

1 **Maggot : An ecosystem for sharing metadata within the** 2 **web of FAIR Data**

3 Daniel Jacob^{1†}, François Ehrenmann², Romain David³, Joseph Tran⁴, Cathleen Mirande-Ney⁵, Philippe
4 Chaumeil²

5 †Corresponding author

6 Institutional addresses:

7 ¹ INRAE, Université de Bordeaux, UMR BFP, 71 av E Bourlaux, F-33140 Villenave d'Ornon, France

8 ² INRAE, Université de Bordeaux, UMR BIOGECO, 69 route d'Arcachon, F-33610 Cestas, France

9 ³ European Research Infrastructure on Highly Pathogenic Agents (ERINHA AISBL), 98 rue du Trône
10 B-1050 Bruxelles, Belgium

11 ⁴ INRAE, Université de Bordeaux, Bordeaux Sciences Agro, ISVV, UMR EGFV, F-33140 Villenave
12 d'Ornon, France

13 ⁵ Université de Bordeaux, INRAE, UMR BFP, 71 av E Bourlaux, F-33140 Villenave d'Ornon, France

14 Email addresses, ORCID:

15 DJ: daniel.jacob@inrae.fr, ORCID: 0000-0002-6687-7169

16 FE: francois.ehrenmann@inrae.fr, ORCID:0000-0003-2727-0070

17 RD: romain.david@erinha.eu, ORCID: 0000-0003-4073-7456

18 JT: joseph.tran@inrae.fr, ORCID: 0000-0002-4624-0363

19 CMN: cathleen.mirande-ney@inrae.fr, ORCID: 0000-0001-9760-8482

20 PC: philippe.chaumeil@inrae.fr

21

22 **Abstract**

23 **Background** : Descriptive metadata are crucial for the discovery, reporting and mobilisation of
24 research datasets. Addressing all metadata issues within the Data Management Plan often poses
25 challenges for data producers. Organising and documenting data within data storage entails
26 creating various descriptive metadata. Subsequently, data sharing involves ensuring metadata
27 interoperability in alignment with FAIR principles. Given the tangible nature of these challenges,
28 a real need for management tools has to be addressed to assist data managers to the fullest
29 extent. Moreover, these tools have to meet data producers requirements and be user-friendly as
30 well with minimal training as prerequisites.

31

32 **Results** : We developed Maggot which stands for Metadata Aggregation on Data Storage,
33 specifically designed to annotate datasets by generating metadata files to be linked into storage
34 spaces. Maggot enables users to seamlessly generate and attach comprehensible metadata to
35 datasets within a collaborative environment. This approach seamlessly integrates into a data
36 management plan, effectively tackling challenges related to data organisation, documentation,

37 storage, and frictionless FAIR metadata sharing within the collaborative group and beyond.
38 Furthermore, for enabling metadata crosswalk, metadata generated with Maggot can be
39 converted for a specific data repository or configured to be exported into a suitable format for data
40 harvesting by third-party applications.

41
42 **Conclusion** : The primary feature of Maggot is to ease metadata capture based on a carefully
43 selected schema and standards. Then, it greatly eases access to data through metadata as
44 requested nowadays in projects funded by public institutions and entities such as Europe
45 Commission. Thus, Maggot can be used on one hand to promote good local versus global data
46 management with open data sharing in mind while respecting FAIR principles, and on the other
47 hand to prepare the future EOSC FAIR Web of Data within the framework of the European Open
48 Science Cloud.

49
50 **Keywords** : FAIR, Data management, Metadata, Interoperability, Crosswalk, Controlled
51 vocabulary, Ontologies, Thesaurus, Semantic artefacts

52

53

54 **Background**

55 In the realm of scientific research, metadata plays a key yet often overlooked role. Despite their
56 crucial importance for the discovery, reporting, and mobilisation of research datasets, metadata
57 remains insufficiently known within scientific communities. Yet being data themselves, metadata
58 have to be managed with the same level of rigour as the data produced and consumed by
59 research processes. This lack of awareness persists at a time when the sharing of research data
60 has emerged as a cornerstone of open or at least reproducible science initiatives. As the scientific
61 landscape increasingly emphasises transparency and collaboration, understanding the
62 significance of metadata becomes imperative [1, 2].

63

64 Producing comprehensive metadata is not a task to be taken lightly. It requires time, effort and
65 expertise. Data producers tasked with generating datasets may be reluctant if they see no tangible
66 return on investment in creating metadata [3]. To overcome this hurdle, proactive efforts are
67 required to raise awareness among data producers about the benefits of open data practices [4].
68 However, crafting metadata poses challenges beyond incentivization. Data Management Plans
69 (DMPs), which outline strategies for managing research data throughout their lifecycle, often pose
70 non-trivial questions for data producers. These questions may be time-consuming or complex,
71 particularly when datasets span diverse scientific domains and require input from individuals with
72 varied skill sets. Consequently, collaborative efforts involving domain experts, data managers,
73 and information specialists are essential for navigating the intricacies of DMPs effectively,
74 furthermore when projects are involving multiple partners (e.g. [5]). The sheer diversity of
75 research data and possible dimensions - i.e., the type of characteristics they describe - further
76 complicate metadata management [6]. From omics data to images to experimental data tables,
77 the spectrum of data types is vast and multifaceted.

78

79 Given the complexity of the matter, it is suitable to differentiate between various types and
80 functions of metadata. While not delving into every category, we can simplify them into two main
81 groups: general metadata and specialised metadata. Within the latter category, we encounter
82 structural metadata, which serve to depict the organisation, arrangement, and interconnections
83 within a dataset. For instance, when considering different types of data such as experimental data
84 tables, it becomes obvious that structural metadata are essential for optimising their utility [7].
85 Conversely, general metadata (descriptive, administrative, rights) apply to all data types
86 generated within studies with similar experimental contexts or even an entire project. The
87 subsequent sections of this article will focus on these general metadata.

88
89 How can we collect such metadata while ensuring that they ultimately meet the requisite criteria
90 for interoperability? Indeed, standardisation is the key towards interoperability and consistency in
91 metadata practices. Sustainable metadata have to adhere to established standards and be
92 described using controlled vocabularies endorsed broadly by the scientific community [8].
93 However, the responsibility for metadata creation predominantly falls upon data producers who
94 possess intimate knowledge of the data intricacies. This is challenging as data producers may
95 lack familiarity with metadata standards and best practices, and so reinforces the importance of
96 roles within the data management ecosystem. Data managers and data stewards, equipped with
97 expertise in applying FAIR (Findable, Accessible, Interoperable, Reusable) principles [9] and
98 metadata standards, play a key role in guiding metadata creation and management processes.
99 Conversely, scientists, serving as data producers, possess deep domain knowledge essential for
100 contextualising and enriching metadata. Recognizing the complementary of these roles,
101 collaborative partnerships between data managers and scientists are indispensable for ensuring
102 the effective and sustainable management of research data [10].

103
104 Attempting to find a one-size-fits-all data warehouse capable of accommodating every data type
105 proves to be a futile endeavour. Invariably, some data types remain unsupported or inadequately
106 represented within existing data repositories. To address this challenge, we propose a pragmatic
107 approach that involves managing data directly from storage spaces and then depositing them
108 when the time comes into repositories tailored to requirements of each data type. To this end, we
109 have developed Maggot (Metadata Aggregation on Data Storage), a specialised tool for
110 aggregating metadata into data storage. It is specifically designed for annotating datasets by
111 generating metadata files to attach to the data storage. It has been designed to help data
112 managers in solving as many of the aforementioned challenges as feasible, all while catering to
113 the needs of data producers with minimal training on their part.

114
115

116 **Design Considerations**

117 The development and implementation of Maggot followed a structured approach, involving
118 multiple steps and actors (**Fig. 1**). These steps encompass in particular the identification of
119 metadata fields, terms, vocabularies, and standards. Inspiration for this approach was drawn from
120 an online document detailing the implementation of a descriptive metadata plan [11]. An approach

121 with some similarities has been adopted with the FAIR-DS application [12], dedicated to
122 nucleotide sequences.

123
124 Within any research team, collaboration between the data manager and data producers is
125 essential to select and customise the minimum metadata required and an associated metadata
126 schema suitable for the specific scientific domain. While this process may present challenges, it
127 is crucial to meticulously construct and adapt the schema to align with existing data. Rather than
128 mandating data conformity, the schema should be flexible to accommodate pre-existing data.
129 Thus, the adoption of a schema, such as the one implemented within the Harvard Dataverse
130 software [13] (<https://dataverse.harvard.edu>), followed by an iterative and progressive
131 adjustment, is the approach embraced by Maggot. Indeed, Harvard Dataverse itself is built upon
132 the standard DDI (Document, Discover and Interoperate) metadata schema
133 (<https://ddialliance.org>), which has been expanded to accommodate its requirements. The
134 advantage of the DDI schema is that it encompasses a wealth of general information for
135 describing datasets of any type.

136
137 In the same way, Maggot advocates an iterative and progressive approach regarding the
138 management of controlled vocabulary. Recognizing the impossibility of achieving exhaustiveness
139 in the initial stages, Maggot facilitates a process of continuous improvement. This involves starting
140 with a simple vocabulary dictionary sourced locally and consolidating community-used vocabulary
141 within the related scientific domain. Subsequently, consideration could be given to the creation of
142 a thesaurus (or at least a controlled vocabulary), with or without mapping to existing ontologies.
143 Maggot is seamlessly based on the SKOSMOS web application [14] (<https://skosmos.org>) to
144 query thesauri directly, streamlining the process. Furthermore, ontologies can be chosen
145 progressively by selecting those which are truly relevant for the collective and by drawing up an
146 understandable landscape of the context in which they fit. In the same way for thesauri, Maggot
147 offers the opportunity to enrich metadata using ontologies seamlessly accessible through the
148 OntoPortal web applications such as BioPortal [15] (<https://biportal.bioontology.org>) or
149 AgroPortal [16] (<https://agroportal.lirmm.fr>).

150
151 In our view, creating metadata should not compel data producers to engage in training in topics
152 or concepts outside their expertise, such as FAIR principles, the semantic web, or metadata
153 schema, unless they choose to do so. The data manager needs to recognize that data producers
154 may not possess extensive knowledge of data management practices, especially in the realm of
155 open science and FAIR data. Therefore, pedagogy should be prioritised by refraining from
156 overwhelming data producers with technical details specific to their field. Instead, the data
157 manager should focus on raising awareness and encouraging data producers to improve the
158 quality and reusability of their data [3]. This includes providing guidance on relevant metadata
159 and controlled vocabulary within their scientific domain, as well as training on best practices such
160 as the use of permanent identifiers like DOI, ORCID, RoR, ePICs, Handle and other well
161 recognized identifying systems. Furthermore, data producers should be informed about the
162 selection of appropriate licences, such as CC-BY licences (<https://creativecommons.org>). One
163 of the features of Maggot is precisely to allow the data manager to document each of the terms,
164 providing examples and links for additional information if necessary. This contextual online

165 assistance is thus accessible during entry to allow data producers to fill in each field in the most
166 relevant way, ensuring optimal support throughout the process.

167

168 For downstream metadata management, Maggot provides functionality enabling transformation
169 of metadata in two ways: to targeted data repositories with prerequisites as defined by the Core
170 Trust Seal [17] (<https://www.coretrustseal.org>), or to an export format suitable for harvesting data
171 by third-party applications via an application programming interface (API). These functionalities
172 are built upon a metadata crossing approach based on the mapping defined upstream by the data
173 manager. They ensure compatibility with systems allowing content indexing, thus aligning with
174 the FAIR principles. Adopting this approach improves organisations' data-management practices
175 by effectively using metadata throughout the data lifecycle and facilitates data linkage. It
176 enhances data interoperability and reusability, optimising the value derived from data assets. It
177 also provides practical accessibility through the Web of FAIR Data - i.e. data which meet FAIR
178 principles - by increasing data linkage possibilities, as envisioned by international consortiums
179 like the European Open Science Cloud (EOSC, <https://open-science-cloud.ec.europa.eu>).

180

181 Finally, let us emphasise that a software application rarely meets a need alone but is part of a
182 more global approach, involving several roles. Namely, the data manager is the person who
183 defines the data policy, i.e., its implementation and governance, while data stewards are
184 responsible for data quality, and therefore have a role in data curation. Unavoidably, the
185 development and dissemination of metadata, involving a metadata-centric culture, underline the
186 need for ongoing training initiatives and data stewards are pivotal in educating data producers.
187 Despite advancements in tool intuitiveness and automation, the growing complexity and scale of
188 data across the Web of Data call for an increase in the number of skilled data stewards within
189 research organisations to ensure maximum data reusability [18].

190

191

192 **Results**

193 Maggot (Metadata Aggregation on Data Storage) was developed to address the need for a multi-
194 purpose data management tool capable of supporting the wide data diversity range within a
195 collective. Its primary objectives are to provide visibility into the collective's data assets, facilitate
196 better data description and increase early adoption of FAIR principles. It contributes to guarantee
197 the sustainability of data reusability, especially for those produced by fixed-term personnel
198 (limited-term contract, doctoral students and postdoctoral fellows). It helps to raise awareness
199 among newcomers and students about the importance of robust data description practices which
200 is crucial for fostering a culture of data management excellence [3].

201

202 Maggot enables users to achieve these objectives through various features and functionalities
203 which can be divided into three parts: creation, sharing, dissemination (**Fig. 2**). Maggot supports
204 by default the common metadata standards of Harvard Dataverse (based on DDI) serving as a
205 starting point which can be extended/specialised to suit individual contexts. It then offers
206 scalability and flexibility to enrich the core metadata to adapt any experimental context. Controlled
207 vocabulary management is another key aspect of Maggot, offering options ranging from simple

208 dictionaries to ontologies as discussed above. The tool includes enrichment functionalities of
209 existing resources (e.g., dictionary editing, adding additional ontologies), ensuring that users can
210 effectively manage and utilise controlled vocabularies relevant to their data. To this end, it offers
211 great flexibility in configuration, allowing organisations to tailor the tool to their specific needs and
212 requirements (**Fig. 3**).

213
214 Implementing a data management plan (DMP) entails certain prerequisites, including data
215 externalisation to preserve them outside users' disk space. This ensures data is secured in one
216 location and serves as an initial backup, which becomes particularly crucial when fixed-term
217 personnel are involved in data production. Consequently, considerations arise regarding the
218 organisation of storage spaces, such as harmonising folder and file naming conventions, setting
219 up folder structure, and using README files to provide essential information. Maggot precisely
220 answers these DMP challenging questions related to organising, documenting, storing and
221 sharing data from various sources. In our approach the data storage space becomes a local
222 reference data repository, mitigating the risk of data duplication or divergence to another medium.
223 Then, only metadata need to be added to this centralised space, streamlining data management
224 processes, and enhancing efficiency. Indeed, once the metadata file is generated in JSON format,
225 it has to be placed in the storage space reserved for this purpose alongside the corresponding
226 dataset. This metadata file can be seen as a README file adapted to machines (**Additional file**
227 **1**), but still readable by humans. In contrast, with an internal structure, it offers better coherence
228 and consistency of information than a simple README file with a completely free and therefore
229 unstructured text format. In this way, the storage space becomes a data asset which can therefore
230 be efficiently leveraged using metadata. Indeed, all the JSON metadata files are scanned and
231 parsed according to a fixed time interval (30 min) then loaded into a database. This allows users
232 to query datasets based on metadata filters. The search form, in a compact shape, is almost the
233 same as the entry form. Matching datasets are returned as a list, and for each of them a provided
234 link helps to access the detailed metadata.

235
236 Metadata entry can be initiated at the outset of a project or study, without requiring the completion
237 of all data acquisition or processing, or waiting until the data need to be published. Indeed, the
238 ability to reload a metadata file facilitates gradual and iterative metadata addition across the
239 project, thereby spanning the research data lifecycle to the greatest extent possible. Maggot
240 supports the input of both descriptive and administrative metadata for any type of data, including
241 datasets, images, sequences, and more, with customizable field definitions to suit diverse user
242 requirements. Moreover, Maggot emphasises ease of use and adaptability. It offers guided
243 assistance through drop-down menus and vocabulary lists featuring autocompletion, greatly
244 speeding up the process of filling in numerous descriptive metadata. Crucially, Maggot does not
245 restrict the choice of data repository, ensuring compatibility with currently supported platforms
246 knowing that others may be supported in the future (e.g. Dryad [19], RO-Crate [20]). This also
247 does not prejudge the use of metadata. It is entirely possible, for example, to set up an internal
248 metadata harvesting process to automatically fill in another data source (e.g., FAIRDOME-SEEK
249 data management platform [21]). It is essential to highlight that opting for Maggot to generate
250 metadata does not confine the data to an isolated silo. In case one day the Maggot tool was no
251 longer supported, all metadata will persist in disk space in a format accessible to both humans

252 and machines. This ensures that future applications/services are able to continue to use legacy
253 metadata and therefore warranty data reuse. For this purpose, Maggot enables data scientists or
254 data repositories to harvest data. The OAI-PMH (Open Archives Initiative Protocol for Metadata
255 Harvesting, <https://www.openarchives.org>) allows for listing all datasets based on the DublinCore
256 schema (<https://www.dublincore.org>), while the metadata of each dataset can be harvested in
257 JSON-LD format (JSON for Linking Data, <https://json-ld.org>), mainly adhering to the schema.org
258 standard (<https://schema.org>). This aspect is particularly critical for linking metadata within the
259 realm of linked data, thereby ensuring their interoperability. For instance, we plan in the near
260 future to support DCAT-based harvesting (<https://www.w3.org/TR/vocab-dcat-3/>).

261
262 Maggot also provides a solution to data fragmentation. Indeed, data is often scattered across
263 various platforms, databases, and file formats, making it challenging to locate and access.
264 Moreover, non-standardized metadata and inconsistent data organisation hinder effective data
265 discovery and reuse. Therefore, Maggot allows data producers to specify resources, i.e., data in
266 the broader sense, whether external or internal, to centralise all links towards data (**Fig. 2**).
267 External resources must be specified by an URL with preference for a permanent identifier (e.g.,
268 DOI) but also any URL pointing to data whether they comply with the FAIR principle or not.
269 Furthermore, in the case of local data management, it is wise to indicate in which space the data
270 is located if it is not located in the same place as metadata (e.g., NAS unit or data cloud). Maggot
271 can thus become a data hub by gathering all references to several data sources in one place at
272 hand.

273
274 While the primary focus is managing metadata linked to data stored within a given collective,
275 Maggot also facilitates data openness through metadata, especially in projects funded by public
276 institutions. By setting up metadata schemas that facilitate crosswalks with established data
277 repositories, users can seamlessly push metadata along with the corresponding data without the
278 need for additional data entry. This promotes data openness and accessibility in accordance with
279 international standards and community norms. Such functionality empowers organisations to
280 share their data with external stakeholders while ensuring consistency and interoperability.
281 Maggot thus offers a comprehensive and open solution for metadata management, catering to
282 the diverse requirements of organisations and promoting best practices in data description and
283 dissemination.

284
285

286 **Implementation and Documentation**

287 The deployment of Maggot requires two infrastructure components: a dedicated server for the
288 web application and a designated data storage space. Regarding the server, it must be capable
289 of running an operating system compatible with Linux. In addition, it should support
290 containerization using Docker. This latter aspect offers a simplified approach to installation and
291 administration, but also ease of use and flexibility. Regarding data storage, any technology is
292 suitable. Data storage can be local (e.g., NAS unit) or remote (e.g., data cloud). Successful tests
293 have been performed by implementing a server on our institutional data center and data storage

294 on another data center. Access to the storage space can easily be done using the rclone tool
295 (<https://rclone.org>), a real Swiss army knife for disk sharing.

296
297 Maggot is a web-based PHP application as a front-end using MongoDB
298 (<https://www.mongodb.com>) to index all scanned metadata from disk storage every 30 minutes.
299 Moreover, Maggot mobilises various vocabularies (thesauri and ontologies), most of which are in
300 remote resources. So the utilisation of APIs plays a significant role, particularly for integrating
301 these vocabularies. This extensive use of APIs facilitates real-time imports, thus reducing the
302 need for pre-updating information.

303
304 Documentation is available via <https://inrae.github.io/pgd-mmdt/> and from within the app. It
305 includes technical information on how to configure Maggot, but also a quick overview of how to
306 use it. For data managers, it explains in detail how to construct the terminology with the associated
307 vocabularies.

308
309

310 **Conclusion**

311 Maggot is specifically designed to annotate datasets by generating metadata files to be linked
312 into storage spaces, to tackle challenges related to data organisation, documentation, storage,
313 and frictionless FAIR metadata sharing within the collaborative group and beyond. Indeed,
314 Maggot meets the Open Data requirements beyond the simple provision of data with unlimited
315 access. This essentially implies: i) to ensure search and access to metadata that define data
316 access and usage conditions, and ii) to foster metadata and data interoperability to break down
317 silos, highlighting the necessity of embracing FAIR principles even when complete openness is
318 not achievable.

319 By covering as much of the research data lifecycle as possible, Maggot ensures effective and
320 sustainable research data management and significantly simplifies the adoption of FAIR principles
321 thereby empowering organisations to elevate the value and usability of their own data assets.
322 Moreover, its ability via crosswalk approaches to distribute metadata based on standard schemas
323 while being machine-readable, expands the toolbox needed to prepare the future EOSC FAIR
324 Web of Data within the framework of the European Open Science Cloud.

325
326

327 **Availability of Source Code and Requirements**

- 328 ● Project name: Maggot
- 329 ● Project homepage: <https://pmb-bordeaux.fr/maggot/>
- 330 ● Project code repository: <https://github.com/inrae/pgd-mmdt>
- 331 ● Documentation: <https://inrae.github.io/pgd-mmdt/>
- 332 ● Operating system(s): Platform independent
- 333 ● Programming languages: PHP, python, javascript
- 334 ● Licence: GNU GPL v3

- 335 • [Maggot, RRID: SCR_025261](#)
- 336 • [Biotoools: https://bio.tools/maggot](https://bio.tools/maggot)

337

338

339 **Abbreviations**

340 API: Application Programming Interface; DDI: Document, Discover and Interoperate; DMP: Data
341 Management Plan; EOSC: European Open Science Cloud; FAIR: Findable, Accessible,
342 Interoperable, Reusable; JSON: JavaScript Object Notation; JSON-LD: JSON for Linking Data;
343 OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting; NAS: Network Attached
344 Storage;

345

346

347 **Competing interest**

348 The authors declare that they have no competing interests.

349

350

351 **Funding**

352 D.J. was partly supported by the MetaboHUB project funded by the French National Research
353 Agency (ANR-11-INBS-0010) and by the WAPNMR project funded by the French National
354 Research Agency (ANR-21-CE21-0014). D.J., F.E., C.M.N. and J.T. were partly supported by
355 the Bordeaux Plant Science (BPS) project funded by the Université de Bordeaux. D.J., F.E,
356 C.M.N, J.T and P.C were partly supported by the French National Research Institute for
357 Agriculture, Food and the Environment (INRAE). R.D. was partly supported by the European
358 Research Infrastructure on Highly Pathogenic Agents (ERINHA AISBL).

359

360

361 **Authors' Contributions**

362 Conceptualization: D.J, F.E, PC.; funding acquisition: D.J, F.E.; methodology: D.J., F.E, R.D.;
363 software: D.J, F.E.; writing—original draft: D.J., R.D.; writing—review and editing: All authors. All
364 authors read and approved the final manuscript.

365

366

367 **Acknowledgments**

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

372

373

374 Additional files

375

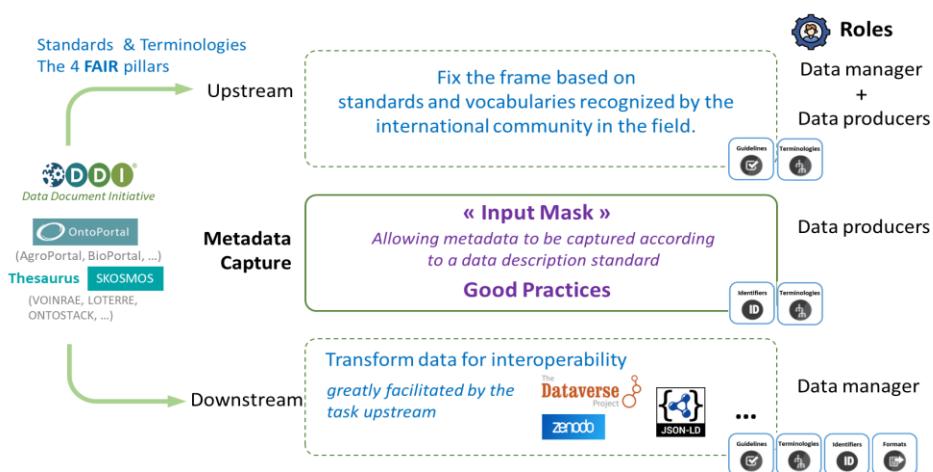
376 **Additional file 1** : Examples of metadata files along with corresponding definitions files within
377 an [Excel workbook](#).

378

379

380 Figures

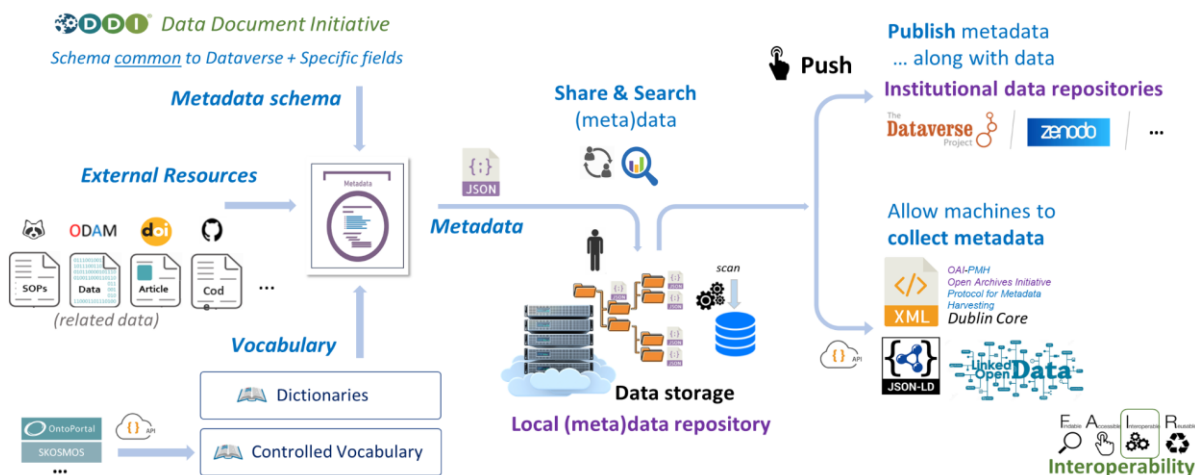
381



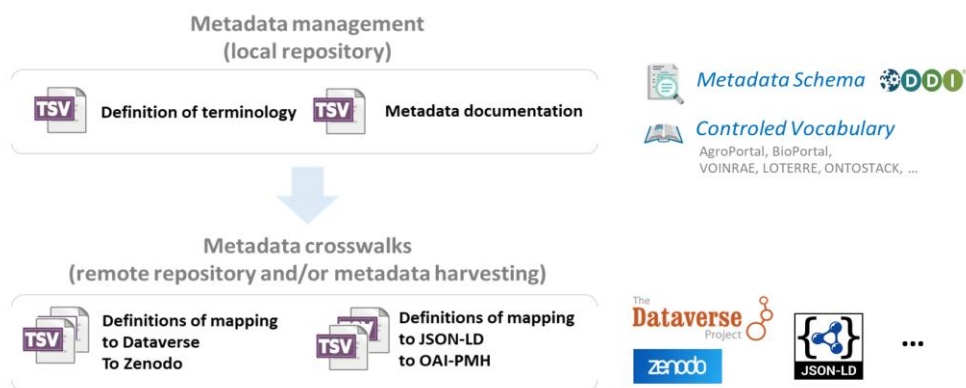
382

383 **Figure 1:** Development and implementation of Maggot tool structured into three steps, namely before,
384 during and after metadata capture: i) Upstream, collaboration between the data manager and the data
385 producers was essential to select and customise a flexible metadata schema adapted to the scientific
386 domain as well as the identification of terms and vocabularies (dictionaries, thesauri, ontologies). Therefore,
387 Maggot proposes the Dataverse schema serving as a fundamental model, itself based on the standard DDI
388 metadata schema. ii) For metadata entry, data producers must be trained on good practices such as the
389 proper use of permanent identifiers or the choice of licences. iii) Downstream, data can easily be pushed
390 into a support data repository without any addition or can be harvested based on a dedicated protocol (OAI-
391 PMH). JSON-LD format is also supported for linking metadata within the realm of linked data, thereby
392 ensuring their interoperability. The complementarity of roles between data manager and data producers
393 ensures effective and sustainable research data management.

394



395
 396 **Figure 2:** Main functionalities of Maggot split into three parts: creation, sharing, dissemination. First,
 397 producing a document with metadata sets of data within a collective of people, thus allowing i) to answer
 398 certain questions of the Data Management Plan (DMP) concerning data organisation, documentation,
 399 storage and sharing in the data storage space, ii) to meet certain data and metadata requirements, listed
 400 for example by Open Research Europe in accordance with FAIR principles. Next, searching for datasets
 401 by their metadata. Indeed, the descriptive metadata thus produced can be associated with the
 402 corresponding data directly in the storage space and then it is possible to perform a search on the metadata
 403 to find one or more sets of data. Only descriptive metadata is accessible by default. Finally, publishing the
 404 metadata of the datasets as well as their data files in a European-approved repository, with the possibility
 405 either to directly harvest the metadata via the OAI-PMH protocol, or to export the associated metadata with
 406 their semantic context for full interoperability.
 407
 408



409
 410 **Figure 3 :** Maggot tool flexibility in configuration. Maggot allows users to choose all the metadata describing
 411 their data with two levels of definition files. The first level concerns the definition of metadata similar to a
 412 descriptive metadata plan. This category is more akin to configuration files, and constitutes the heart of the
 413 configuration around which everything else is based. The input and search interfaces are completely
 414 generated from these definition files, thus defining each of the fields, their input type and the associated
 415 controlled vocabulary. The second level concerns the definitions of the mapping to a differently structured
 416 metadata schema (metadata crosswalk, i.e a specification for mapping one metadata standard to another),

417 used either for metadata export to a remote repository (e.g. Dataverse, Zenodo) or for metadata harvesting
418 (e.g. JSON-LD, OAI-PMH).

419
420

421 **References**

422

- 423 1. Ulrich H, Kock-Schoppenhauer A, Deppenwiese N, et al. Understanding the Nature of Metadata:
424 Systematic Review. *J Med Internet Res*. 2022;24(1):e25440. <https://doi.org/10.2196/25440>
- 425 2. Manninen L. Describing Data: A Review of Metadata for Datasets in the Digital Commons
426 Institutional Repository Platform: Problems and Recommendations. *J Libr Metadata*. 2018.
427 <https://doi.org/10.1080/19386389.2018.1454379>
- 428 3. David R, Mabile L, Specht A, et al. FAIRness Literacy: The Achilles' Heel of Applying FAIR
429 Principles. *CODATA Data Sci J*. 2020. <https://doi.org/10.5334/dsj-2020-032>
- 430 4. Popkin G. Data sharing and how it can benefit your scientific career. *Nature* 2019;569:445-447.
431 <https://doi.org/10.1038/d41586-019-01506-x>
- 432 5. David R, Baumann K, LeFranc Y, et al. Converging on a Semantic Interoperability Framework for the
433 European Data Space for Science, Research and Innovation (EOSC), 2nd Workshop on Ontologies
434 for FAIR and FAIR Ontologies. Sherbrooke, Québec (Canada). 2023a.
435 <https://doi.org/10.5281/zenodo.8102786>
- 436 6. Çaldağ MT, Gökalp E. Understanding barriers affecting the adoption and usage of open access data
437 in the context of organisations. *Data Inform Manag*. 2023. <https://doi.org/10.1016/j.dim.2023.100049>
- 438 7. Jacob D, David R, Aubin S, Gibon Y. Making experimental data tables in the life sciences more
439 FAIR: a pragmatic approach. *GigaScience* 2020;9(12):giaa144.
440 <https://doi.org/10.1093/gigascience/giaa144>
- 441 8. David R, Richard AS, Connellan C, et al. Umbrella Data Management Plans to integrate FAIR data:
442 lessons from the ISIDORE and BY-COVID consortia for pandemic preparedness. *Data Sci J*.
443 2023b;22:35, pp. 1–15. <https://doi.org/10.5334/dsj-2023-0350>
- 444 9. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data
445 management and stewardship. *Sci Data* 2016;3:160018. <http://dx.doi.org/10.1038/sdata.2016.18>
- 446 10. David R, Rybina A, Burel J-M, et al. “Be sustainable”: EOSC-Life recommendations for
447 implementation of FAIR principles in life science data handling. *EMBO J*. 2023c. e115008.
448 <https://doi.org/10.15252/emboj.2023115008>
- 449 11. Bingo S, Montenegro M. How to Create a Descriptive Metadata Plan. Sustainable Heritage Network.
450 Accessed April 13, 2024. [Sustainable Heritage Network](https://www.sustainableheritagenetwork.org/)
- 451 12. Nijssse B, Schaap P, Koehorst J. FAIR data station for lightweight metadata management and
452 validation of omics studies. *GigaScience* 2023(12):giad014,
453 <https://doi.org/10.1093/gigascience/giad014>
- 454 13. King G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociol*
455 *Methods Res*. 2007;36(2):173-199. <https://doi.org/10.1177/0049124107306660>
- 456 14. Suominen O, Ylikotila H, Pessala S, et al. Publishing SKOS vocabularies with Skosmos. *Manuscript*
457 *submitted for review*. June 2015.
- 458 15. Noy NF, Shah NH, Whetzel PL, et al. BioPortal: ontologies and integrated data resources at the click
459 of a mouse. *Nucleic Acids Res*. 2009;37(suppl_2):W170–W173. <https://doi.org/10.1093/nar/gkp440>
- 460 16. Jonquet C, Toulet A, Arnaud E, et al. AgroPortal: A vocabulary and ontology repository for
461 agronomy. *Comput Electron Agric*. 2018;144:126-143. <https://doi.org/10.1016/j.compag.2017.10.012>

- 462 17. CoreTrustSeal Standards and Certification Board. CoreTrustSeal Trustworthy Digital Repositories
463 Requirements 2023-2025 Extended Guidance (V01.00). Zenodo. 2022.
464 <https://doi.org/10.5281/zenodo.7051096>
- 465 18. Mons B. Invest 5% of research funds in ensuring data are reusable. *Nature* 2020;578:491.
466 <https://doi.org/10.1038/d41586-020-00505-7>
- 467 19. Isard M, Budiu M, Yu Y, et al. Dryad: distributed data-parallel programs from sequential building
468 blocks. *ACM SIGOPS Oper Syst Rev.* 2007;41(3). <https://doi.org/10.1145/1272998.1273005>
- 469 20. Soiland-Reyes S., Sefton P., Crosas M. et al. Packaging research artefacts with RO-Crate, *Data*
470 *Science* 5(2) <https://doi.org/10.3233/DS-210053>
- 471 21. Wolstencroft K, Owen S, Krebs O, et al. SEEK: A systems biology data and model management
472 platform. *BMC Syst Biol.* 2015;9:33. <http://doi.org/10.1186/s12918-015-0174-y>
473