




Collaborative creation and management of rich FAIR metadata: Two case studies from robotics field research


Christian Backe ¹Veit Briken ¹Atefeh Gooran Orimi ²Rayen Hamlaoui ²Malte Wirkus ¹Bilal Wehbe ¹Frank Kirchner ¹

1. Robotics Innovation Center (RIC), Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Bremen, Germany.

2. Institut für Produktentwicklung und Gerätebau (IPeG), Leibniz Universität Hannover (LUH), Hannover, Germany.

**Date Submitted:**

2025-01-19

Licenses:This article is licensed under: **Keywords:**

Research Data Management, Data Lifecycle, Metadata, Field Data, FAIR, Robotics

Data availability:

[RoBivaL data corpus](#),
[Sonar-to-RGB Image Translation for Diver Monitoring in Poor Visibility Environments](#),
[Fusion of Underwater Camera and Multibeam Sonar for Diver Detection and Tracking](#),
[Metadata of a Large Sonar and Stereo Camera Dataset Suitable for Sonar-to-RGB Image Translation](#)

Software availability:**Corresponding Author:**Christian Backe
christian.backe@dfki.de**Abstract.**

This paper presents lessons learned from the creation and management of FAIR (Findable, Accessible, Interoperable, Reusable) data and metadata in two recent robotics projects, in order to derive principles and building blocks for collaborative (meta)data management in field research. First, an inventory of metadata purposes and topics is presented, distinguishing between executive metadata necessary for data producers, and rich reusable metadata satisfying the FAIR principles. A model of the metadata creation process is developed and compared with the Metadata4Ing ontology. Second, social aspects of FAIR research data management (RDM) are discussed in the project context and beyond. The primary tasks of a FAIR research data manager are analyzed in three domains: data production team, research domain, and FAIR RDM community. Third, some improvements on prominent data lifecycle models are proposed to support the requirements of collaborative RDM, and to foster an iterative improvement of RDM systems.

1 Introduction**2 1.1 Motivation**

3 The FAIR principles are formulated in generic terms and without reference to any particular
4 scientific discipline [1]. According to Jacobsen et al. (2020), "[t]his has likely contributed to
5 [their] broad adoption [...], because individual stakeholder communities can implement their
6 own FAIR solutions. However, it has also resulted in inconsistent interpretations that carry the
7 risk of leading to incompatible implementations. Thus, [...] for true interoperability we need to
8 support convergence in implementation choices that are widely accessible and (re)-usable." [2]
9 Communities who wish to develop their own FAIR conventions may greatly benefit from

10 successful real-world examples that allow researchers and practitioners to identify, evaluate, and
11 select best practices for research data management (RDM). However, the application of RDM in
12 practical applications is still lacking in literature. Many approaches and models exist in theory,
13 but their value to the community remains low if their practice is not systematically examined.

14 This paper presents lessons learned from the creation and management of FAIR data and metadata
15 in two recent robotics field research projects, RoBivaL and DeeperSense, conducted at the
16 Robotics Innovation Center (RIC) of the German Research Institute for Artificial Intelligence
17 (DFKI) [3] [4]. The paper is an extended and revised version of a presentation given at the
18 NFDI4Ing Conference 2023 [5].

19 We chose to present RoBivaL and DeeperSense together for two complementary reasons. On
20 the one hand, their commonalities allow us to generalize RDM principles to some degree in
21 certain circumstances: Both projects are field studies, conducted over extended periods of time
22 by diverse research teams from multiple institutions and disciplines. Their differences, on the
23 other hand, allow us to examine the application of RDM principles in different research scenarios:
24 RoBivaL is a terrestrial robotics project studying the performances of different hardware systems
25 at a single location in the context of agriculture. DeeperSense is an underwater robotics project
26 developing an AI for the translation of sonar outputs into camera-like images based on training
27 data collected both at a laboratory and at multiple field locations.

28 Openly available datasets are crucial to advancing the field of robotics. One reason is that robotics
29 research often requires access to specialized hardware, sensors, and environments. Making high-
30 quality data accessible to more researchers, including those without the means to collect such
31 data on their own, fosters innovation from a wider range of perspectives. Furthermore, open
32 datasets provide common ground for the community to develop and benchmark algorithms
33 collaboratively on standardized data.

34 Our discussion of RDM, however, is not supposed to be applicable just to robotics. In principle,
35 our findings can be applied to any collaborative (field) research project employing humans and
36 technical systems for data acquisition in multiple steps and iterations. The purpose of our study
37 is to derive requirements and strategies for the creation and management of "rich" metadata in
38 the FAIR sense.

39 1.2 Outline

40 Section 2 presents related work on FAIR RDM in general, on open data in robotics in particular,
41 and on formal knowledge representation in robotics.

42 Section 3 features brief summaries of RoBivaL and DeeperSense. We present their overall project
43 objectives and contrast their base data requirements from a high-level perspective.

44 The main body of the paper is divided into three parts (Sections 4 - 6). Specific desiderata
45 and additional related work are introduced at the beginning of each part if necessary. The
46 models and concepts presented in these parts were extracted from the experience in RoBivaL
47 and DeeperSense and shall be applied in future projects.

48 The first main part (Section 4) discusses the content dimension of FAIR RDM. We distinguish
49 between executive metadata necessary for producers to achieve their project goals, and reusable

50 metadata necessary for reusers to satisfy the FAIR principles. We introduce this distinction into
51 the FAIR data debate, because we believe that data producers are more likely to adopt FAIR
52 principles if the specific needs of producers are taken into account by the FAIR community.
53 Both metadata types are illustrated with examples from RoBivaL and DeeperSense. Further, this
54 section introduces base elements for a model of the metadata creation process at the micro level
55 in the context of collaborative metadata management. We relate our model to the "processing
56 step" class of the Metadata4Ing (M4I) ontology. Though M4I acknowledges the existence of
57 metadata, it does not appear to address the process of metadata creation.

58 The second main part (Section 5) expands the distinction between different stakeholder groups
59 from the previous section and explores the social dimension of collaborative FAIR RDM more
60 broadly. We argue that a FAIR research data manager acts as a link between three social domains
61 where they perform different primary tasks. We are not aware of a discussion about the social
62 implications of FAIR RDM, but we believe such a discussion to be indispensable for a definition
63 of the FAIR data manager role.

64 The third main part (Section 6) examines the time dimension of collaborative and iterative FAIR
65 RDM at the macro level. Based on a critical appraisal of prominent data lifecycle models, we
66 suggest a model of a self-improving data lifecycle geared towards collaborative and iterative
67 RDM. We introduce a data provision phase which is necessary for internal collaboration. Fur-
68 ther, we introduce an evaluation phase at the end of the lifecycle complementing the planning
69 phase at its beginning, to foster iterative improvement of the data management system. Lastly,
70 we recognize that planning and evaluation are different kinds of activities than data creation,
71 provision, processing, publishing, etc., which gives rise to a lifecycle model with two nested
72 loops. Our model is illustrated with lessons learned from RoBivaL and DeeperSense.

73 2 Related work

74 This section presents some recent developments in open data and FAIR RDM with applications
75 in robotics, and on formal knowledge representation in robotics.

76 The need for large-scale, real-world datasets in robotics is highlighted by the rise of Robotics
77 Foundation Models (RFMs) [6]. Extensive multimodal training data can lead to high-level task
78 performance in many different scenarios, as illustrated by the RT model class presented by
79 Brohan et al. (2023) [7]. The RT-1-X model was trained on the Open X-Embodiment dataset,
80 developed and published by Google DeepMind in 2024 in collaboration with over 20 research
81 institutions [8] [9]. This initiative demonstrates the benefits of shared open data. Still, according
82 to Firoozi et al. (2024), the scarcity of robot-relevant training data remains a major open research
83 challenge in the improvement of RFMs [10].

84 Robotics-enabled marine research has seen some advancements towards FAIR RDM. Schoening
85 et al. (2022) observe that published marine image datasets have been lacking metadata to describe
86 their high technical heterogeneity; the authors propose a concept for image FAIR digital objects
87 (iFDOs) as a remedy [11]. Motta et al. (2023) present a method for the creation of FAIR marine
88 robotic telemetry data and metadata about marine robotic missions; they observe a general lack of
89 controlled vocabularies in robotics [12]. In the context of space robotics, Dominguez et al. (2020)

90 developed a modular framework for multisensor data fusion including a suite of data management
91 tools; their approach for describing complex data processing systems might feed into FAIR
92 metadata components for robotics [13]. Arundel et al. (2023) offer a data management case
93 study focused on conveyance of big data over multiple stages; though motivated by geospatial
94 data processing, their methods seem applicable to many domains, including robotics [14].

95 Several initiatives in science and industry are working to represent robotics knowledge using for-
96 mal ontologies and terminologies. Olivares et al. (2019) review five ontology-based approaches
97 to autonomous robotics and quote four additional ontological efforts in robotics that either don't
98 address autonomy or lack relevant qualities [15]. The IEEE 1872 group of standards comprises
99 six ontologies which were released between 2015 and 2024 [16]–[18]. They define more than
100 100 terms addressing general robotics concepts, robot parts, pose, tasks, and autonomy. An
101 additional IEEE ontology about reasoning on multiple robots is in development [19]. A parallel
102 standardization enterprise at the terminological level are the ISO vocabularies for robotics [20]
103 and for mobile robots [21]. Though ISO 8373 is a normative reference in IEEE 1872, there are
104 considerable differences which must be navigated by practitioners. The ROS middleware offers
105 the URDF specification for a formal description of kinematic and dynamic aspects of individual
106 robots that consist of rigid links connected by joints [22]. A proposal to cure some shortcomings
107 and limitations of URDF has not yet been addressed [23]. Considering applications of ontologies
108 and terminologies in robotics research, Jorge et al. (2015) evaluate the POS ontology in a use
109 case where heterogeneous robots and humans interact in a manufacturing task [24]. Neto et al.
110 (2019) apply the CORA ontology in a simulated robotics reconnaissance mission with interaction
111 between autonomous aerial and ground robots [25]. Yüksel (2023) explores the application of
112 ontologies at the robot component level for the automated design of robotic systems in relation
113 to robot tasks and capabilities; the work introduces the Korcut ontology family [26].

114 DeeperSense and RoBivaL did not employ any formal ontologies or terminologies for devel-
115 opment or data management. This is in line with the usual practice in the respective research
116 teams. Applying ontologies on top of the primary project requirements would have posed a major
117 challenge exceeding the available resources. Formal knowledge representation with ontologies
118 and terminologies will therefore be explored in future work.

119 **3 Project summaries**

120 This section gives brief summaries of the projects RoBivaL and DeeperSense, focusing on
121 general project objectives and the base data that was created.

122 **3.1 RoBivaL**

123 The project RoBivaL [3] [27] was conducted between August 2021 and October 2023 by an
124 interdisciplinary and multi-institutional team of roboticists and agriculture researchers in Ger-
125 many. The project compared different robot locomotion concepts both from space research and
126 agricultural applications on the basis of experiments conducted under agricultural conditions.
127 The goal was to promote knowledge and technology transfer between space and agriculture
128 research. While the experiment designs were inspired by the standards ISO 18646-1 [28] and
129 ISO 18646-2 [29], the environmental properties were adapted to the agricultural context, and the

130 main evaluation focus was on soil interaction. Four robots were used: Two having their origins
 131 in space applications, the other two developed for agriculture. The robots were subjected to
 132 six experiments addressing different agricultural challenges and requirements. Soil conditions
 133 were controlled and varied on the two dimensions moisture (dry, moist, wet) and density (tilled,
 134 compacted). Figure 1 gives an impression of selected experiments and robots in the field.

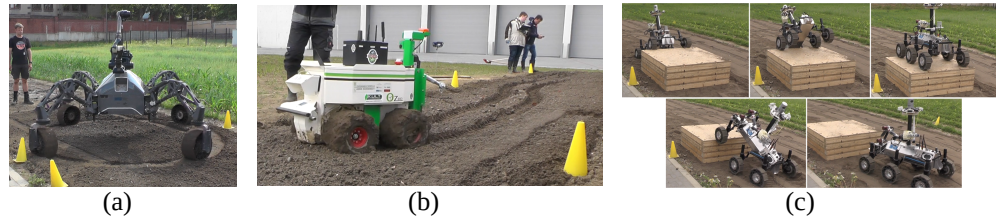


Figure 1: Selected RoBivaL field experiments and robots: (a) Turn around with SherpaTT, (b) Straight travel with Naio Oz, (c) Sill crossing with ARTEMIS. © DFKI, Malte Wirkus. License: CC BY 4.0 International

135 Field conditions and robot behavior were monitored with various sensors and measuring devices,
 136 partly on the robots and partly in the field, in order to document the experiment execution and
 137 to determine the robot performance. The data capturing devices, their roles and deployments
 138 are summarized in Table 1. (Video camera and Lidar on the system are greyed out, because,
 139 although available, they were not used in the project.) The entire dataset including comprehensive
 140 metadata is publicly available on the Zenodo platform [30].

	Device on System	Device on System and in Field	Device in Field
System Monitoring	<ul style="list-style-type: none"> • IMU • Force sensor 	<ul style="list-style-type: none"> • RTK-GPS 	<ul style="list-style-type: none"> • Stopwatch • Compass
System and Field Monitoring			<ul style="list-style-type: none"> • Video camera • Ruler
Field Monitoring	<ul style="list-style-type: none"> • Video camera • Lidar 		<ul style="list-style-type: none"> • Tilt laser scanner • Penetrometer • Moisture meter

Table 1: RoBivaL data capturing devices by purpose and deployment

141 3.2 DeeperSense

142 The project DeeperSense [4] [31] was conducted between January 2021 and December 2023 by
 143 an international, interdisciplinary, and multi-institutional team of researchers and domain experts
 144 in Germany, Spain, and Israel. This paper focuses on the German use case, which employed
 145 roboticists, sensor experts, and technical divers. The objective was to improve the safety of the
 146 divers, who work under dangerous conditions and therefore require constant monitoring and
 147 assistance. Existing safety systems rely on cameras, which is a problem in turbid water that limits
 148 visibility – just when the divers most need outside support. Sonars are more robust to turbidity,
 149 but conventional sonar output is difficult to interpret. DeeperSense therefore developed a neural
 150 network which translates sonar output into images that appear camera-like, thus combining the
 151 best aspects of both modalities. Figure 2 illustrates the sonar-to-image translation.

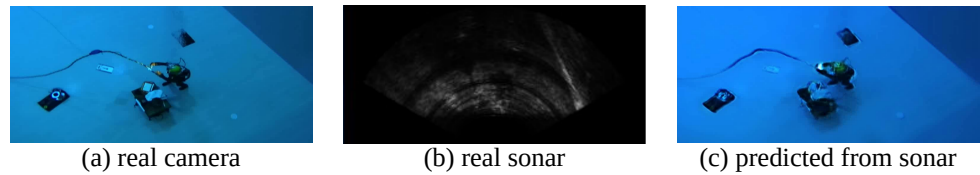


Figure 2: DeeperSense sonar-to-camera translation. The pair (a) and (b) is a training input sample. The generated image (c) is predicted in production from (b) alone. © DFKI, Bilal Wehbe. License: CC BY 4.0 International

152 To gather training data, divers performing typical work tasks were recorded underwater with
 153 sonar and camera simultaneously. Figure 3 shows the training data collection setup schematically.

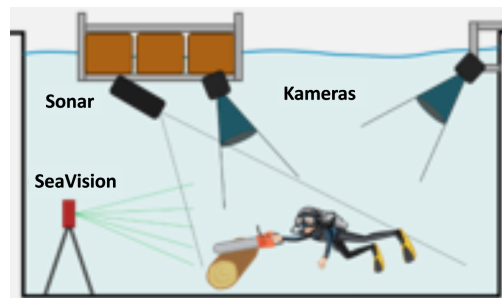


Figure 3: DeeperSense training data collection setup. © DFKI, Bilal Wehbe. License: CC BY 4.0 International

154 For the neural network to be able to handle different types and degrees of turbidity, the training
 155 data had to be varied accordingly. Since this is difficult to establish and control efficiently at
 156 a single time and location, data was captured during six sessions at four different locations,
 157 covering inside and outside conditions, natural and artificial water bodies, and different seasons.
 158 Figure 4 gives an impression of the field locations.



Figure 4: DeeperSense field locations. From left to right: Martime Exploration Hall, Bremen; Chalk Lake, Hemmoor; Tank Wash Basin, Neu-Ulm; Starnberg Lake, Percha. © DFKI, Bilal Wehbe / Christian Backe. License: CC BY 4.0 International

159 Selected parts of the sensor data were published on the Zenodo platform [32] [33]. Due to the
 160 size of the sensor data, it is currently impractical to make the entire corpus available online.
 161 Instead, the metadata was published as a standalone database [34], allowing researchers to select
 162 portions relevant for their use cases, which are made available on demand. This is an effort to
 163 comply with the FAIR principle A2 to make metadata accessible independently of the base data.
 164

165 3.3 Comparison

166 Table 2 summarizes and contrasts the data-related properties of RoBivaL and DeeperSense as
 167 presented in Sections 3.1 and 3.2. An immediate takeaway is that the scope and form of the base
 168 data, its purpose and handling can be quite different between projects even at a single institute.
 169 A data management solution should be flexible enough to accommodate such variance.

	RoBivaL	DeeperSense
Objective and method	<ul style="list-style-type: none"> • Compare robot performances using statistics and visualization 	<ul style="list-style-type: none"> • Train neural network using multimodal machine learning
Base data	<ul style="list-style-type: none"> • Ten distributed sensor outputs and manual measurements observing robot behavior and field characteristics before, during, and after an experiment 	<ul style="list-style-type: none"> • Indefinite stream of synchronized, co-located camera and sonar snapshots showing divers working underwater
Data acquisition in the field	<ul style="list-style-type: none"> • Multiple sessions at one location • Deliberate field preparation 	<ul style="list-style-type: none"> • Multiple sessions and locations • Adapt to given field conditions

Table 2: Summary of data-related project properties of RoBivaL and DeeperSense

170 4 Content dimension: Executive metadata and rich reusable metadata

171 This section discusses the content dimension of metadata creation and management in RoBivaL
 172 and DeeperSense from the perspectives of data producers on the one hand, and potential reusers
 173 as characterized by the FAIR principles on the other. Subsection 4.1 introduces necessary
 174 background about high-level purposes of metadata, metadata semantics in the context of robotics
 175 and engineering in general, and the concept of metadata "richness" according to the FAIR
 176 principles. Subsection 4.2 lays out a collection of metadata topics from RoBivaL and DeeperSense
 177 for different purposes, and divides it into executive metadata relevant for producers and reusable
 178 metadata for public consumers based on the different motives of both parties. This analysis
 179 foreshadows the discussion of social aspects of metadata management in Section 5. Subsection
 180 4.3 attempts to model the process of metadata creation abstractly and at the micro level for
 181 production purposes. We illustrate our model with examples from DeeperSense and RoBivaL,
 182 and compare it to the communication-oriented "processing step" class from the Metadata4Ing
 183 (M4I) ontology.

184 4.1 Metadata purposes, semantics, and richness

185 Virtually every general metadata definition starts with the assertion that metadata is "data about
 186 data" [35]–[39]. A common purpose-based classification distinguishes at a high level between
 187 descriptive, administrative, and structural metadata [40]–[42]: Descriptive metadata "enables
 188 discovery, identification, and selection of resources", administrative metadata "facilitates the
 189 management of resources", structural metadata "describes relationships among various parts of a
 190 resource", and is "generally used in machine processing" [42].

191 There are many domain-specific approaches to model metadata semantics. For the communica-
 192 tion of metadata in engineering disciplines including robotics, the NFDI4Ing community has

193 developed the Metadata4Ing (M4I) ontology [43]–[45]. It features a generalized process model,
 194 centered around the "processing step" class. This is an attempt at communicating multi-stage
 195 data processing to satisfy the FAIR principle R1.2 of detailed provenance tracking. We compare
 196 the M4I processing step class with our own metadata creation process model in Section 4.3.

<ul style="list-style-type: none"> • F1. (meta)data are assigned a globally unique and persistent identifier • F2. data are described with rich metadata (defined by R1 below) • F3. metadata clearly and explicitly include the identifier of the data it describes • F4. (meta)data are registered or indexed in a searchable resource
<ul style="list-style-type: none"> • A1. (meta)data are retrievable by their identifier using a standardized communications protocol • A1.1 the protocol is open, free, and universally implementable • A1.2 the protocol allows for an authentication and authorization procedure, where necessary • A2. metadata are accessible, even when the data are no longer available
<ul style="list-style-type: none"> • I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. • I2. (meta)data use vocabularies that follow FAIR principles • I3. (meta)data include qualified references to other (meta)data
<ul style="list-style-type: none"> • R1. meta(data) are richly described with a plurality of accurate and relevant attributes • R1.1. (meta)data are released with a clear and accessible data usage license • R1.2. (meta)data are associated with detailed provenance • R1.3. (meta)data meet domain-relevant community standards

Table 3: The FAIR principles [1]

197 The FAIR principles [1] (see Table 3) require metadata to include the identifier of the base
 198 data (principle F3) and to be independently accessible (A2). Special emphasis is put on "rich"
 199 metadata. The term is associated with findability (F2), but defined in the context of reusability
 200 (R1). In fact, from the formulation of R1 it appears that rich metadata is the *essence* of reusability.
 201 Its definition is left vague, which is likely intentional to allow the concept to be applied in various
 202 domains. Richness implies "a plurality of accurate and relevant attributes". The only specific
 203 attributes mentioned are a data usage license (R1.1) and provenance (R1.2). Further attributes
 204 must "meet domain-relevant community standards" (R1.3). In our view, this means it is both
 205 possible and necessary to develop community-specific interpretations of metadata richness. The
 206 focus on reusability implies that richness must be explained from a user perspective.

207 4.2 Executive metadata and reusable metadata

208 While the FAIR principles promote the development of metadata for data reusers, data producers
 209 already create and manage metadata routinely for their own purposes. This does not imply that
 210 they would describe their own practice in these terms or use specific tools and methods. It means
 211 that some forms of metadata creation and management are just an innate part of being an effective
 212 researcher. Examples will be given below. What is the relationship between the executive
 213 metadata necessary for data production and the rich FAIR metadata supporting, enabling, or
 214 facilitating data reuse? This question has a content aspect and a form aspect: Which metadata
 215 topics are relevant for producers or reusers? And which formal requirements are demanded by
 216 either group? Since these questions address two different stakeholders, they foreshadow the
 217 discussion of social aspects of FAIR RDM in Section 5.

	Both projects	RoBivaL	DeeperSense
Producer	<ul style="list-style-type: none"> • Logistics (DAS) • Production standards, tools, workflows (A) • Team coordination (A) • Errors (DAS) 	<ul style="list-style-type: none"> • Robot maintenance and development (A) 	<ul style="list-style-type: none"> • High performance computing (A)
Producer and Reuser	<ul style="list-style-type: none"> • Field spec. (DAS) • Sensor spec. (DS) • Software spec. (DA) • Data format (DS) • Data statistics (A) • Related work (D) 	<ul style="list-style-type: none"> • Experiment spec. (D) • Robot spec. (D) • Key robot properties, perf. metrics (D) • Measuring methods (D) • Source categorization (DAS) 	<ul style="list-style-type: none"> • Machine learning methodology (DAS) • Scene description (D) • Sensor configuration (DS)
Reuser	<ul style="list-style-type: none"> • DOI, URL (A) • Usage license (A) • Provenance (A) • Public ontology (DS) • Extra use cases (DA) 	<ul style="list-style-type: none"> • Tag unused data: by-catch, invalid runs (D) 	<ul style="list-style-type: none"> • Typical examples (D)

Table 4: Metadata topics relevant for producers or reusers in RoBivaL and DeeperSense. Predominant purposes: descriptive (D), administrative (A), or structural (S).

218 Table 4 presents the metadata topics of RoBivaL and DeeperSense categorized by project and by
 219 relevance for producers or reusers. Intersections are possible on both dimensions. Each topic is
 220 labeled with its dominant purpose(s), i.e., descriptive (D), administrative (A), or structural (S).

221 The assignment of topics to producers or reusers is guided by the assumption that either group has
 222 a different primary motive: Producers want a correct execution of their project plan to achieve
 223 their primary research goal. Reusers want a sufficient understanding of the base data to assess its
 224 utility, and to integrate it into their own work flow. Our assumption about producers is primarily
 225 based on our personal experience, i.e., they reflect the motives and requirements prevalent in
 226 the two examined projects and within our institutions more generally. We believe that these
 227 assumptions are neither surprising nor uncommon. The point here is to observe the contrast
 228 between producers and reusers. Our assumption about reusers is based on our interpretation of
 229 the FAIR principles. Both assumptions are further substantiated in Section 5.

230 The different motives also affect the formal requirements. Data producers care less if all metadata
 231 is specified and captured explicitly and formally, but tolerate tacit expert knowledge, code
 232 logic, informal communication, etc. For the sake of efficiency and expediency, they may limit
 233 content and form of metadata to what is essential to their needs. Reusers on the other hand
 234 require all metadata to be explicit, since they lack the immediate access to the creation context
 235 that producers have. To support efficient machine processing, metadata must be formalized. In
 236 order to cover a broad range of possible reuse cases, it must be rich in the FAIR sense.

237 4.3 Base elements of the metadata creation process

238 This section analyzes the process of metadata creation and derives some process-related metadata
 239 categories. The matter is treated abstractly and at the micro level, i.e., with regard to individual
 240 data elements; the big picture of the data lifecycle is discussed in Section 6. The analysis yields

241 elements for the design of metadata production workflows. This is useful in a collaborative
 242 setting with a division of labor, where responsibilities must be communicated effectively.
 243 The data flow diagram in Figure 5 illustrates the first order of metadata creation on a single data
 244 processing stage. Base data processing is represented vertically from top to bottom, metadata
 245 processing horizontally from left to right. The Output represents a piece of base data which is
 246 generated by some Procedure. Output and Procedure are the subjects of metadata. For both,
 247 metadata creation has two phases: Before the subject exists, it is designed; after it exists, it may
 248 be documented. The design is metadata that is injected into the Procedure; the documentation is
 249 metadata that is extracted either from the Procedure or from the Output.

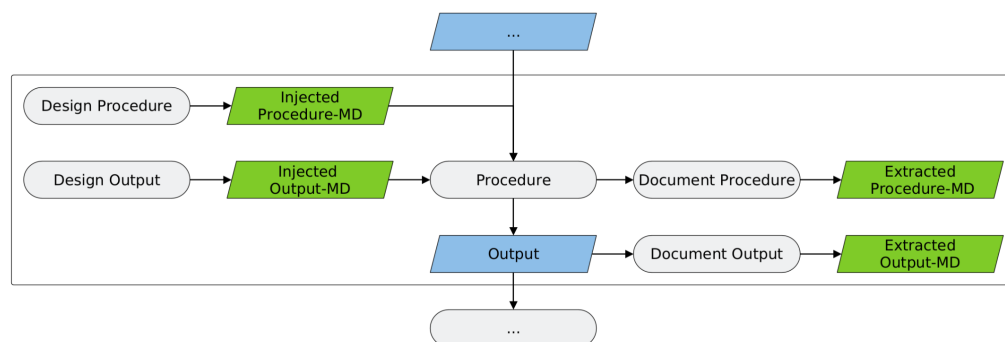


Figure 5: Data flow on a single stage of first order metadata creation. Rounded box: Process. Angled box: (Meta)data. Blue: Data. Green: Metadata. © DFKI, Christian Backe. License: CC BY 4.0 International

250 The entire model can be stacked vertically to represent multi-stage data transformation, i.e., the
 251 Procedure may receive output of a previous stage as its input, the Output may serve as input to
 252 another procedure on a subsequent stage. This model facilitates division of labor by modularizing
 253 metadata both in the content domain (distinguishing metadata subjects Procedure and Output)
 254 and across time (distinguishing design and documentation phase).

255 Table 5 lists examples for each of the four first order metadata categories taken from the
 256 DeeperSense project. They are related to the same Procedure (“Capture camera and sonar
 257 images of a diver”) and corresponding Output (“Logfiles with raw camera and sonar data”).

	Procedure-related	Output-related
Injected	<ul style="list-style-type: none"> • Middleware (ROS 2.0) • Start time 	<ul style="list-style-type: none"> • Raw data structure (“topic”) • File name (Identifier)
Extracted	<ul style="list-style-type: none"> • Scene description • Event documentation 	<ul style="list-style-type: none"> • Number of recorded samples • File size

Table 5: Examples of the four first order metadata categories from the DeeperSense project

258 The assertion that metadata is data, as mentioned in Section 4.1, implies that metadata creation
 259 may be recursive: Higher orders of metadata can treat metadata of lower orders as their base
 260 data. Visually, this means we can stack the first order metadata creation model not just vertically,
 261 but also horizontally. This is illustrated in Figure 6. It contains a condensed version of Figure 5:
 262 The process “Create MD” represents all four metadata creation processes of the first order, which
 263 are applied to the base Procedure and Output. “First order MD” represents all four first order

264 metadata types. The recursion recognizes that "Create MD" and "First order MD" themselves
 265 are a procedure-and-output pair, hence they become subjects of meta-metadata creation.

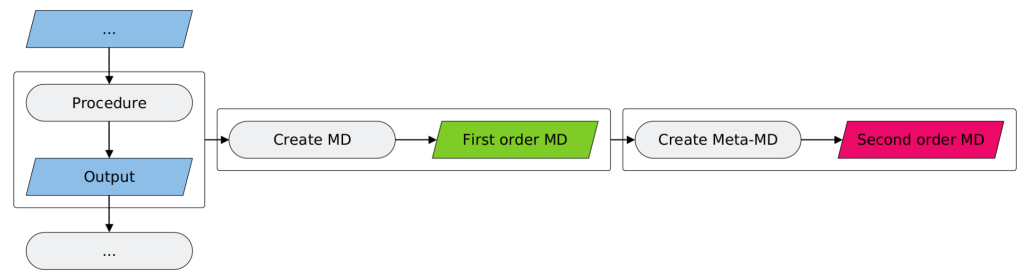


Figure 6: Recursive creation of higher order metadata. © DFKI, Christian Backe. License: CC BY 4.0 International

266 Table 6 gives two sets of generic examples for higher order metadata on multiple levels. The
 267 first example features pieces of literal base data and metadata: A speed measurement is taken at
 268 a certain time; the time stamp formatting is expressed in C string format notation; syntax and
 269 semantics of this formatting are governed by an ISO standard. The second example has a similar
 270 application pattern, but references files, which support structured data and semantic networking.

Base data	Metadata	Meta-Metadata	Meta-Meta-Metadata
5.3 m/s	2023-09-27 09:37:51	%Y-%m-%d %H:%M:%S	ISO 8601
camera.mp4	metadata.json	schema.json	https://json-schema.org

Table 6: Generic examples of higher order metadata

271 How does our metadata process model compare to the processing step class of the Metadata4Ing
 272 (M4I) ontology depicted in Figure 7? The M4I model acknowledges that each data output is
 273 generated by a process, and that data processing may be chained, which corresponds to the
 274 vertical direction of our model. But the M4I model does not appear to cover the process of
 275 metadata creation itself, i.e., our model's horizontal direction (injected vs. extracted metadata,
 276 higher order metadata). We assume this absence is at least partly a result of the purpose of
 277 the M4I model, which is communication of metadata to data consumers after the base data
 278 and metadata have been created. As mentioned above, the modularization of metadata in our
 279 model serves to design workflows for metadata creation by a data production team during project
 280 execution. A further difference between the M4I processing step and our model is that the former
 281 specifies a fixed set of attributes, while the latter is agnostic in this regard. Finally, the M4I
 282 processing step model provides the opportunity to encapsulate multiple substeps into a single
 283 step of larger scale. So far, our model does not feature a similar means of abstraction.

284 5 Social dimension: Collaborative FAIR data management in field research

285 This section discusses the social dimension of metadata creation and management from the
 286 perspective of a research data manager who follows FAIR principles. We argue that a FAIR
 287 manager acts as a link between three social domains, where they perform different primary tasks.

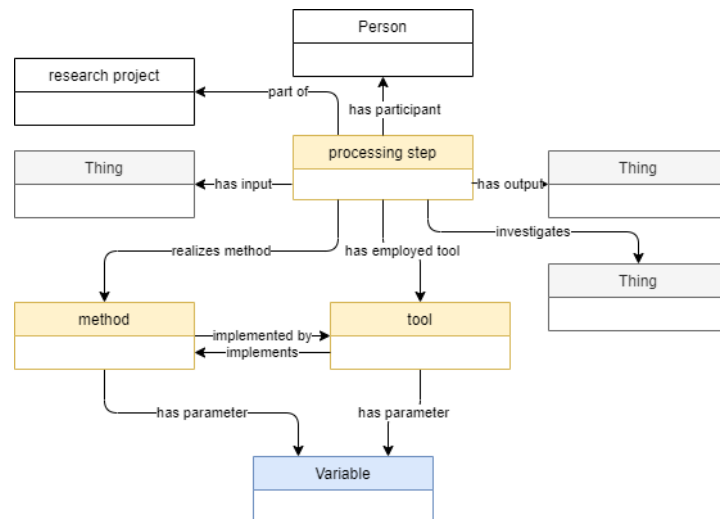


Figure 7: Processing step class of the Metadata4Ing ontology [43]. © Metadata4Ing Workgroup. License: CC BY 4.0 International

288 5.1 Collaboration with the data production team

289 The first social domain is the data production team. Here, the primary task of any data manager
290 (irrespective of FAIRness considerations) is collaboration.

291 Collaborative research in general is challenging, because it involves a multitude of people who
292 must be coordinated and accommodated. If they come from different disciplines and institutions,
293 they may have different motivations, goals, expertise, responsibilities, standards, and practices.
294 These individual attributes may not be equally transparent for everybody, and not be equally
295 present in everyone's mind, which can complicate intra-group communication.

296 Collaborative *field* research is particularly challenging: The pressure to perform is very high,
297 because there are limited opportunities to go into the field; field conditions can be difficult
298 and unpredictable, which often leads to unforeseen problems; equipment and people are put to
299 unusual stress. The main priority is to get all people and systems to work at all at the designated
300 time and place, and to capture the primary data that serves the project goal. This often requires
301 improvisation and adaptation, because prototype systems may break or deviate from specification,
302 and the captured data may not match earlier expectations. Figure 8 illustrates these notions in
303 the context of DeeperSense.

304 This assessment has two immediate implications for effective RDM in field research: First, RDM
305 must be reliable and unobtrusive. A field research team wants their RDM to ease the effort, not
306 stand in the way or cause extra concerns. Second, RDM must capture unforeseen events, so they
307 can be factored into the preparation of future field missions.

308 5.2 Mediation between producers and reusers

309 The previous section dealt with RDM in general. For a FAIR data manager in particular, there is
310 a second social domain, namely the larger research domain. Here, their primary task is mediation
311 between conflicting requirements of their data production team on the one hand, and potential

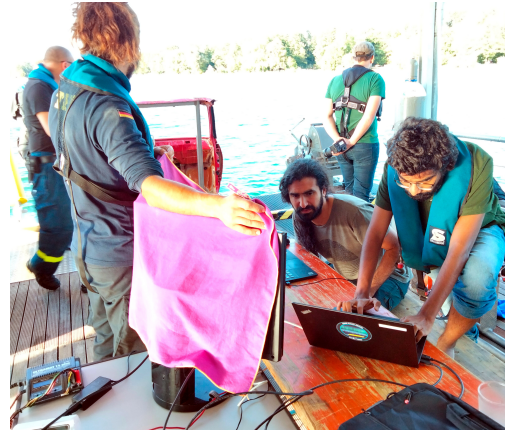


Figure 8: Field team work in DeeperSense. On the last day of data collection, the team is on a boat on a lake, gathering a critical piece of data necessary for the final demonstration event. The underwater system keeps failing. Error messages on the computer screens are difficult to read due to the glaring sun. © DFKI, Christian Backe. License: CC BY 4.0 International

312 data reusers (as characterized by the FAIR principles) on the other.

313 We hinted at this conflict in Section 4.2 and can fully express it in light of Section 5.1: Reusers
 314 require explicit, formal, rich metadata to thoroughly understand the data that is foreign to them,
 315 easily interface with it using machines, and have it serve a broad spectrum of potential use cases.
 316 But this demands extra effort from the producers, who not only have the privilege of being more
 317 implicit, informal, and brief in their internal communication, but who may actually be forced to
 318 cut corners, especially under field conditions, in order to reach their primary research goal.

319 Table 7 summarizes this proposition and adds two aspects derived from experience in RoBivaL
 320 and DeeperSense: In a collaborative setting with division of labor, executive metadata may be
 321 distributed over many places convenient for different contributors; to become reusable, it must be
 322 consolidated. While research is ongoing, the executive metadata design may need to evolve to
 323 adapt to changing circumstances; reusers prefer reliable APIs.

Data producers	Data reusers
Research execution	Data understanding and interoperation
Tacit common knowledge	Explicit metadata files
Ad-hoc communication	Formal specification, Ontologies
Single actual use case	Several potential use cases
Distributed information	Coherent information
Flexible, evolving designs	Static APIs (keywords, structures)

Table 7: Different priorities and requirements of data producers and reusers

324 The conflicting priorities and requirements of data producers and reusers have two implications
 325 for a FAIR research data manager: First, they must motivate their team to apply the extra effort.
 326 One possible incentive may be that today's producers are their own reusers tomorrow, so the
 327 investment in more elaborate metadata will pay off directly towards themselves. There is an
 328 indirect version of this: By creating metadata they would be happy to receive if they were reusers,
 329 producers influence the standards of their community to their own benefit. Another incentive may

330 be increased impact of their research if the underlying data is broadly adopted in the community.
331 A second implication is that FAIR research data managers must design the workflow of their
332 team such that the extra effort necessary to satisfy reuser requirements does not coincide with
333 peak effort towards the primary research goal, because the latter will always have precedence.

334 5.3 Standardization in the FAIR RDM community

335 The third social domain for a FAIR research data manager is the FAIR RDM community. Here,
336 their primary task is to participate in the standardization of FAIR practices in a particular research
337 domain and maybe across domains. We believe the outcome of this activity can be conceptualized
338 as higher order metadata.

339 From a purely theoretical perspective, the metadata recursion could go on to unlimited orders.
340 But in practice, of course, a cut-off is made, from which on the participants (i.e., data producers
341 either among themselves or in relation to reusers) regard all higher metadata orders as common
342 knowledge to be inferred from context or prior convention. Still, the communication relies in
343 principle on the assumption that all higher metadata orders *could* be delivered explicitly. One
344 core role of the FAIR RDM community is to underwrite this assumption, i.e., to work towards a
345 codification of common knowledge (including standards and open vocabularies) to which all
346 participants can refer in their communication.

347 6 Time dimension: A self-improving data lifecycle

348 This section divides FAIR research data management into different tasks and organizes them
349 across time. Subsection 6.1 discusses the concept of a data lifecycle and proposes some modifica-
350 tions to the type of lifecycle used by NFDI4Ing and similar parties. The two main modifications
351 are the introduction of an internal data provision phase necessary for collaborative research, and
352 the introduction of an evaluation phase to drive an iterative improvement of the RDM system.
353 Subsection 6.2 presents some lessons learned from RoBivaL and DeeperSense in each phase.

354 6.1 Model of a self-improving data lifecycle

355 There is no consensus which phases constitute a data lifecycle and how the phases shall be ordered.
356 In their survey of 76 data lifecycles, Shah et al. identify at least 14 phases [46]. NFDI4Ing uses
357 a model with six phases, named Planning, Production, Analysis, Storage, Access, and Re-Use
358 [47]. It is similar to other six-phase models prevalent in the FAIR RDM community [48]–[50]
359 but there are still differences about the naming and ordering of the phases. These models have
360 two shortcomings regarding their application to collaborative and iterative research.

361 First, while there is a phase in these models near the end of the cycle for making data available
362 externally to the public (called "Publication", "Access", "Sharing", or "Disclosure"), there
363 is no equivalent phase dedicated to making the data available internally to the research team
364 immediately after creation. In our experience, such a phase is necessary in collaborative research,
365 and it has different requirements than the publication phase. We propose to call it Provision.

366 Second, almost all phases are actions that apply to data (data is produced, analyzed, ...), except
367 for Planning which is the only phase that applies to other actions (production is planned, analysis

368 is planned, ...). Another oddity about Planning is that it has no corresponding phase for looking
 369 into the past. In iterative research, comparing how things were planned to how they turned out
 370 would enable an iterative improvement of the RDM system. Since our research is in fact iterative,
 371 such a self-improving data lifecycle would be welcome. Therefore, we propose two additional
 372 phases called Execution and Evaluation. Together with Provision, they apply to each data-related
 373 action and thus form a separate loop nested with the data-related loop.

374 In summary, our proposed model has six data-related phases: Creation, Provision, Processing,
 375 Publication, Reuse, and Archiving. Each of these is divided into three process-related phases:
 376 Planning, Execution, and Evaluation. The model is illustrated in Figure 9.

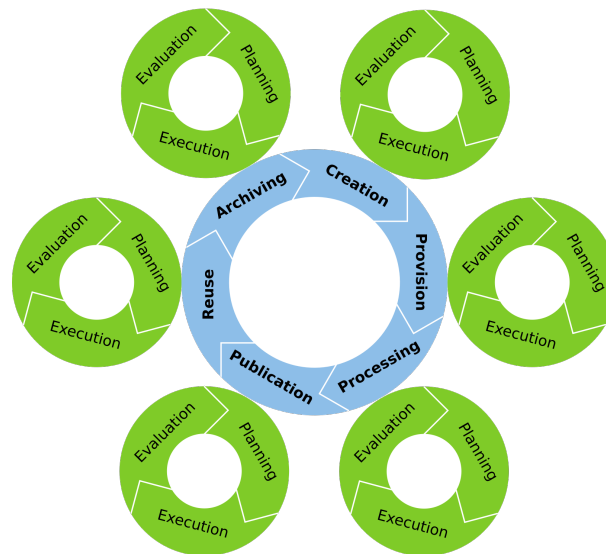


Figure 9: Self-improving nested data lifecycle. © DFKI, Christian Backe. License: CC BY 4.0 International

377 The term Creation is chosen over Collection or Acquisition to emphasize the designed and
 378 fabricated nature of data and metadata. The broader term Processing is preferable over the
 379 narrow term Analysis, because we encounter a range of data processing activities in our practice,
 380 both primary (e.g. machine learning model development is synthesis, rather than analysis) and
 381 secondary (e.g., data cleaning, fusion, performance tuning, or quality assurance). The term
 382 Planning is to include Preparation. A single word is used here for brevity, but one should be
 383 aware that it does not signify purely cerebral activity, but also, e.g., handling of hardware.

384 6.2 Lessons learned from RoBivaL and DeeperSense

385 This section serves to illustrate the data lifecycle model discussed in Section 6.1 by presenting
 386 lessons learned in the different lifecycle phases of RoBivaL and DeeperSense. Many of the
 387 lessons are derived from failures, either to perform a task or to anticipate a challenge. Due to
 388 space constraints, we focus on consequences and leave specifics of the failures mostly implicit.
 389 Methods and strategies from Sections 4 and 5 are addressed where appropriate.

390 6.2.1 Creation

391 **Planning** The planning of data creation deserves special care, because errors made during
392 creation are typically difficult to repair. In the field, errors may not be repairable at all if the field
393 conditions cannot be replicated or the cost of another deployment is prohibitive.

394 Data management needs to specify the scope and form of the metadata set, and provide tools
395 and procedures for metadata creation. Workflows and responsibilities can be made transparent
396 by sorting the planned metadata items into the classes discussed in Section 4, i.e., executive
397 vs. reusable, injected vs. extracted, process- vs. output-related. The specification of first-order
398 metadata items involves the creation of higher-order metadata content. In field research, the
399 environmental conditions and the data creation process need to be documented more extensively
400 than in the lab, because there are fewer means of control and more chances of surprise.

401 Terms coined during planning will propagate through a growing corpus of communication,
402 documentation, and implementation. To avoid costly changes later, it is advisable to stabilize the
403 terminology early on. FAIR terminologies must reflect community practices in their domains.
404 This can either facilitate planning if a communal terminology already exists, or it can complicate
405 planning if a terminology first needs to be compiled from scholarly sources. FAIR data managers
406 may need to advocate for the requirements of reusers as discussed in Section 5.2.

407 Terminology requirements may be different for humans and machines. Machines need more
408 consistency, less ambiguity, and may accept only restricted token sets. Inconsistency and
409 ambiguity may arise, e.g., in interdisciplinary settings when different communities use different
410 terms for the same thing or the same term for different things. Since humans are more flexible, a
411 machine-consumable version is preferable for co-processed information, e.g., file and directory
412 names. In this case, human collaborators must be educated on machine requirements.

413 Concerning the tensions discussed in Sections 4.2 and 5.2, the creation of purely executive or
414 purely reusable metadata is relatively easy: Producers are intrinsically motivated to fulfill their
415 own needs, and pure reuse issues can be handled by the data manager alone. For metadata
416 concerning both producers and reusers, however, the data manager again has to advocate for the
417 reusers' requirements and possibly bear the additional effort (or part of it) during execution.

418 **Execution** In a collaborative setting, different people may observe different features of an
419 object. This is an opportunity for the data manager to be a team player, as discussed in Section
420 5.1: Having one person responsible to record all observations avoids misalignment and ensures
421 consistency, completeness, and uniform compliance with standards, e.g., related to accuracy or
422 measuring units. The information relay requires structured communication and routines to avoid,
423 detect, and correct miscommunication.

424 A designated record keeper can also take note of problems and unforeseen events which may
425 help improve the planning of future data creation sessions. This may be performed proactively
426 by looking out for and trying to prevent errors in the first place. As discussed in Section 5.1,
427 field data creation can be cognitively very taxing, so it is easy to miss, e.g., a critical failure of a
428 single component. Therefore, having someone specifically focused on error detection is useful.

429 **Evaluation** DeeperSense and RoBivaL each had multiple data creation sessions, so there was
430 reason and occasion to improve the data creation system during the project, e.g., by capturing
431 additional metadata items, or by simplifying the creation process. This was partially countered
432 by the requirement to have data and metadata be compatible between all sessions. To avoid this
433 tension, it is advisable to perform pre-trials where the data creation system can be tested.

434 If the metadata recording task is delegated, e.g., due to illness, the recording tools must be
435 usable for the delegate, who may be unfamiliar with the task and have additional responsibilities.
436 Mandatory and important metadata items must be indicated. Content requirements must be
437 clearly communicated. Number and complexity of items should be kept at a minimum.

438 6.2.2 Provision

439 **Planning** Provision is dedicated to the needs of the original research team, in contrast to
440 publication which caters to reusers. Therefore, provision seems to require executive metadata,
441 while publication requires reusable metadata (see Section 4.2). Still, to ensure a smooth transition
442 between the phases, it may be wise to gather reusable metadata already during provision.

443 The internal data repository must be laid out physically: How much data will be stored where
444 and for which purpose? For example, there may be storage embedded in sensor platforms to
445 collect raw data; file servers to consolidate, backup, and exchange data; database servers to
446 validate, merge, filter, and aggregate data; workstations of different contributors to process and
447 analyze data parts; high-performance servers for compute-intensive tasks.

448 Logically, the repository can be specified with different resolutions, on multiple layers and
449 domains. Aspects to consider may be file trees, database schemas, and request APIs; encodings,
450 types, and formats; sources and processing stages; separation of base data and metadata; auxiliary
451 assets (e.g., documentation, specification, schemas, logs, errors). The terminology should be
452 consistent between layers and domains, and be compatible with terminologies of the other phases.
453 Again, in case of tensions between the needs of producers and reusers, the data manager may
454 have to advocate for the reuser perspective (see Section 5.2).

455 Governance and administration of the internal repository as a shared resource must be specified.
456 Who gets access to what? How are safety, security, availability, quality, and privacy established?
457 Who is in charge for which procedures? Examples are consolidation of data from different
458 sources, sessions, or processing stages; deduplication of redundant data; replication to prevent
459 data loss; data removal to free resources; consistency checking and error management.

460 **Execution** An explicit specification of the physical and logical layout can improve team
461 alignment. User onboarding is an opportunity to check if the specification is properly understood
462 and reflects the actual requirements. The layout and its specification may need to be updated to
463 account for e.g., larger volume, changing pipelines, different data formats, etc. Such adjustments
464 during execution may never fully be prevented; still they should be noted for evaluation.

465 DeeperSense and RoBivaL developed dedicated metadatabases to facilitate reporting (e.g.,
466 volume per data layer, sample count per sensor type and session, runs per experiment and robot).
467 As standalone items, they can be transmitted separately from the large base data corpora. They

468 provide information for decision-making both by the executing researchers (e.g., are there critical
469 data gaps?) and reusers (e.g., is this dataset suitable for my use case?). Thus, the metadatabases
470 are a further example of shared concerns as discussed in Section 4.2.

471 **Evaluation** Physical and logical layouts emerge even if they are not expressly designed. They
472 are implemented by contributors out of necessity to accomplish particular tasks, and are reinforced
473 by continued use. To achieve interoperability and consistency in a collaborative setting, a
474 patchwork of individual approaches must be consolidated.

475 But there is tension: Research data processing must be flexible enough to adjust to new findings
476 and changing views. In interdisciplinary research, practices from different domains must be
477 accommodated. Too much specification too early or too rigidly may lower the acceptance and
478 adoption of a layout. Further, writing a comprehensive, accurate, and understandable specification
479 may be difficult and time-consuming, thus conflicting with other priorities. On the other hand,
480 working with undocumented, inconsistent layouts that need to be reverse engineered and might
481 change without notice, lowers productivity and risks producing bad results. This dilemma shows
482 that it may not always be obvious for a data manager how to follow the maxim expressed in
483 Section 5.1 to ease the effort of the production team.

484 6.2.3 Processing

485 **Planning** The processing phase is typically comprised of multiple stacked processing sub-
486 stages, as discussed in Section 4.3 and depicted in Figure 5. Therefore, the processing phase
487 may give rise to a lot of first-order metadata content. Injected metadata may already be created
488 during planning, both related to the (sub-)processes and to their outputs. Extracted metadata
489 will normally be created during execution. Some extracted metadata may be created during
490 evaluation, e.g., if problems with the processes or their outputs must be documented.

491 Processing resources must be supplied for different tasks and stages. This includes individual
492 workstations for all team members, and high performance servers that are used as a shared
493 resource. If there are multiple contributors, it is important to specify who is responsible for
494 which processing job, and what are the interfaces between consecutive steps in a processing
495 pipeline.

496 **Execution** One core responsibility of the data manager is metadata processing. In RoBivaL
497 and DeeperSense, this was done in the context of developing and maintaining a metadatabase,
498 involving schema design, metadata extraction, fusion, and aggregation. (These tasks require
499 also creation and provision, and are planned before execution; they are highlighted here to mark
500 the metadatabase as a processing tool.) The data manager may also be tasked with (meta)data
501 quality assurance, which affects all other processing jobs. This involves the conception of error
502 cases, error logging, escalation of errors, and resolution management.

503 If the results of a processing step need to be persisted for later consumption by other processing
504 steps, this produces a feedback loop between processing and provision.

505 **Evaluation** In case the input or output requirements of a processing step change, updates to the
506 interface with its predecessor or successor steps may need to be negotiated.

507 6.2.4 Publication

508 **Planning** The data repository where the data and metadata are to be published should match
509 the given content. Where will the intended reusers be likely to look for data to match a certain
510 use case? The journal Scientific Data recommends various data repositories geared towards
511 particular natural and social sciences, as well as some generalist repositories [51]. For large
512 datasets, space constraints by different repositories may have to be considered.

513 **Execution** Data from RoBivaL and DeeperSense was published on Zenodo. The publisher
514 requires filling out a form with platform-specific metadata, i.e., authors and contributors with
515 affiliations and IDs, a summary description of the dataset, references to related publications, etc.

516 **Evaluation** Typically, only a part of all data and metadata created during a project will be
517 published. To facilitate the separation, it is advisable to store the parts dedicated for publication
518 at a separate place from the beginning, or at least design the internal storage such that these parts
519 are clearly marked and can be easily extracted.

520 6.2.5 Reuse

521 Reuse is different from the other data lifecycle phases, because its planning, execution, and
522 evaluation are outside the purview of the data production team. We did not get any feedback
523 from data reusers yet, so we currently cannot report any experiences about the reuse of data from
524 RoBivaL or DeeperSense.

525 6.2.6 Archiving

526 The data from RoBivaL and DeeperSense has not been archived yet, so there is no experience to
527 report.

528 7 Conclusion

529 This paper discussed the collaborative creation and management of rich FAIR metadata on three
530 dimensions: the metadata content, the social relationships between metadata stakeholders, and
531 the phases of metadata management over time. The discussion was illustrated with examples
532 from the robotics field research projects RoBivaL and DeeperSense.

533 On the content dimension, we categorized metadata by different purposes, presented a broad
534 spectrum of metadata topics, and discussed the relationship between executive metadata for data
535 producers, and rich reusable metadata to satisfy the FAIR principles. We modeled the process of
536 metadata creation at the micro level, introducing the concepts of injected and extracted metadata,
537 and of higher order metadata.

538 One risk to consider here is the possibility of scope explosion in multiple directions: Firstly, since
539 executive metadata covers many areas, metadata management for internal purposes might soon

540 turn into general knowledge management. Secondly, since rich metadata lacks a comprehensive
541 definition and is grounded in potential needs of data reusers, it is difficult to judge what must be
542 included and what may be omitted. Thirdly, higher order metadata implies an infinite recursion
543 which must be capped at a level that is reasonable for different stakeholders.

544 The purpose of higher order metadata is to create formal and accessible expressions of common
545 knowledge and practice which may exist primarily in the heads of practitioners. This is difficult
546 for multiple reasons, not least because it entails a social process: Who may contribute their
547 expertise and how? Does everyone agree with an expression and how are conflicts resolved?

548 To get broadly adopted, FAIR RDM practices must make sense to data producers. We argued
549 that this is more likely if producers see their own requirements and challenges taken into account.
550 Still, caring for reusability may appear to many researchers as a burden that interferes with their
551 primary goals. Therefore, we presented FAIR data production as team work where someone
552 takes on the role of a dedicated FAIR RDM expert who at the same time provides immediate
553 value to their research team. We attempted to contribute to a definition of this role and explain
554 its competing demands, and we presented tools for the design and communication of FAIR RDM
555 workflows that facilitate collaboration in data production teams.

556 Trust is a social aspect we omitted in our discussion, because it is a broad topic in itself and
557 involves additional stakeholders. Data reuse depends on the assumption that the delivered data
558 is not manufactured to deceive. Though not a FAIR principle, this is certainly a maxim of
559 scientific fairness in a broader sense. But even if their intentions are pure, producers may deceive
560 themselves in thinking their data is accurate and represents reality. This problem is compounded
561 when data is processed by different people on multiple stages, or fused from multiple providers.
562 At the end of the data supply chain are people who apply, consume, or are otherwise affected by
563 products derived from data. For them, trustworthiness may literally be a life-and-death issue.
564 The DeeperSense sonar-to-camera translation is an example from our own research. Diving
565 companies have expressed their motivation to solve the trustworthiness problem in this case.

566 On the time dimension, we divided the prevalent image of a simple data lifecycle into an outer
567 and an inner cycle: The phases of the outer cycle are actions that apply to data (i.e., creation,
568 provision, etc.). The phases of the inner cycle are actions that apply to each outer phase, namely
569 planning, execution, and evaluation. Evaluation allows the data management system to improve
570 over multiple research iterations.

571 One important challenge here is to find the right balance between flexibility and stability of the
572 data management system. Flexibility is necessary to eliminate errors and inefficiencies in the
573 system itself, and to be able to adapt to new insights and requirements for the primary research.
574 Stability of the system facilitates its adoption, provides backwards compatibility, and allows
575 one to devote more energy to primary research. The trick is to know when the system is good
576 enough, and to stop improving when the marginal benefit becomes too small.

577 **8 Acknowledgements**

578 The authors would like to thank the Federal Government and the Heads of Government of the
579 Länder, as well as the Joint Science Conference (GWK), for their funding and support within the

580 framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) -
581 project number 442146713.

582 The RoBivaL project was funded by the Federal Ministry for Economic Affairs and Climate
583 Action with grant number 50RP2150.

584 The DeeperSense project was funded by the European Union. Program: H2020-ICT-2020-2
585 ICT-47-2020. Project Number: 101016958.

586 9 Roles and contributions

587 **Christian Backe:** Conceptualization, Data curation, Investigation, Software, Visualization,
588 Writing - original draft, Writing - review & editing

589 **Veit Briken:** Conceptualization, Writing - review & editing

590 **Atefeh Gooran Orimi:** Investigation, Project administration, Writing - review & editing

591 **Rayen Hamlaoui:** Investigation, Writing - review & editing

592 **Malte Wirkus:** Data curation, Funding acquisition, Investigation, Project administration, Soft-
593 ware, Writing - review & editing

594 **Bilal Wehbe:** Data curation, Funding acquisition, Investigation, Project administration, Visual-
595 ization, Writing - review & editing

596 **Frank Kirchner:** Funding acquisition, Supervision, Writing - review & editing

597 References

598 [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The FAIR Guiding Principles for
599 scientific data management and stewardship,” *en, Scientific Data*, vol. 3, no. 1, p. 160 018,
600 Mar. 2016, ISSN: 2052-4463. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18). [Online]. Available:
601 <https://www.nature.com/articles/sdata201618> (visited on 10/13/2023).

602 [2] A. Jacobsen, R. de Miranda Azevedo, N. Juty, *et al.*, “FAIR Principles: Interpretations and
603 Implementation Considerations,” pp. 10–29, Jan. 2020. DOI: [10.1162/dint_r_00024](https://doi.org/10.1162/dint_r_00024).

604 [3] DFKI GmbH, *RoBivaL - Robot Soil Interaction Evaluation in Agriculture*, en-US, website,
605 2021. [Online]. Available: [https://robotik.dfki-bremen.de/en/research/proj
606 ects/robival/](https://robotik.dfki-bremen.de/en/research/projects/robival/) (visited on 12/10/2023).

607 [4] DFKI GmbH, *DeeperSense - Deep-Learning for Multimodal Sensor Fusion*, en-US,
608 website, 2021. [Online]. Available: [https://robotik.dfki-bremen.de/en/resear
609 ch/projects/deepersense/](https://robotik.dfki-bremen.de/en/research/projects/deepersense/) (visited on 12/10/2023).

610 [5] C. Backe, A. Gooran Orimi, V. Briken, R. Hamlaoui, and H. Görner, *Creating Rich
611 Metadata for Collaborative Research: Case Studies and Challenges*, eng, Sep. 2023. DOI:
612 [10.5281/zenodo.8430753](https://doi.org/10.5281/zenodo.8430753). [Online]. Available: [https://zenodo.org/records/84
613 30753](https://zenodo.org/records/8430753) (visited on 12/10/2023).

- 614 [6] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans, *Foundation Models*
615 *for Decision Making: Problems, Methods, and Opportunities*, 2023. arXiv: 2303.04129
616 [cs.AI]. [Online]. Available: <https://arxiv.org/abs/2303.04129>.
- 617 [7] A. Brohan, N. Brown, J. Carbajal, *et al.*, *RT-1: Robotics Transformer for Real-World*
618 *Control at Scale*, 2023. arXiv: 2212.06817 [cs.R0]. [Online]. Available: <https://arxiv.org/abs/2212.06817>.
- 620 [8] Q. Vuong and P. Sanketi. “Scaling up learning across many different robot types.” (Oct.
621 2023), [Online]. Available: [https://deepmind.google/discover/blog/scaling-](https://deepmind.google/discover/blog/scaling-up-learning-across-many-different-robot-types/)
622 [up-learning-across-many-different-robot-types/](https://deepmind.google/discover/blog/scaling-up-learning-across-many-different-robot-types/) (visited on 11/14/2024).
- 623 [9] Embodiment Collaboration, A. O’Neill, A. Rehman, *et al.*, *Open X-Embodiment: Robotic*
624 *Learning Datasets and RT-X Models*, 2024. arXiv: 2310.08864 [cs.R0]. [Online].
625 Available: <https://arxiv.org/abs/2310.08864>.
- 626 [10] R. Firoozi, J. Tucker, S. Tian, *et al.*, “Foundation models in robotics: Applications, chal-
627 lenges, and the future,” *The International Journal of Robotics Research*, vol. 0, no. 0,
628 p. 02 783 649 241 281 508, 0. DOI: 10.1177/02783649241281508. eprint: [https://d](https://doi.org/10.1177/02783649241281508)
629 [oi.org/10.1177/02783649241281508](https://doi.org/10.1177/02783649241281508). [Online]. Available: [https://doi.org/10](https://doi.org/10.1177/02783649241281508)
630 [.1177/02783649241281508](https://doi.org/10.1177/02783649241281508).
- 631 [11] T. Schoening, J. M. Durden, C. Faber, *et al.*, “Making marine image data FAIR,” *Scientific*
632 *Data*, vol. 9, no. 1, p. 414, Jul. 2022, ISSN: 2052-4463. DOI: 10.1038/s41597-022-01
633 491-3. [Online]. Available: <https://doi.org/10.1038/s41597-022-01491-3>.
- 634 [12] C. Motta, S. Aracri, R. Ferretti, *et al.*, “A framework for FAIR robotic datasets,” *Scientific*
635 *Data*, vol. 10, no. 1, p. 620, Sep. 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-023-
636 02495-3. [Online]. Available: <https://doi.org/10.1038/s41597-023-02495-3>.
- 637 [13] R. Dominguez, M. Post, A. Fabisch, R. Michalec, V. Bissonnette, and S. Govindaraj,
638 “Common Data Fusion Framework: An open-source Common Data Fusion Framework
639 for space robotics,” *International Journal of Advanced Robotic Systems*, vol. 17, no. 2,
640 p. 1 729 881 420 911 767, 2020. DOI: 10.1177/1729881420911767. eprint: [https://d](https://doi.org/10.1177/1729881420911767)
641 [oi.org/10.1177/1729881420911767](https://doi.org/10.1177/1729881420911767). [Online]. Available: [https://doi.org/10.1](https://doi.org/10.1177/1729881420911767)
642 [177/1729881420911767](https://doi.org/10.1177/1729881420911767).
- 643 [14] S. T. Arundel, K. G. McKeegan, B. B. Campbell, A. N. Bulen, and P. T. Thiem, “A guide
644 to creating an effective big data management framework,” *Journal of Big Data*, vol. 10,
645 no. 1, p. 146, Sep. 2023, ISSN: 2196-1115. DOI: 10.1186/s40537-023-00801-9.
646 [Online]. Available: <https://doi.org/10.1186/s40537-023-00801-9>.
- 647 [15] A. Olivares-Alarcos, D. Beßler, A. Khamis, *et al.*, “A review and comparison of ontology-
648 based approaches to robot autonomy,” *The Knowledge Engineering Review*, vol. 34, e29,
649 2019. DOI: 10.1017/S0269888919000237.
- 650 [16] “IEEE Standard Ontologies for Robotics and Automation,” *IEEE Std 1872-2015*, pp. 1–60,
651 Apr. 2015. DOI: 10.1109/IEEESTD.2015.7084073. [Online]. Available: [https://ie](https://ieeexplore.ieee.org/document/7084073)
652 [eexplore.ieee.org/document/7084073](https://ieeexplore.ieee.org/document/7084073).
- 653 [17] “IEEE Standard for Robot Task Representation,” *IEEE Std 1872.1-2024*, pp. 1–32, Jun.
654 2024. DOI: 10.1109/IEEESTD.2024.10557559. [Online]. Available: [https://ieeex](https://ieeexplore.ieee.org/document/10557559)
655 [plore.ieee.org/document/10557559](https://ieeexplore.ieee.org/document/10557559).

- 656 [18] “IEEE Standard for Autonomous Robotics (AuR) Ontology,” *IEEE Std 1872.2-2021*,
657 pp. 1–49, May 2022. DOI: [10.1109/IEEESTD.2022.9774339](https://doi.org/10.1109/IEEESTD.2022.9774339). [Online]. Available:
658 <https://ieeexplore.ieee.org/document/9774339>.
- 659 [19] “IEEE Standard Proposal for Ontology Reasoning on Multiple Robots,” *IEEE Std P1872.3*,
660 Sep. 2022. [Online]. Available: [https://standards.ieee.org/ieee/1872.3/1103](https://standards.ieee.org/ieee/1872.3/11037/)
661 [7/](https://standards.ieee.org/ieee/1872.3/11037/).
- 662 [20] ISO 8373:2021, “Robotics - Vocabulary,” International Organization for Standardization,
663 Standard, 2021. [Online]. Available: <https://www.iso.org/standard/75539.html>.
- 664 [21] ISO 19649:2017, “Mobile robots - Vocabulary,” International Organization for Standard-
665 ization, Standard, 2017. [Online]. Available: [https://www.iso.org/standard/6565](https://www.iso.org/standard/65658.html)
666 [8.html](https://www.iso.org/standard/65658.html).
- 667 [22] Open Robotics, *URDF XML Specifications*, Sep. 2022. [Online]. Available: [http://wiki](http://wiki.ros.org/urdf/XML)
668 [i.ros.org/urdf/XML](http://wiki.ros.org/urdf/XML) (visited on 10/16/2023).
- 669 [23] S. Chitta, *URDF 2.0: Update the ROS URDF Format*, Jan. 2016. [Online]. Available:
670 <https://sachinchitta.github.io/urdf2> (visited on 01/16/2025).
- 671 [24] V. A. Jorge, V. F. Rey, R. Maffei, *et al.*, “Exploring the IEEE ontology for robotics and
672 automation for heterogeneous agent interaction,” *Robotics and Computer-Integrated*
673 *Manufacturing*, vol. 33, pp. 12–20, 2015, Special Issue on Knowledge Driven Robotics
674 and Manufacturing, ISSN: 0736-5845. DOI: [https://doi.org/10.1016/j.rcim.20](https://doi.org/10.1016/j.rcim.2014.08.005)
675 [14.08.005](https://doi.org/10.1016/j.rcim.2014.08.005). [Online]. Available: [https://www.sciencedirect.com/science/arti](https://www.sciencedirect.com/science/article/pii/S0736584514000660)
676 [cle/pii/S0736584514000660](https://www.sciencedirect.com/science/article/pii/S0736584514000660).
- 677 [25] A. B. d. O. Neto, J. A. Silva, and M. E. Barreto, “Prototyping and Validating the CORA
678 Ontology: Case Study on a Simulated Reconnaissance Mission,” in *2019 Latin American*
679 *Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019*
680 *Workshop on Robotics in Education (WRE)*, Oct. 2019, pp. 341–345. DOI: [10.1109](https://doi.org/10.1109/LARS-SBR-WRE48964.2019.00066)
681 [/LARS-SBR-WRE48964.2019.00066](https://doi.org/10.1109/LARS-SBR-WRE48964.2019.00066).
- 682 [26] M. Yüksel, “An expert knowledge representation based on real world experiences in
683 proof-of-concept robot system design,” Available at [https://doi.org/10.26092/eli](https://doi.org/10.26092/elib/2337)
684 [b/2337](https://doi.org/10.26092/elib/2337), PhD thesis, Universität Bremen, Jun. 2023.
- 685 [27] M. Wirkus, S. Hinck, C. Backe, *et al.*, “Comparative study of soil interaction and driving
686 characteristics of different agricultural and space robots in an agricultural environment,”
687 *Journal of Field Robotics*, pp. 1–34, DOI: <https://doi.org/10.1002/rob.22347>.
688 eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.22347>.
689 [Online]. Available: [https://onlinelibrary.wiley.com/doi/abs/10.1002/rob](https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.22347)
690 [.22347](https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.22347).
- 691 [28] ISO 18646-1:2016, “Robotics — Performance criteria and related test methods for service
692 robots. Part 1: Locomotion for wheeled robots,” International Organization for Standard-
693 ization, Standard, 2016. [Online]. Available: [https://www.iso.org/standard/6312](https://www.iso.org/standard/63127.html)
694 [7.html](https://www.iso.org/standard/63127.html).
- 695 [29] ISO 18646-2:2019, “Robotics — Performance criteria and related test methods for service
696 robots. Part 2: Navigation,” International Organization for Standardization, Standard,
697 2019. [Online]. Available: <https://www.iso.org/standard/69057.html>.

- 698 [30] C. Backe, M. Wirkus, S. Hinck, *et al.*, *RoBivaL data corpus*, Experiment Data, Zenodo,
699 Jun. 2024. DOI: [10.5281/zenodo.12547116](https://doi.org/10.5281/zenodo.12547116).
- 700 [31] B. Wehbe, N. Shah, M. Bande, and C. Backe, “Sonar-to-RGB Image Translation for
701 Diver Monitoring in Poor Visibility Environments,” in *OCEANS 2022, Hampton Roads*,
702 ISSN: 0197-7385, Oct. 2022, pp. 1–9. DOI: [10.1109/OCEANS47191.2022.9977024](https://doi.org/10.1109/OCEANS47191.2022.9977024).
703 [Online]. Available: <https://ieeexplore.ieee.org/document/9977024> (visited
704 on 12/10/2023).
- 705 [32] B. Wehbe, N. Shah, M. Bande, and C. Backe, *Sonar-to-RGB Image Translation for Diver*
706 *Monitoring in Poor Visibility Environments*, Experiment Data, Zenodo, Mar. 2023. DOI:
707 [10.5281/zenodo.7728089](https://doi.org/10.5281/zenodo.7728089). [Online]. Available: [https://zenodo.org/records/77](https://zenodo.org/records/7728089)
708 [28089](https://zenodo.org/records/7728089) (visited on 12/10/2023).
- 709 [33] O. Köken and B. Wehbe, *Fusion of Underwater Camera and Multibeam Sonar for Diver*
710 *Detection and Tracking*, Experiment Data, version 1, Zenodo, Dec. 2023. DOI: [10.5281](https://doi.org/10.5281/zenodo.10220989)
711 [/zenodo.10220989](https://doi.org/10.5281/zenodo.10220989). [Online]. Available: [https://doi.org/10.5281/zenodo.1022](https://doi.org/10.5281/zenodo.10220989)
712 [0989](https://doi.org/10.5281/zenodo.10220989).
- 713 [34] C. Backe, B. Wehbe, M. Bande, N. Shah, D. Cesar, and M. Pribbernow, *Metadata of a*
714 *Large Sonar and Stereo Camera Dataset Suitable for Sonar-to-RGB Image Translation*,
715 Experiment Data, Zenodo, Dec. 2023. DOI: [10.5281/zenodo.10373154](https://doi.org/10.5281/zenodo.10373154). [Online].
716 Available: <https://doi.org/10.5281/zenodo.10373154>.
- 717 [35] J. Furner, “Definitions of ”Metadata”: A Brief Survey of International Standards,” en,
718 *Journal of the Association for Information Science and Technology*, vol. 71, no. 6, Jun.
719 2020, ISSN: 2330-1635, 2330-1643. DOI: [10.1002/asi.24295](https://doi.org/10.1002/asi.24295). [Online]. Available:
720 <https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24295> (visited
721 on 10/11/2023).
- 722 [36] K. G. Jefferey and R. Koskela, “RDA Metadata Principles and their Use,” en, Research
723 Data Alliance, Tech. Rep., Nov. 2014, p. 6. [Online]. Available: [https://rd-alliance](https://rd-alliance.org/metadata-principles-and-their-use.html)
724 [.org/metadata-principles-and-their-use.html](https://rd-alliance.org/metadata-principles-and-their-use.html) (visited on 06/05/2023).
- 725 [37] RFII - Rat für Informationsinfrastrukturen, *Leistung aus Vielfalt. Empfehlungen zu Struk-*
726 *turen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*,
727 de. Göttingen, May 2016.
- 728 [38] OECD, *Data and Metadata Reporting and Presentation Handbook*, en. OECD, Jun. 2007,
729 ISBN: 978-92-64-03032-9 978-92-64-03033-6. DOI: [10.1787/9789264030336-en](https://doi.org/10.1787/9789264030336-en).
730 [Online]. Available: [https://www.oecd-ilibrary.org/economics/data-and-met](https://www.oecd-ilibrary.org/economics/data-and-metadata-reporting-and-presentation-handbook_9789264030336-en)
731 [adata-reporting-and-presentation-handbook_9789264030336-en](https://www.oecd-ilibrary.org/economics/data-and-metadata-reporting-and-presentation-handbook_9789264030336-en) (visited on
732 10/11/2023).
- 733 [39] National Institute of Statistical Sciences, “Metadata and Paradata: Information Collection
734 and Potential Initiatives,” en, National Institute of Statistical Sciences, Expert Panel Report,
735 Nov. 2010, p. 36. [Online]. Available: [https://www.niss.org/research/metadata-](https://www.niss.org/research/metadata-and-paradata-information-collection-and-potential-initiatives)
736 [and-paradata-information-collection-and-potential-initiatives](https://www.niss.org/research/metadata-and-paradata-information-collection-and-potential-initiatives) (visited
737 on 10/11/2023).
- 738 [40] W3C, *Data on the Web Best Practices*, en, Jan. 2017. [Online]. Available: [https://www](https://www.w3.org/TR/dwbp/)
739 [.w3.org/TR/dwbp/](https://www.w3.org/TR/dwbp/) (visited on 02/20/2024).

- 740 [41] J. Riley, *Understanding Metadata: What is Metadata, and What is it For?* (NISO Primer
741 Series), en-US. National Information Standards Organization (NISO), 2017, ISBN: 978-
742 1-937522-72-8. [Online]. Available: [https://www.niso.org/publications/under-](https://www.niso.org/publications/understanding-metadata-2017)
743 [standing-metadata-2017](https://www.niso.org/publications/understanding-metadata-2017) (visited on 02/20/2024).
- 744 [42] M. Cofield, *Metadata Basics*, en, Feb. 2024. [Online]. Available: [https://guides.lib](https://guides.lib.utexas.edu/metadata-basics)
745 [.utexas.edu/metadata-basics](https://guides.lib.utexas.edu/metadata-basics) (visited on 02/20/2024).
- 746 [43] S. Arndt, B. Farnbacher, M. Fuhrmans, *et al.*, “Metadata4Ing: An ontology for describing
747 the generation of research data within a scientific activity.,” en, Sep. 2023, Publisher:
748 Zenodo Version Number: 1.2.0. DOI: [10.5281/zenodo.5957103](https://doi.org/10.5281/zenodo.5957103). [Online]. Available:
749 <https://zenodo.org/record/5957103> (visited on 10/11/2023).
- 750 [44] M. Fuhrmans and D. Iglezakis, “Metadata4Ing - Ansatz zur Modellierung interoperabler
751 Metadaten für die Ingenieurwissenschaften,” en, Aug. 2020, Publisher: Zenodo. DOI:
752 [10.5281/zenodo.3982367](https://doi.org/10.5281/zenodo.3982367). [Online]. Available: [https://zenodo.org/record/398](https://zenodo.org/record/3982367)
753 [2367](https://zenodo.org/record/3982367) (visited on 10/11/2023).
- 754 [45] B. Schembera and D. Iglezakis, “EngMeta: Metadata for computational engineering,”
755 en, *International Journal of Metadata, Semantics and Ontologies*, vol. 14, no. 1, p. 26,
756 2020, ISSN: 1744-2621, 1744-263X. DOI: [10.1504/IJMSO.2020.107792](https://doi.org/10.1504/IJMSO.2020.107792). [Online].
757 Available: <http://www.inderscience.com/link.php?id=107792> (visited on
758 10/11/2023).
- 759 [46] S. I. H. Shah, V. Peristeras, and I. Magnisalis, “DaLiF: A data lifecycle framework for
760 data-driven governments,” en, *Journal of Big Data*, vol. 8, no. 1, p. 89, Dec. 2021, ISSN:
761 2196-1115. DOI: [10.1186/s40537-021-00481-3](https://doi.org/10.1186/s40537-021-00481-3). [Online]. Available: [https://jo](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00481-3)
762 [urnalofbigdata.springeropen.com/articles/10.1186/s40537-021-00481-3](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00481-3)
763 (visited on 10/13/2023).
- 764 [47] D. Schmitz and M. Politze, “Forschungsdaten managen – Bausteine für eine dezentrale,
765 forschungsnahe Unterstützung,” de, *o-bib. Das offene Bibliotheksjournal / Herausgeber*
766 *VDB*, vol. 5, no. 3, pp. 76–91, Sep. 2018, Number: 3, ISSN: 2363-9814. DOI: [10.5282](https://doi.org/10.5282/o-bib/2018H3S76-91)
767 [/o-bib/2018H3S76-91](https://doi.org/10.5282/o-bib/2018H3S76-91). [Online]. Available: [https://www.o-bib.de/bib/article](https://www.o-bib.de/bib/article/view/5339)
768 [/view/5339](https://www.o-bib.de/bib/article/view/5339) (visited on 02/23/2024).
- 769 [48] NFDI4Ing, *Trainings zum Datenlebenszyklus (Data Life Cycle, DLC)*. [Online]. Available:
770 [https://nfdi4ing.pages.rwth-aachen.de/education/education-pages/mai](https://nfdi4ing.pages.rwth-aachen.de/education/education-pages/main/html_slides/startpage_dlc.html#/)
771 [n/html_slides/startpage_dlc.html#/](https://nfdi4ing.pages.rwth-aachen.de/education/education-pages/main/html_slides/startpage_dlc.html#/) (visited on 02/22/2024).
- 772 [49] NFDI4Chem, *Research Data Life Cycle*, en-GB. [Online]. Available: [https://knowl](https://knowledgebase.nfdi4chem.de/knowledge_base/docs/data_life_cycle/)
773 [edgebase.nfdi4chem.de/knowledge_base/docs/data_life_cycle/](https://knowledgebase.nfdi4chem.de/knowledge_base/docs/data_life_cycle/) (visited on
774 02/22/2024).
- 775 [50] UK Data Service, *Research data management*, en-US. [Online]. Available: [https://u](https://ukdataservice.ac.uk/learning-hub/research-data-management/)
776 [kdataservice.ac.uk/learning-hub/research-data-management/](https://ukdataservice.ac.uk/learning-hub/research-data-management/) (visited on
777 02/22/2024).
- 778 [51] Scientific Data. “Data repository guidance.” ISSN: 2052-4463. (2021), [Online]. Avail-
779 able: <https://www.nature.com/sdata/policies/repositories> (visited on
780 04/01/2024).