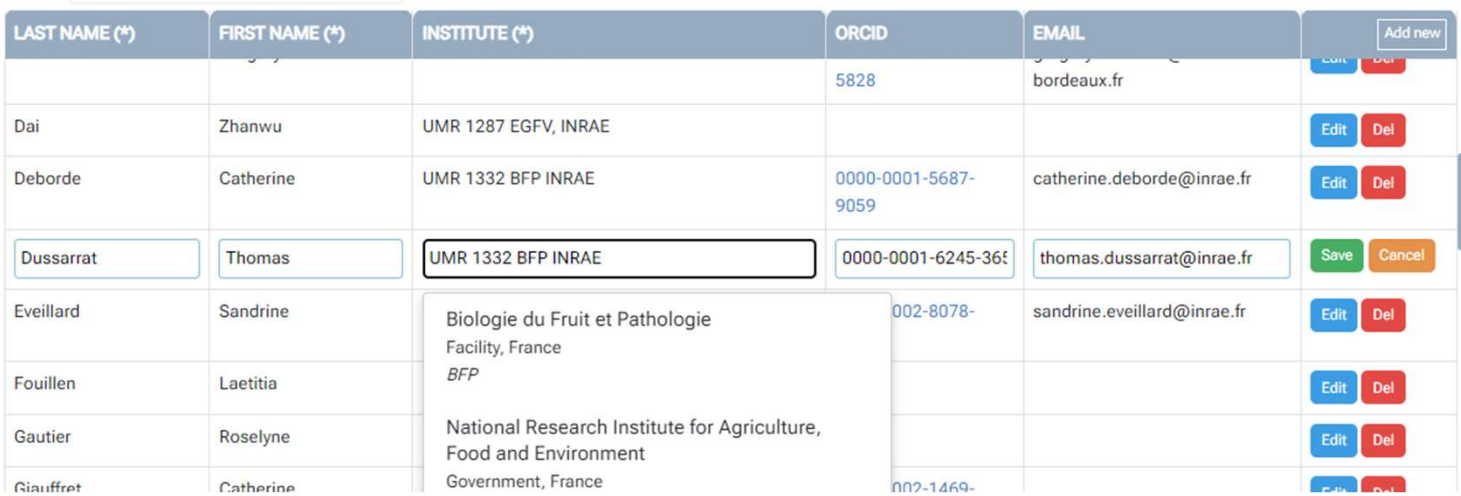
2 *Vocabulary* *(content)*



From what vocabulary ?

Dictionaries

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).



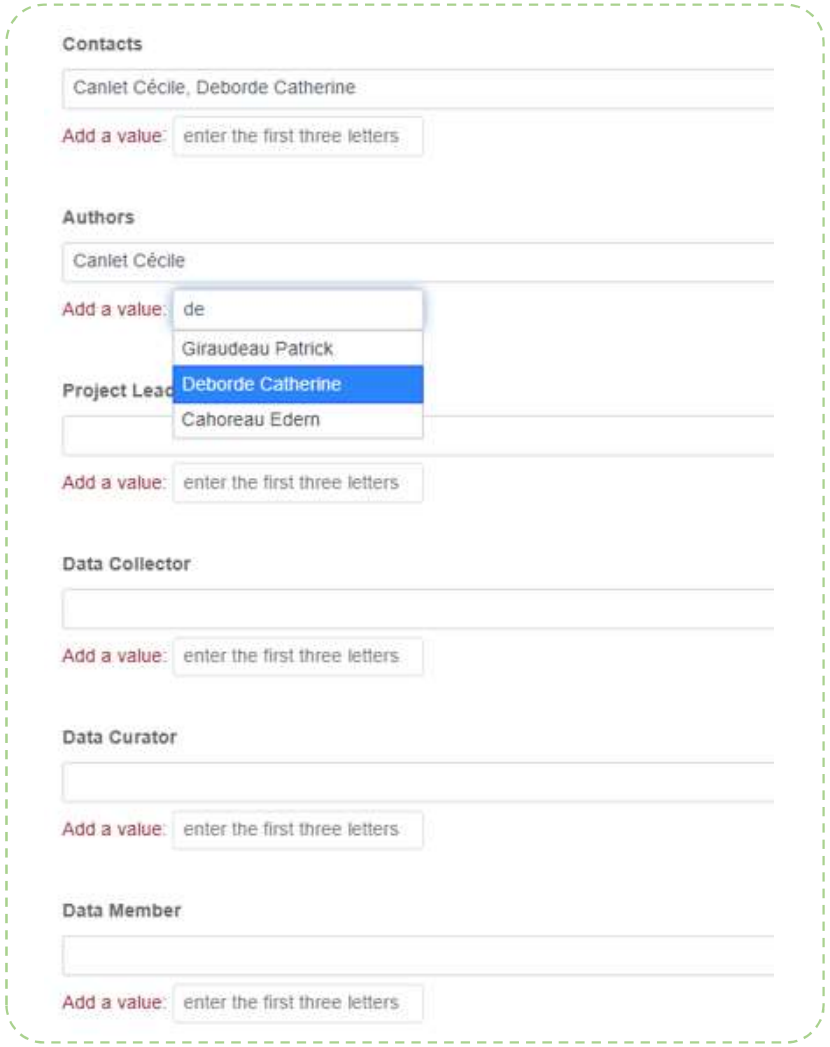
Dictionaries allow you to record multiple information necessary to define an entity, such as the names of people, or even the funders.

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

Maggot Dictionaries

Metadata aggregation on data storage

Example : people



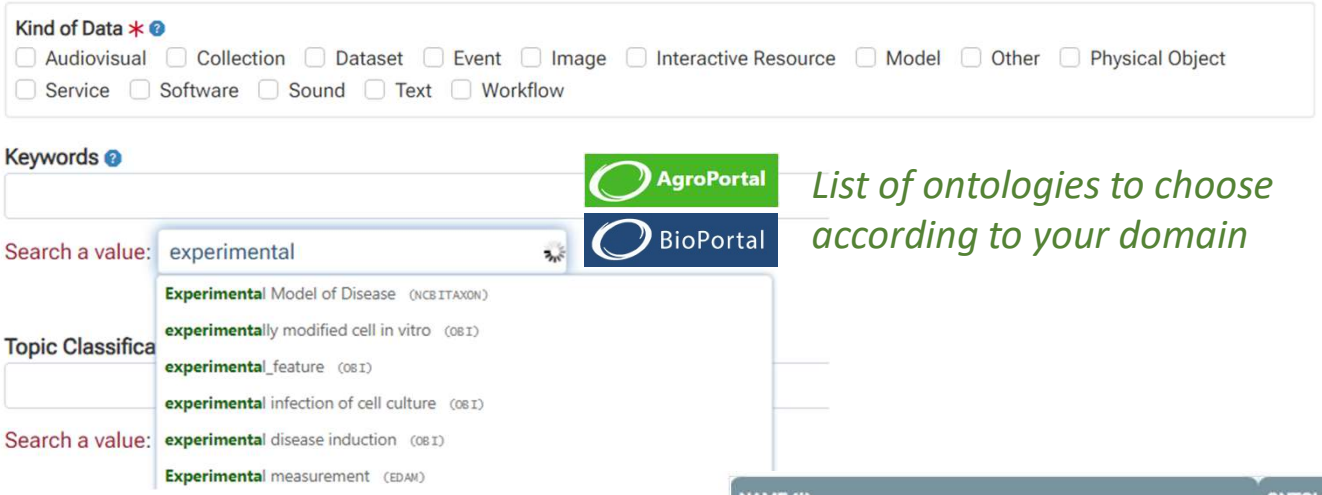
Push

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

Controlled vocabulary

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



AgroPortal
BioPortal

List of ontologies to choose according to your domain

Thesaurus SKOSMOS

(VOINRAE, LOTERRE, ONTOSTACK, …)

V.O.INRAE VOCABULAIRES OUVERTS

Building a vocabulary

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

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

Metadata creation *(3/4)*

1 *Metadata schema (standard)*

(container)

DDI

Data Document Initiative

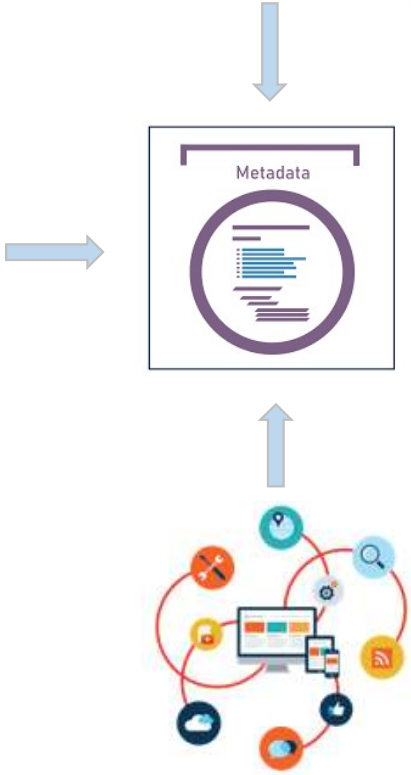
2 *Vocabulary (content)*

- Dictionaries
- Controlled Vocabulary

AgroPortal BioPortal

Thesaurus SKOSMOS

(VOINRAE, LOTERRE, ONTOSTACK, …)



3

External resources

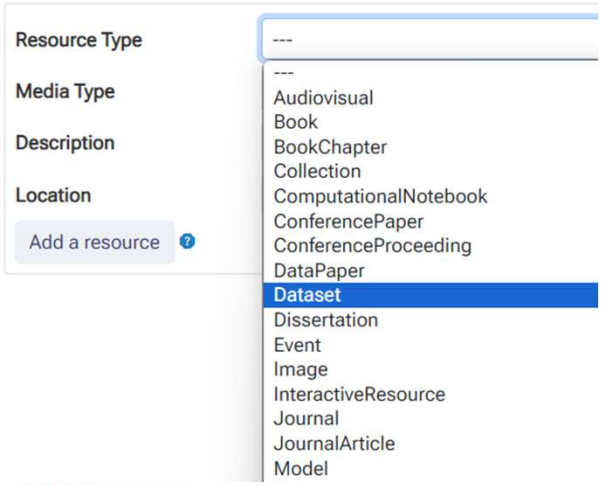
(related data)

Additional information ?

Other metadata / data ?

Maggot External Resources

Metadata aggregation on data storage



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.

Metadata creation *(4/4)*

1 *Metadata schema (standard)*

(container)

DDI

Data Document Initiative

2 *Vocabulary (content)*

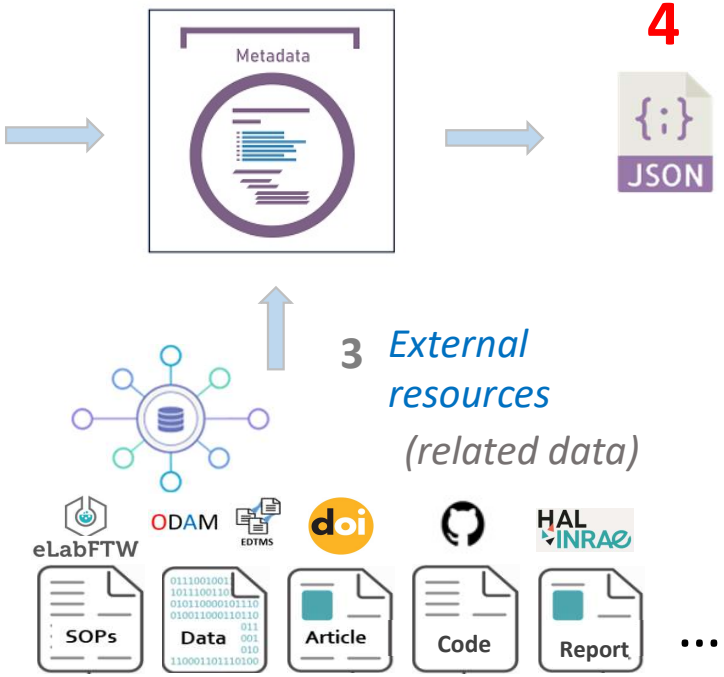
Dictionaries

Controlled Vocabulary

AgroPortal BioPortal

Thesaurus SKOSMOS

(VOINRAE, LOTERRE, ONTOSTACK, …)



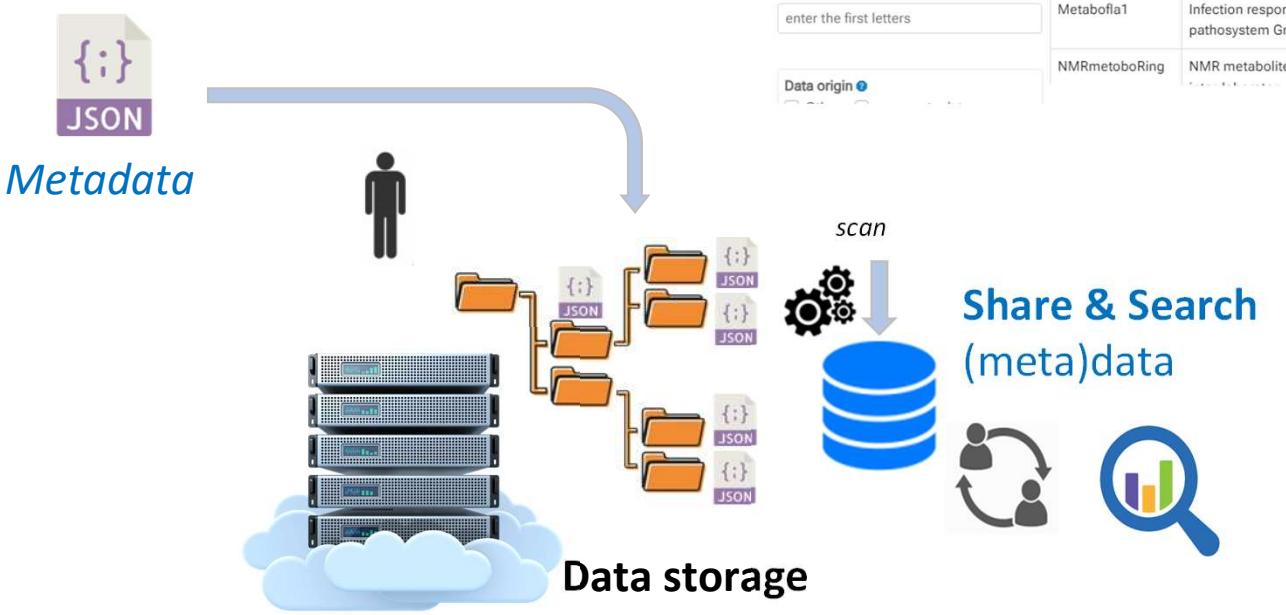
As an output, we produce a file in JSON format that can be read by both humans and machines.

What to do with this file?

Maggot

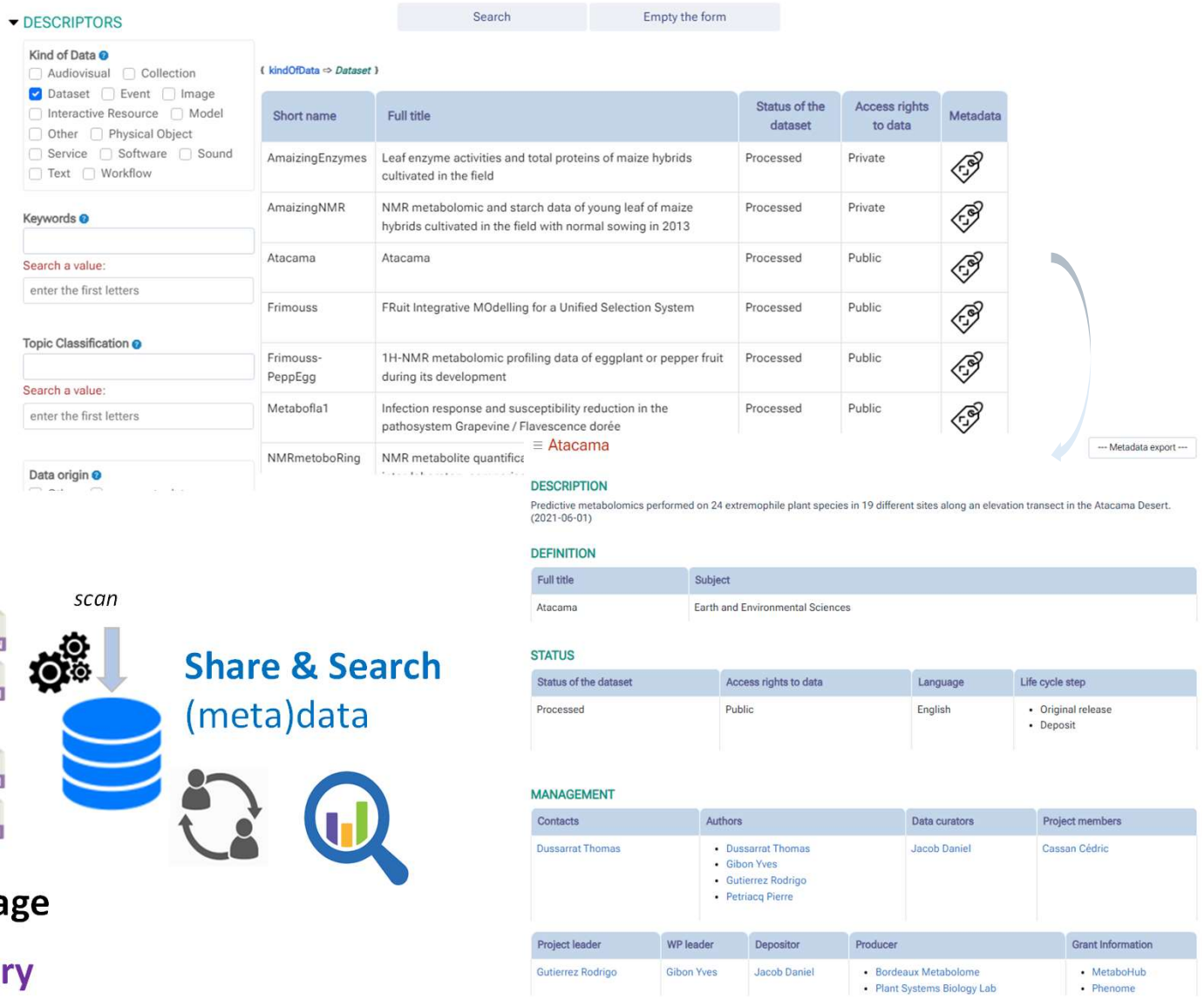
Metadata aggregation on data storage

The deposit must be done by hand because the Maggot tool must only have access to the storage space in read mode. So this guarantees that the user has writing rights to this space without having to manage user accounts on the Maggot side.



Local (meta)data repository

Storage space becomes the data repository

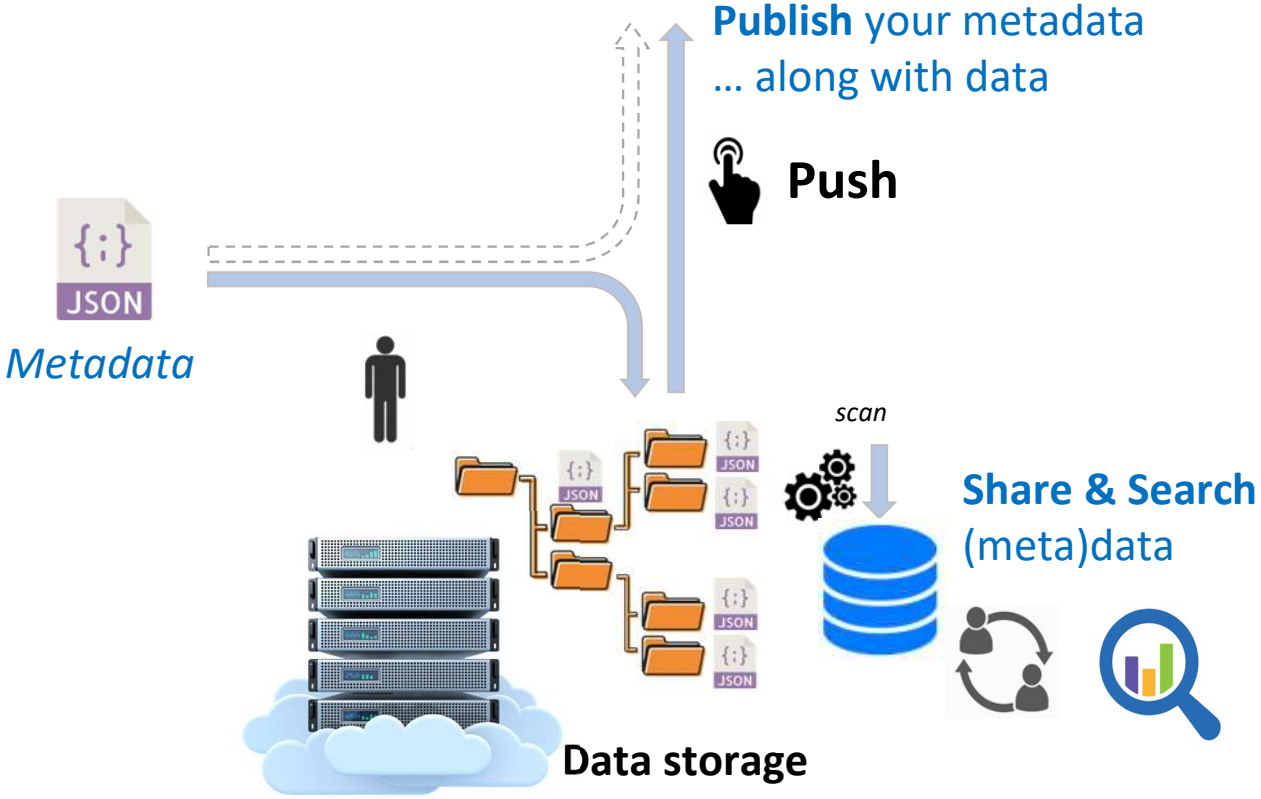


Maggot

Metadata aggregation on data storage

Institutional data repositories

The Dataverse Project / zenodo / ...



Local (meta)data repository

Storage space becomes the data repository

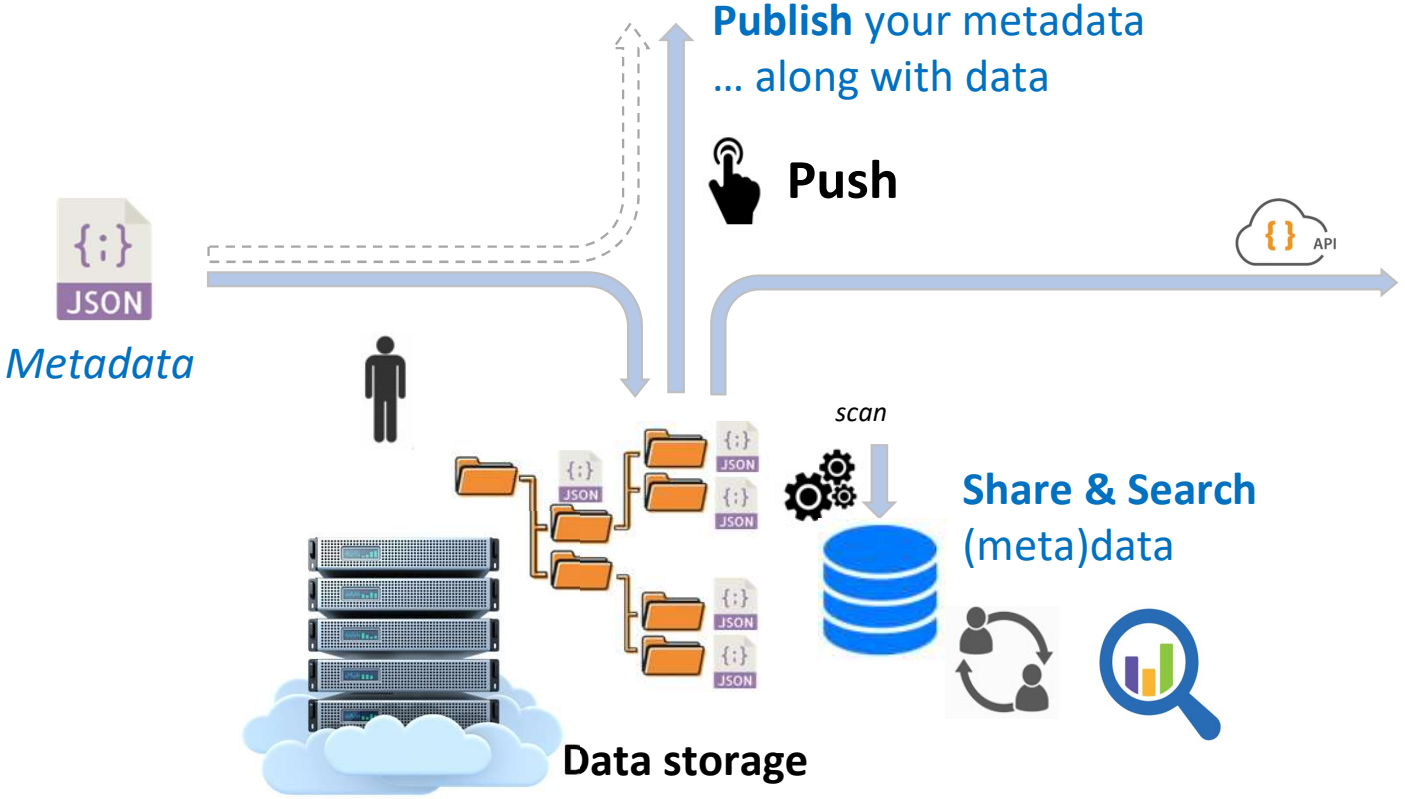
Maggot

Metadata aggregation on data storage

Institutional data repositories

The Dataverse Project / zenodo / ...

Findable Accessible Interoperable Reusable **Interoperability**



Allow machines to collect metadata

XML — *Open Archives Initiative Protocol for Metadata Harvesting*

Dublin Core

JSON-LD

Linked Open Data

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

Local (meta)data repository

Storage space becomes the data repository