



HAL
open science

An ecosystem for producing and sharing metadata within the web of FAIR Data

Daniel Jacob, François Ehrenmann, Romain David, Joseph Tran, Cathleen
Mirande-Ney, Philippe Chaumeil

► To cite this version:

Daniel Jacob, François Ehrenmann, Romain David, Joseph Tran, Cathleen Mirande-Ney, et al.. An ecosystem for producing and sharing metadata within the web of FAIR Data. *GigaScience*, 2025, 14, pp.1-8. 10.1093/gigascience/giae111. hal-04873759

HAL Id: hal-04873759

<https://hal.inrae.fr/hal-04873759v1>

Submitted on 9 Jan 2025

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 4.0 International License

An ecosystem for producing and sharing metadata within the web of FAIR Data

Daniel Jacob ^{1,2,3,*}, François Ehrenmann ⁴, Romain David ⁵, Joseph Tran ⁶, Cathleen Mirande-Ney ⁷, and Philippe Chaumeil⁴

¹INRAE, Université de Bordeaux, F-33140 Villenave d'Ornon, France

²INRAE, UR1268 BIA, Centre INRAE Pays de Loire–Nantes, F-44000 Nantes, France

³INRAE, BIBS facility, PROBE research infrastructure, F-44300 Nantes, France

⁴INRAE, Université de Bordeaux, F-33610 Cestas, France

⁵European Research Infrastructure on Highly Pathogenic Agents (ERINHA AISBL), B-1050 Bruxelles, Belgium

⁶INRAE, Bordeaux Sciences Agro, Université de Bordeaux, F-33140 Villenave d'Ornon, France

⁷INRAE, Université de Bordeaux, F-33140 Villenave d'Ornon, France

*Correspondence address. Daniel Jacob, INRAE, UR1268 BIA, Centre INRAE Pays de Loire–Nantes, F-44000 Nantes, France. E-mail: daniel.jacob@inrae.fr

Abstract

Background: Descriptive metadata are vital for reporting, discovering, leveraging, and mobilizing research datasets. However, resolving metadata issues as part of a data management plan can be complex for data producers. To organize and document data, various descriptive metadata must be created. Furthermore, when sharing data, it is important to ensure metadata interoperability in line with FAIR (Findable, Accessible, Interoperable, Reusable) principles. Given the practical nature of these challenges, there is a need for management tools that can assist data managers effectively. Additionally, these tools should meet the needs of data producers and be user-friendly, requiring minimal training.

Results: We developed Maggot (Metadata Aggregation on Data Storage), a web-based tool to locally manage a data catalog using high-level metadata. The main goal was to facilitate easy data dissemination and deposition in data repositories. With Maggot, users can easily generate and attach high-level metadata to datasets, allowing for seamless sharing in a collaborative environment. This approach aligns with many data management plans as it effectively addresses challenges related to data organization, documentation, storage, and the sharing of metadata based on FAIR principles within and beyond the collaborative group. Furthermore, Maggot enables metadata crosswalks (i.e., generated metadata can be converted to the schema used by a specific data repository or be exported using a format suitable for data collection by third-party applications).

Conclusion: The primary purpose of Maggot is to streamline the collection of high-level metadata using carefully chosen schemas and standards. Additionally, it simplifies data accessibility via metadata, typically a requirement for publicly funded projects. As a result, Maggot can be utilized to promote effective local management with the goal of facilitating data sharing while adhering to the FAIR principles. Furthermore, it can contribute to the preparation of the future EOSC FAIR Web of Data within the European Open Science Cloud framework.

Keywords: data management, FAIR, high-level metadata

Background

In scientific research, metadata play a crucial yet often overlooked role. Despite being essential for the discovery, reporting, and mobilization of research datasets, metadata remain poorly understood in scientific communities. However, since metadata are data themselves, they must be managed with the same level of rigor as the research data produced and consumed by research processes. This lack of awareness persists even in an era where sharing research data has become the cornerstone of open science initiatives and reproducible science. As transparency and collaboration become increasingly important to the scientific process, understanding the importance of metadata becomes imperative [1, 2].

However, the production of metadata requires effort and expertise, and data producers and curators may be reluctant to make this additional time investment unless they see a tangible return [3]. Therefore, proactive approaches are needed to overcome this hurdle and educate data producers about the benefits of open data practices [4].

Furthermore, the creation of metadata poses challenges for data producers. Data management plans (DMPs) that describe strategies for managing research data throughout their life cycle often ask nontrivial questions. For example, they may inquire about the interoperability of data or about the type of metadata schema used. These questions can be difficult to answer, especially when datasets span various scientific domains and require input from people with varying skills [5]. The great diversity of research data and the wide variety of characteristics they describe further complicate metadata management [6].

Given the complexity of the issue, it is important to differentiate between the different types and functions of metadata. To keep things simple, we can divide metadata into 2 main groups: high-level and specialized metadata. The latter comprises structural metadata that describe the organization and interconnections within a dataset. For example, structural metadata are essential to optimize the reuse of experimental data tables [7]. In contrast, high-level metadata (descriptive, administrative, rights)

Received: May 17, 2024. Revised: September 10, 2024. Accepted: November 27, 2024

© The Author(s) 2025. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

apply to all types of data generated within similar experimental contexts.

These 2 types of metadata can be considered either within the same data warehouse or separately. In the first case, the warehouse must be able to accommodate the data along with all associated metadata, typically resulting in repositories that are highly specialized (e.g., MetaboLights, [8]). In the second case, the different types of data produced can be distributed across various warehouses. Complementing successful initiatives describing experimental data arrays with ODAM [7], complex experiments with CEDAR [9], or omics data with the ISA-Tools suite [10], there are generalist data repositories such as Zenodo [11] or repositories based on the Harvard Dataverse software [12, 13], allowing users to deposit both high-level metadata and data that are not supported or insufficiently represented in existing data repositories.

High upstream in the metadata creation and curation chain (i.e., well before any dissemination), data—in all their diversity—need to be managed locally. Even within the same project, the production of data can be spread out over several years and involve several partners and numerous individuals, including fixed-term staff such as doctoral students and postdoctoral fellows. In this situation, data management practices must provide transparency and ensure access to the collective's data assets while adopting best practices such as the FAIR (Findable, Accessible, Interoperable, Reusable) principles [14].

Therefore, high-level metadata, a requirement for later data dissemination, are also highly relevant during the initial local data management. This should motivate data producers to provide a high level of documentation of their data, especially once they have been made aware of their benefits and the fact that this type of data only needs to be created once. In this work, we focused on the use of high-level metadata to locally manage a data catalog, with the prospect of later being able to distribute the data more easily in a data repository. Our approach is based on the Maggot (Metadata Aggregation on Data Storage) software, specifically designed to facilitate the documentation of datasets using high-level metadata in the form of files that can be attached to the storage space.

Design Considerations

Maggot (Metadata Aggregation on Data Storage) (RRID: SCR 025, 261) was developed to meet the need for a versatile data management tool that can support diverse annotation requirements. Its main objectives are to provide visibility of a collective's data assets, enable the general description of data, and promote the early adoption of FAIR principles. Furthermore, it ensures that data are kept in a format that is sustainable and facilitates reusability, particularly if data are produced by fixed-term staff (doctoral students and postdoctoral fellows), as it helps to create an awareness among less experienced staff members of the importance of good data description practices, thus fostering a culture of excellence in data management [3].

The wide range of scientific data is often managed separately using dedicated tools or repositories (omics data, experimental data tables, images, etc.). While each data type typically requires highly specific structural metadata, it may be possible to define a common set of high-level metadata descriptors (i.e., descriptive, administrative, rights) that apply across a wide variety of data types and usage scenarios, including contexts that require collective data sharing. To address the challenge of managing all data, Maggot was developed by relying primarily on high-level meta-

DMPs usually call for a centralized approach to data storage to ensure data safety (backup) and security (controlling access) (i.e., outside of users' disk space), which becomes particularly important when fixed-term staff are involved in the data production. However, data managers must consider how such centralized storage spaces can be organized (e.g., through harmonizing folder and file naming conventions and the use of README files to provide relevant information). The only constraint imposed by using Maggot is that each dataset must be associated with 1 root directory regardless of whether this contains the entire dataset or just a subset. If some data are stored elsewhere, they could be referenced using a URL, for example. Regarding directory trees, they can be created in any fashion that suits data producers (e.g., by project, theme, or team or using a combination of these). Instead of (or in addition to) a README file, a high-level metadata file is deposited in each dataset directory to clearly identify the particular dataset. These high-level metadata files allow Maggot to create a data catalog directly from the data repository. Maggot software has been specifically designed to provide effective answers regarding the choice, format, and means of creating relevant high-level metadata.

While some tools (e.g., CEDAR [9], FAIRDOME-SEEK [15]) use both high-level and structural metadata descriptors, Maggot relies on high-level metadata only because they apply to all the data while encouraging us to rely on other tools or online platforms specialized in a particular area or data type. This therefore leaves open to other tools the description of the data themselves (specialized, structural metadata), which can be of great diversity. For example, an image management tool like OMERO (<https://www.openmicroscopy.org/omero/>), being dedicated to this type of data, is therefore more able to describe them than a tool like Maggot. On the other hand, Maggot makes it possible to make the link between all the data. This approach allows metadata to be managed by mobilizing tools each dedicated to a particular type of metadata, thus leaving open the choice of possibilities. This approach is particularly advantageous for projects with a diversity of data types to process. For high-level metadata, the metadata schema should be chosen with some degree of foresight based on the data repository where the final data are to be deposited. In France, the national data repository [16] uses the Harvard Dataverse software, which is largely based on the standard DDI (Data Documentation Initiative) metadata schema [17]. The advantage of the DDI schema is that it encompasses a wealth of background information that can describe data sets of any type. It is also more extensive than the DataCite [18] or DublinCore [19] schemas. Due to these advantages, Maggot uses the DDI schema by default, although it is possible to employ one of the other schemas mentioned.

Typically, the metadata schema implemented by generalist data repositories only offers a small set of metadata, which may serve as a starting point. Consequently, it will be necessary to add other pertinent metadata to facilitate the local data management. As the choice of high-level metadata is an important step, collaboration between the data manager and data producers is essential to agree on an adequate minimum set of metadata. Although challenging, it is crucial to meticulously build and adapt the schema to align with existing and future data needs [20].

The high-level metadata span several types of information to be documented (descriptive, administrative, rights), implying different types of terminology. Since standardization is key to interoperability, metadata must meet established standards and be described using controlled vocabulary widely accepted by the scientific community [21]. Hence, both the metadata and sources

of vocabulary (ontologies, thesauri, dictionaries) must be agreed upon by data managers and producers. On the one hand, choosing metadata involves primarily the data producers who have direct knowledge of the data. This may represent a challenge because data producers may not be familiar with metadata standards and best practices. On the other hand, since data managers and data stewards have a high level of expertise in the application of FAIR principles [14] and metadata standards, they can play an important role in guiding the choice and management of metadata. In recognition of the complementarity of these roles, collaborative partnerships between data managers and scientists are essential to ensure effective and sustainable management of research data [22].

Data managers must raise awareness and encourage data producers to improve the quality and reusability of their data without requiring them to become subject matter experts [3]. This guidance is therefore only intended to provide recommendations on relevant metadata and controlled vocabulary for the relevant scientific area, as well as training data producers on best practices such as the use of permanent identifiers like DOI, ORCID, RoR, and other systems. Additionally, data producers should be informed about selecting appropriate licenses such as CC-BY [23], data policies [24], or data repositories [25].

Results

Maggot was designed for maximum usability, which meant keeping the user interface simple and automating metadata entry based on auto-completion whenever possible. However, the configuration remains slightly more complex, consisting of 2 configuration levels and several configuration tables in a spreadsheet (Fig. 1) [26].

Metadata definition

At configuration level 1, the high-level metadata must be defined, a crucial step that will affect all future data management. The input and search interfaces are entirely generated based on the terminology file, which defines each field, the corresponding input type, and associated vocabulary. Another level 1 configuration file serves to document each term through examples and links to additional information if necessary. This provides contextual assistance accessible during data entry, guiding data producers as they fill in each form field. A complete example is provided (Additional File 1), and a detailed documentation of the setup of configuration files is available online [27].

As mentioned above, Maggot uses the Harvard Dataverse metadata standards (DDI based) by default, serving as a useful starting point from which users can make customizations. Because other choices are possible, users should consult the MIT online metadata documentation before making any changes to the schema [28]. Furthermore, although metadata schemas should be linked to the FAIR principles [29], data producers should be able to modify any chosen schema if necessary to adequately describe the data. Maggot allows users to divide metadata into several sections, each section constituting a tab in the interface. Hence, a section dedicated to specific metadata not included in the standard metadata schema could be created to facilitate efficient searches in the catalog adapted to the given data type. The choice to extend the original metadata schema largely depends on the scientific field and the criteria that each collective wishes to establish for an adequate general description. Maggot possesses the necessary scal-

ability and flexibility to allow the creation of high-level metadata tailored to any experimental context.

To ensure the effective management of controlled vocabularies, Maggot provides users with a choice of dictionaries and ontologies and allows them to create their own custom dictionaries [30]. As it is unlikely that users will be able to conceive a complete set of appropriate terminologies right from the outset of a project, Maggot allows for an iterative and progressive approach. For example, users can start out with a simple dictionary based on local sources. As they start to consolidate their vocabulary, they can create a thesaurus (or a controlled vocabulary) that is separate from existing ontologies. To facilitate a quick start, Maggot can query thesauri directly from the SKOSMOS web application [31]. In addition, ontologies can be chosen gradually as data producers gain a better understanding of the relevant terminology and usage context. Indeed, Maggot allows users to enrich their metadata using ontologies publicly accessible via OntoPortal web applications such as BioPortal [32] and AgroPortal [33] but also via the EMBL-EBI Ontology Lookup Service [34].

Metadata crosswalks

At configuration level 2, users need to establish definitions of how their schema can be mapped onto a differently structured metadata format. This is termed “metadata crosswalk.” The DDI metadata schema that is used by default in Maggot is sufficiently rich to allow for mapping to other schemas such as DublinCore, for example. Maggot allows users to transform high-level metadata for deposition in data repositories. This can either be the default repositories that also use Maggot’s native DDI format such as Dataverse, or high-level metadata can be exported to other formats suitable for data harvesting (e.g., XML, JSON-LD) by third-party applications via an application programming interface (API) (e.g., OAI-PMH) [35]. These functionalities use a metadata crosswalk approach [36] based on the mapping files defined at configuration level 2. These files map the metadata defined at level 1 onto the output metadata schema (Fig. 1). This ensures maximum compatibility with other systems, in line with FAIR principles. Furthermore, this approach ensures long-term data preservation and facilitates a potential future migration away from Maggot. Maggot allows organizations to improve their data management practices, guaranteeing effective metadata use throughout their life span while facilitating data dissemination. Maggot also improves the interoperability and reusability of data while increasing the possibilities for data coupling, as envisaged by international consortia such as the European Open Science Cloud [37].

Features

Maggot’s functionalities can be divided into 3 parts: creation, sharing, and distribution (Fig. 2).

High-level metadata capture can be initiated from the very start of a project and does not require all data to be available or processed. In fact, metadata can be added in an iterative fashion throughout the duration of the project. Maggot supports descriptive and administrative metadata for any type of data, relying on user-defined custom fields where necessary. The metadata entry occurs via a form (Fig. 3), which is auto-generated based on the level 1 configuration. A minimum set of fields is mandatory to ensure compatibility for later data deposits. The selection of mandatory fields can be changed at any time by the data manager, who also defines the data policy and its implementation and governance. In contrast, data stewards are responsible for the data quality and curation. Before depositing the high-level metadata

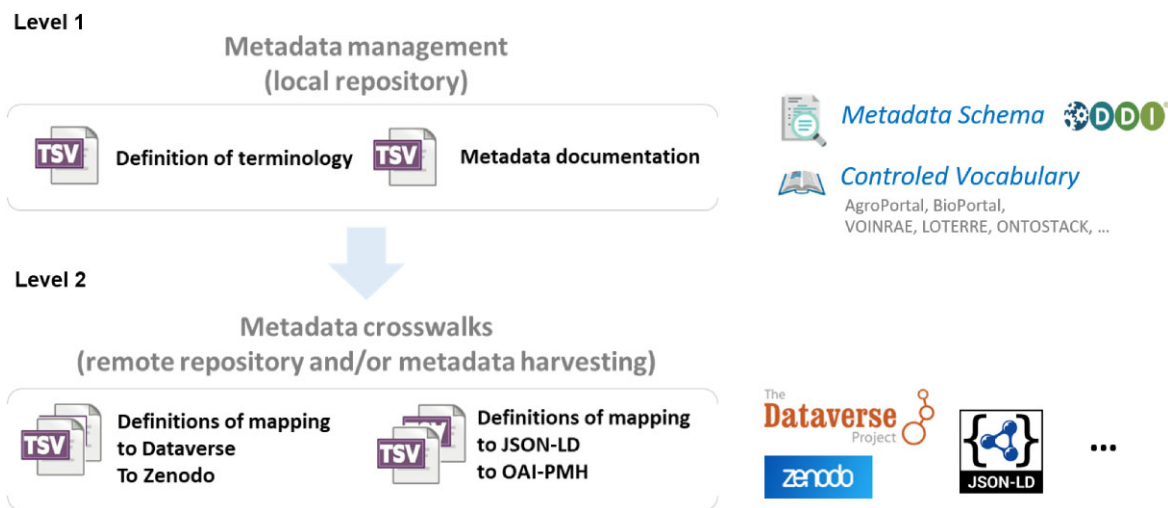


Figure 1: Maggot allows users to choose all the high-level metadata describing their data with 2 levels of definition files. The first level concerns the definition of metadata similar to a descriptive metadata plan. This category is more akin to configuration files and constitutes the heart of the configuration around which everything else is based. The input and search interfaces are completely generated from these definition files, thus defining each of the fields, their input type, and the associated controlled vocabulary. The second level concerns the definitions of the mapping to a differently structured metadata schema (metadata crosswalk, i.e., a specification for mapping 1 metadata standard to another), used either for metadata export to a remote repository (e.g., Dataverse, Zenodo) or for metadata harvesting (e.g., JSON-LD, OAI-PMH). The documentation gives the detailed workflow for the metadata dissemination in the Dataverse and Zenodo repositories showing the articulation of these 2 levels of configuration (<https://inrae.github.io/pgd-mmdt/publish/>).

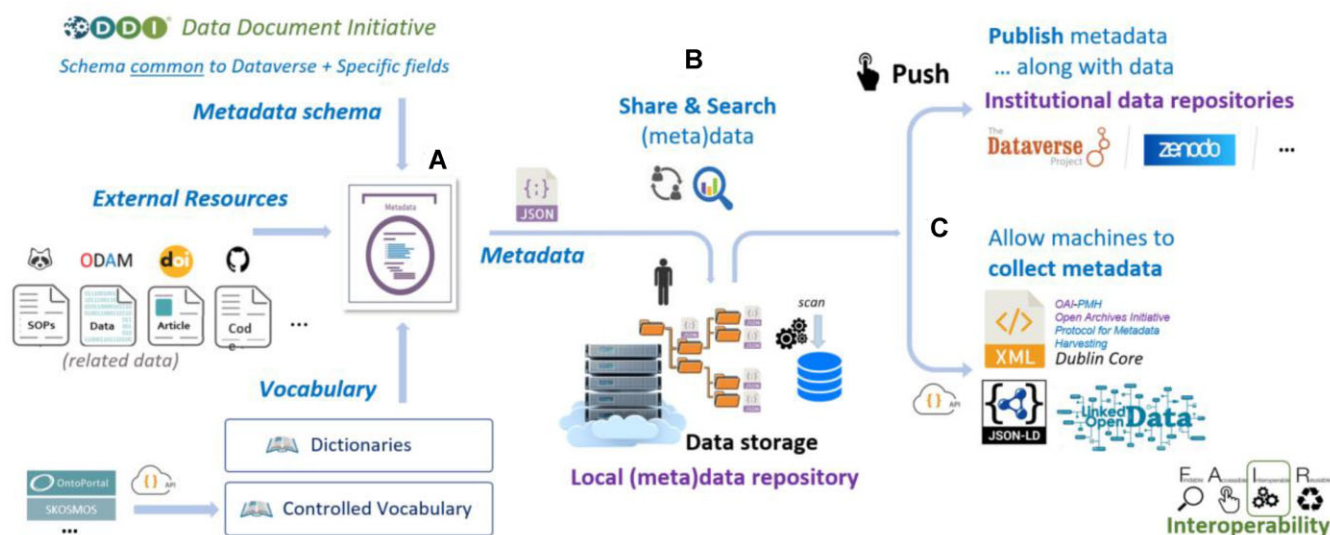


Figure 2: Main functionalities of Maggot split into 3 parts: creation, sharing, and dissemination. (A) First, producing a document with metadata sets of data within a collective of people, thus allowing users to (i) answer certain questions of the DMP concerning data organization, documentation, storage, and sharing in the data storage space and (ii) meet certain data and metadata requirements, listed, for example, by Open Research Europe in accordance with FAIR principles. (B) Next, searching for datasets by their metadata. Here, the descriptive metadata thus produced can be associated with the corresponding data directly in the storage space, making it possible to perform a search on the metadata to find 1 or more sets of data. Only high-level metadata are accessible by default. (C) Finally, publishing the high-level metadata of the datasets, as well as their data files, in a European-approved repository, with the possibility either to directly harvest the metadata via the OAI-PMH protocol or to export the associated metadata with their semantic context for full interoperability.

file in the storage space, Maggot allows users to send the metadata file to the data stewards for validation purposes and quality control. Data stewards can also handle the final submission if the designated storage space has limited write access. For each metadata input field, a help dialogue is available to provide a definition of the field and instructions for completing it. It is the responsibility of data managers to maintain high-quality documentation for data entry based on project requirements and users'/data stewards' feedback.

The sharing of data relies on both storage space and high-level metadata. To establish the search criteria, a form that closely resembles the entry form is provided (see Fig. 4). The quality of this research is dependent on the high-level metadata, which plays a crucial role at 2 levels. First, it is important to carefully choose relevant metadata to effectively target a specific set of data from the storage space. Second, attention must be given to the data entry process and any subsequent curation. Therefore, it is recommended to minimize the use of open fields for free text en-

▼ Pre-fill the form

Select a JSON file META_frim1.json **B**

E

Generate the metadata file Empty the form

DEFINITION *

STATUS

MANAGEMENT *

DESCRIPTORS *

OTHER

RESOURCES

A

Kind of Data *

Audiovisual Collection Dataset Event Image Interactive Resource Model Other Physical Object

Service Software Sound Text Workflow

Keywords

tomato, fruit growth, experimental measurement, plant trait

Search a value: **C**

Topic Classification

fruit growth, p

Search a value: **C**

Data origin

Other simulation

Experimental Fa

fruit developm

DESCRIPTORS

D

This section allows you to define elements characterizing the data themselves and certain experimental conditions for obtaining them.

Keywords

- A key term that describes an important aspect of the dataset and information about any controlled vocabulary used. Based on [BioPortal ontologies](#) : EFO, JERM, EDAM, MS, NM, NCI, OBI, PO, PTO, AGRO, ECOCORE, IOBC, NCBITAXON.

Topic Classification

- Indicates a broad, important topic or subject that the dataset covers and information about any controlled vocabulary used. Based on [Thesaurus-INRAE](#).

Kind of Data

- The type of data included in the files (e.g. survey data, machine-readable text, experimental data tables). The list of terms is imposed by Dataverse. Adding, modifying or deleting them may prevent uploading to the repository. See [DDI-CV](#)

* mandatory fields

Figure 3: The metadata input form is generated based on the terminology file created during the initial system setup. (A) Metadata fields are distributed between several tabs where related input fields are grouped together. (B) By loading a previously created metadata file, all form fields will be initialized to the predefined values. Mandatory fields are marked with a red asterisk. (C) Controlled vocabulary can be entered as free-form text, although the system offers to autocomplete entries based on a list of terms retrieved from the web. (D) Help for each input field can be obtained by clicking on a “?” icon, providing users with a definition of the corresponding field. (E) Once completed, the form can be saved as a file (an example has been provided as [Additional File 1](#)).

▼ **DESCRIPTORS** **A**

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

B

(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	
Frimouss-PeppEgg	1H-NMR metabolomic profiling data of eggplant or pepper fruit during its development	Processed	Public	
Metabofla1	Infection response and susceptibility reduction in the pathosystem Grapevine / Flavescence dorée	Processed	Public	
NMRmetoboRing	NMR metabolite quantification of a synthetic urine sample: an	Processed	Public	

Figure 4: Maggot allows users to manage a data catalog in the designated local storage space. The dataset search is divided into 2 parts: (A) a form almost identical to the entry form is provided to establish the search criteria using all user-provided metadata and (B) a table containing the default data sets. Only datasets meeting the search criteria will remain. Clicking on each of the columns in this table displays the datasets sorted accordingly. The column headers can be customized in the terminology definition file (<https://inrae.github.io/pgd-mmdt/definitions/terminology/>).

try. Whenever possible and appropriate, a controlled vocabulary should be used and enforced. However, since it is impossible to predict all scenarios in advance, Maggot allows for open fields for data entry over a long period of time. Any new entries can be regulated either by the data stewards through additions to dictionaries or by the system itself. In the latter case, Maggot provides fields where prerecorded options can be selected while still allowing users to enter new entries. These new entries are immediately recorded in the system and can then be proposed for selection to other users or datasets [38].

Sharing metadata does not imply sharing data. However, there are various ways to provide access to the data themselves before their distribution. One option is to install a file browser, either with or without passwords [39]. Another possibility is to deposit all the data in the collective's data center and include a link to this resource in the metadata file. Maggot allows for data fragmentation, meaning that data can be dispersed across different platforms, databases, and file formats. This allows data producers to specify resources, both external and internal, and centralize all links to the data (Fig. 2). External resources should be specified using a URL, with a preference for a permanent identifier such as a DOI. Any URL that points to data and respects the FAIR principle can also be used. Additionally, in cases where local data management is applicable, it is advisable to indicate the location of the data if it is different from that of the metadata (e.g., NAS unit or data cloud). By bringing together all references to multiple data sources in 1 place, Maggot can function as a data hub.

The distribution of data requires high-level metadata. It is crucial to have the mapping file properly configured upstream so that metadata can be crosswalked from the original internal schema to the schema of the target repository, which may not use all metadata fields originally defined. Although Maggot is currently limited to 2 repository platforms (Dataverse and Zenodo), there may be support for others in the future (e.g., Dryad [40] and RO-Crate [41]). This also does not prevent the reuse of metadata. It is entirely possible, for instance, to establish an internal metadata harvesting process to automatically populate another data source, such as the FAIRDOM-SEEK data management platform [15]. By choosing Maggot, users are not restricted to this system as they can export the generated metadata to other formats and platforms, which ensures that future applications/services can still make use of legacy metadata, thereby avoiding data loss. Maggot facilitates this by allowing data scientists and data repositories to harvest data. Through the OAI-PMH protocol, users can retrieve all datasets based on the DublinCore schema, while the metadata can be collected in JSON-LD format [42], which adheres to the schema.org standard [43]. This aspect is particularly critical for linking metadata in the linked data domain and ensuring interoperability. Future releases of Maggot will support DCAT-based harvesting [44].

High-level metadata alone are insufficient to fully describe a dataset, and structural metadata are needed as well. For instance, when dealing with experimental data tables that are managed using ODAM [7], the structural metadata are provided in the "Frictionless data package" standard format [45], which enables data users to easily parse the data. As a result, this data package file can be deposited in a data repository along with the high-level metadata (e.g., [46]). It is important to note that Maggot only handles high-level metadata for ODAM resources. Knowing that with ODAM data, management also relies on storage space, perfect complementarity exists between these 2 tools, each managing a specific level of metadata.

Another usage scenario for Maggot is the production of high-level metadata that are directly pushed to a data repository without requiring a local storage. In this case, the dataset is not registered in the local data catalog. This allows users to utilize the Maggot web interface, thus benefiting from all its contributions facilitating the entry of metadata (including dictionaries, controlled vocabulary, etc.) instead of the web interface provided by the data repository.

Implementation and Documentation

Deploying Maggot requires 2 infrastructure components: (i) a server to host the web application and (ii) a data storage space. The server must be capable of running a Linux-based operating system and support containerization using Docker. The latter facilitates easy installation and administration. The data storage can be local (e.g., NAS unit) or remote (e.g., cloud based). Data access can be managed via the rclone tool [47].

Maggot is a web-based PHP application that uses MongoDB [48] to index all metadata obtained by scanning the disk storage at 30-minute intervals. In addition, Maggot utilizes several remote vocabularies (thesauri and ontologies) that it queries via API to facilitate real-time imports, reducing the need to manually update information. For example, Maggot uses Twitter's Typeahead library [49], which allows data managers to easily implement a new vocabulary. The SKOSMOS thesauri and EMBL-EBI Ontology Lookup Service (OLS) have also been implemented in this way. While API access to vocabularies uses a caching mechanism to speed up the export of metadata to other formats (e.g., JSON-LD) or to push to a data repository (Dataverse, Zenodo), this caching mechanism is disabled when searching for a term by auto-completion in the input interface. As Maggot allows different vocabulary sources (e.g., BioPortal and EBI OLS), it is possible for 2 versions of the same ontologies to coexist. Hence, the list of ontologies for each source and field needs to be specified with care to prevent them from overlapping.

Additional documentation is available at [50] and within the application itself, with detailed explanations of how the terminology should be constructed using associated vocabularies.

Conclusion and Perspectives

Maggot is a tool designed specifically for annotating datasets by generating high-level metadata files that can be linked to storage spaces. It addresses challenges related to data organization, documentation, storage, and sharing of metadata in line with FAIR principles. By covering as much of the research data life cycle as possible, Maggot ensures efficient and sustainable management of research data and simplifies the adoption of FAIR principles. This enables organizations to increase the value and accessibility of their data assets. Additionally, Maggot's ability to disseminate metadata based on standard (machine-readable) schemas make it an important tool for the creation of the future EOSC FAIR Data Web, part of the European Open Science Cloud.

To date, Maggot is primarily used on the intranet of organizations and research units for processing and managing different types of data and metadata. Furthermore, efforts to build a community around this tool are underway (e.g., discussion blog, provision of several configurations for different areas of application). This is intended to help users with the installation, configuration, and use of Maggot. Future releases will allow exports in RO-Crate

[41] format and metadata harvesting based on DCAT. Additionally, there are plans to implement centralized authentication mechanisms (SSO) to cater to the requests of multisite institutions such as universities.

Availability of Source Code and Requirements

- Project name: Maggot
- Project homepage: <https://pmb-bordeaux.fr/maggot/>
- Project code repository: <https://github.com/inrae/pgd-mmdt>
- Documentation: <https://inrae.github.io/pgd-mmdt/>
- Operating system(s): Platform independent
- Programming languages: PHP, python, JavaScript
- Licence: GNU GPL v3
- Maggot, RRID: SCR_025261
- Biotools: <https://bio.tools/maggot>

Data Availability

Examples of metadata files are included as an **Additional File**, and snapshots of the code are available in Software Heritage [51].

Additional Files

Additional File 1. Examples of metadata files along with corresponding definitions files within a ZIP file (https://inrae.github.io/pgd-mmdt/ex/maggot_file_examples.zip).

Abbreviations

API: Application Programming Interface; DDI: document, discover, and interoperate; DMP: data management plan; EOSC: European Open Science Cloud; FAIR: Findable, Accessible, Interoperable, Reusable; JSON: JavaScript Object Notation; JSON-LD: JSON for Linking Data; OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting; NAS: Network attached storage.

Acknowledgments

We thank Dr. Catherine Deborde (INRAE UR BIA & PROBE Research Infrastructure, BIBS Facility Nantes) and Dr. Annick Moing (INRAE UMR 1332 BFP, Bordeaux) for advice on the manuscript and for constructive reviews. We also thank Edouard Guitton (INRAE, Animal Health Department) for his fruitful feedback in the implementation of Maggot within his research infrastructure.

Author Contributions

Conceptualization: DJ., F.E., P.C.; funding acquisition: DJ., F.E.; methodology: DJ., F.E., R.D.; software: DJ., F.E.; writing—original draft: DJ., R.D.; writing—review and editing: all authors. All authors read and approved the final manuscript.

Funding

DJ. was partly supported by the MetaboHUB project funded by the French National Research Agency (ANR-11-INBS-0010) and by the WAPNMR project funded by the French National Research Agency (ANR-21-CE21-0014). DJ., F.E., C.M.N., and J.T. were partly supported by the Bordeaux Plant Science (BPS) project funded by the

Université de Bordeaux. DJ., F.E., C.M.N., J.T., and P.C. were partly supported by the French National Research Institute for Agriculture, Food and the Environment (INRAE). R.D. was supported by the European Research Infrastructure on Highly Pathogenic Agents (ERINHA AISBL).

Competing Interests

The authors declare that they have no competing interests.

References

1. Ulrich H, Kock-Schoppenhauer A, Deppenwiese N, et al. Understanding the nature of metadata: systematic review. *J Med Internet Res* 2022;24(1):e25440. <https://doi.org/10.2196/25440>.
2. Manninen L. Describing data: a review of metadata for datasets in the Digital Commons Institutional repository platform: problems and recommendations. *J Library Metadata* 2018;18:1–11. <https://doi.org/10.1080/19386389.2018.1454379>.
3. David R, Mabile L, Specht A, et al. FAIRness literacy: the Achilles' heel of applying FAIR principles. *CODATA Data Sci J* 2020;19:32. <https://doi.org/10.5334/dsj-2020-032>.
4. Popkin G. Data sharing and how it can benefit your scientific career. *Nature* 2019;569:445–47. <https://doi.org/10.1038/d41586-019-01506-x>.
5. David R, Baumann K, LeFranc Y, et al. Converging on a Semantic Interoperability framework for the European Data Space for Science, Research and Innovation (EOSC). Presented at the 2nd Workshop on Ontologies for FAIR and FAIR Ontologies; Sherbrooke, Québec; July 18th 2023. <https://doi.org/10.5281/zenodo.8102786>.
6. Çaldağ MT, Gökalp E. Understanding barriers affecting the adoption and usage of open access data in the context of organisations. *Data Inf Manage* 2023;100049. <https://doi.org/10.1016/j.di.m.2023.100049>.
7. Jacob D, David R, Aubin S, et al. Making experimental data tables in the life sciences more FAIR: a pragmatic approach. *Gigascience* 2020;9(12):giaa144. <https://doi.org/10.1093/gigascience/giaa144>.
8. EMBL-EBI. MetaboLights. <https://www.ebi.ac.uk/metabolights/>. Accessed 27 November 2024.
9. Musen MA, Bean CA, Cheung KH, et al. The center for expanded data annotation and retrieval. *J Am Med Inform Assoc* 2015;22:1148–52. <https://doi.org/10.1093/jamia/ocv048>.
10. Sansone SA, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. *Nat Genet* 2012;44:121–26. <https://doi.org/10.1038/ng.1054>.
11. Zenodo. <https://zenodo.org/>. Accessed 27 November 2024.
12. Harvarde Dataverse. <https://dataverse.harvard.edu>. Accessed 27 November 2024.
13. King G. An introduction to the dataverse network as an infrastructure for data sharing. *Sociol Methods Res* 2007;36(2):173–99. <https://doi.org/10.1177/0049124107306660>.
14. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. <https://doi.org/10.1038/sdata.2016.18>
15. Wolstencroft K, Owen S, Krebs O, et al. SEEK: a systems biology data and model management platform. *BMC Syst Biol* 2015;9:33. <https://doi.org/10.1186/s12918-015-0174-y>.
16. Recherche Data Gouv. <https://entrepot.recherche.data.gouv.fr>. Accessed 27 November 2024.

17. Document, Discover and Interoperate. <https://ddialliance.org>. Accessed 27 November 2024.
18. Datacite Metadata Schema. <https://schema.datacite.org/>. Accessed 27 November 2024.
19. DublinCore Metadata Initiative. <https://www.dublincore.org/schemas/>. Accessed 27 November 2024.
20. David R, Richard AS, Connellan C, et al. Umbrella Data Management plans to integrate FAIR data: lessons from the ISIDORE and BY-COVID consortia for pandemic preparedness. *Data Sci J* 2023;22:1–15. <https://doi.org/10.5334/dsj-2023-035>
21. Bingo S, Montenegro M. How to create a descriptive metadata plan. Sustainable Heritage Network. <https://sustainableheritagenetwork.org/digital-heritage/how-create-descriptive-metadata-plan>. Accessed April 13, 2024.
22. David R, Rybina A, Burel J-M, et al. “Be sustainable”: eOSC-Life recommendations for implementation of FAIR principles in life science data handling. *EMBO J* 2023;42:e115008. <https://doi.org/10.15252/embj.2023115008>.
23. Creative Commons. <https://creativecommons.org>. Accessed 27 November 2024.
24. FAIRsharing. <https://fairsharing.org/>. Accessed 27 November 2024.
25. Registry of Research Data Repositories. <https://www.re3data.org/>. Accessed 27 November 2024.
26. Maggot:Metadata definition files. <https://inrae.github.io/pgd-mmdt/definitions/>. Accessed 27 November 2024.
27. Maggot:Terminology configuration. <https://inrae.github.io/pgd-mmdt/configuration/>. Accessed 27 November 2024.
28. MIT Libraries: Documentation & metadata. <https://libraries.mit.edu/data-management/store/documentation/>. Accessed 27 November 2024.
29. Musen MA, O'Connor MJ, Schultes E, et al. Modeling community standards for metadata as templates makes data FAIR. *Sci Data* 2022;9:696. <https://doi.org/10.1038/s41597-022-01815-3>.
30. Maggot:Dictionaries. <https://inrae.github.io/pgd-mmdt/dictionaries/>. Accessed 27 November 2024.
31. Suominen O, Ylikotila H, Pessala S, et al. Publishing SKOS vocabularies with Skosmos. Unpublished manuscript, Accessed 27 November 2024. <https://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf>.
32. Noy NF, Shah NH, Whetzel PL, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;37(suppl 2):W170–W73. <https://doi.org/10.1093/nar/gkp440>.
33. Jonquet C, Toulet A, Arnaud E, et al. AgroPortal: a vocabulary and ontology repository for agronomy. *Comput Electron Agric* 2018;144:126–43. <https://doi.org/10.1016/j.compag.2017.10.012>.
34. EMBL-EBI Ontology Lookup Service. <https://www.ebi.ac.uk/ols4>. Accessed 27 November 2024.
35. Open Archives Initiative Protocol for Metadata Harvesting. <https://www.openarchives.org>. Accessed 27 November 2024.
36. Maggot:Metadata Crosswalk Definition. <https://inrae.github.io/pgd-mmdt/chats/chat4/>. Accessed 27 November 2024.
37. EOOSC EU Node. <https://open-science-cloud.ec.europa.eu>. Accessed 27 November 2024.
38. Maggot:Vocabulary. <https://inrae.github.io/pgd-mmdt/definitions/vocabulary/>. Accessed 27 November 2024.
39. Maggot:Installation. <https://inrae.github.io/pgd-mmdt/installation/>. Accessed 27 November 2024.
40. Isard M, Budiu M, Yu Y, et al. Dryad: distributed data-parallel programs from sequential building blocks. *SIGOPS Oper Syst Rev* 2007;41:59–72. <https://doi.org/10.1145/1272998.1273005>.
41. Soiland-Reyes S, Sefton P, Crosas M, et al. Packaging research artefacts with RO-Crate. *Data Sci* 2022;5:97–138. <https://doi.org/10.3233/DS-210053>.
42. JSON for Linking Data. <https://json-ld.org>. Accessed 27 November 2024.
43. Schema.org. <https://schema.org>. Accessed 27 November 2024.
44. Data Catalog Vocabulary. <https://www.w3.org/TR/vocab-dcat-3/>. Accessed 27 November 2024.
45. Frictionless Data. <https://frictionlessdata.io/>. Accessed 27 November 2024.
46. Bénard C, Biais B, Ballas P, et al. FRIM—Fruit Integrative Modelling. November 15th 2024. <https://doi.org/10.15454/95JUTK>.
47. Rclone software. <https://rclone.org>. Accessed 27 November 2024.
48. MongoDB software. <https://www.mongodb.com>. Accessed 27 November 2024.
49. Typeahead JavaScript library. <https://twitter.github.io/typeahead.js/>. Accessed 27 November 2024.
50. Online Documentation of Maggot software. <https://inrae.github.io/pgd-mmdt/>. Accessed 27 November 2024.
51. Jacob D, Ehrenmann F, David R, et al. 2024. Maggot: Metadata Management Tool for Data Storage Spaces. [Computer software]. Software Heritage. <https://archive.softwareheritage.org/swh:1:snp:a7a122107b687fb13dd426822c2984ed93ea304b;origin=https://github.com/inrae/pgd-mmdt>.