ing.grid preprints

RESEARCH ARTICLE

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

Christian Backe [1]
Veit Briken [1]
Atefeh Gooran Orimi [2]
Rayen Hamlaoui [2]
Malte Wirkus [1]
Bilal Wehbe [1]
Frank 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:
2024-05-23

Keywords:
Research Data Management, RDM, 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

The FAIR principles are formulated in generic terms and without reference to any particular scientific discipline [1]. This is an advantage, because it makes them applicable to all sorts of domains. But it also means that communities who wish to implement the FAIR principles must develop their own standards and practices in order to meet their specific challenges and requirements. This development process can benefit from successful real-world examples that allow researchers and practitioners to identify, discuss, and select best practices.

This paper presents lessons learned from the creation and management of FAIR data and metadata in two recent field research projects, RoBivaL and DeeperSense, conducted at the Robotics Innovation Center (RIC) of the German Research Institute for Artificial Intelligence (DFKI). It is an extended and improved version of a presentation from the NFDI4Ing Conference 2023 [2]. The main goal is to derive principles and building blocks for a research data management (RDM) system that responds to the call for “rich” metadata in the FAIR sense [1], to be applied to similar projects, i.e., collaborative research in engineering disciplines producing field data.

After brief summaries of RoBivaL and DeeperSense in Section 2, focusing on their overall project objectives and contrasting their base data requirements, the main body of the paper is divided into three parts. Related work and specific desiderata are introduced separately for each part.

Section 3 discusses the content dimension of FAIR RDM. The primary distinction is between *executive* metadata necessary for producers to achieve their project goals, and *reusable* metadata necessary for reusers to satisfy the FAIR principles. Both metadata types are illustrated with examples from RoBivaL and DeeperSense. This section also develops base elements for a model of the metadata creation process at the micro level, which are compared with the ”processing step” class of the Metadata4Ing ontology.

Section 4 expands the distinction between different stakeholder groups from the previous section and explores the social dimension of *collaborative* FAIR RDM more broadly. We argue that a FAIR research data manager acts as a link between three social domains, where they perform different primary tasks.

Section 5 examines the time dimension of collaborative *and iterative* FAIR RDM at the macro level. Based on a critical appraisal of prominent data lifecycle models, we suggest a model of a self-improving data lifecycle geared towards collaborative and iterative RDM. The model is illustrated with lessons learned from RoBivaL and DeeperSense.

2 Project summaries

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

2.1 RoBivaL

The project RoBivaL [3] [4] was conducted between August 2021 and October 2023 by an interdisciplinary and multi-institutional team of roboticists and agriculture researchers in Germany. The project compared different robot locomotion concepts both from space research and agricultural applications on the basis of experiments conducted under agricultural conditions. Four robots were used: Two have their origins in space applications, the other two were developed specifically for agriculture. The robots were subjected to six experiments, addressing different challenges and requirements for agricultural applications. Soil conditions were controlled and varied on the two dimensions moisture (dry, moist, wet) and density (tilled, compacted). Table 1 gives an impression of selected experiments and robots in the field.

Field conditions and robot behavior were monitored with various sensors and measuring devices, partly on the robots and partly in the field, in order to document the experiment execution, and to determine the robot performance. The data capturing devices, their roles and deployments are

(a) (b) (c)

Table 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

summarized in Table 2. (Video camera and Lidar on the system are greyed out, because, although available, they were not used in the project.) The entire dataset, together with comprehensive metadata, is publicly available on the Zenodo platform [5].

	Device on System	**Device on System and in Field**	**Device in Field**
System Monitoring	• IMU • Force sensor	• RTK-GPS	• Stopwatch • Compass
System and Field Monitoring			• Video camera • Ruler
Field Monitoring	• Video camera • Lidar		• Tilt laser scanner • Penetrometer • Moisture meter

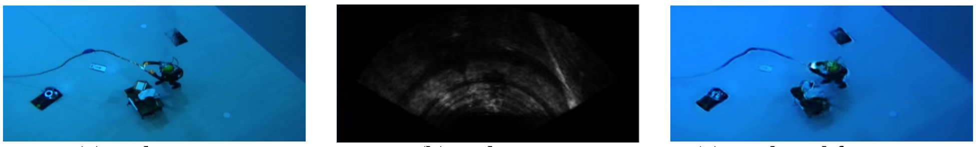
Table 2: RoBivaL data capturing devices by purpose and deployment

2.2 DeeperSense

The project DeeperSense [6] [7] was conducted between January 2021 and December 2023 by an international, interdisciplinary, and multi-institutional team of researchers and domain experts in Germany, Spain, and Israel. This paper focuses on the German use case, which employed roboticists, sensor experts, and technical divers. The objective was to improve the safety of the divers, who work under dangerous conditions and therefore require constant monitoring and assistance. Existing safety systems rely on cameras, which is a problem in turbid water that limits visibility – just when the divers most need outside support. Sonars are more robust to turbidity, but conventional sonar output is difficult to interpret. DeeperSense therefore developed a neural network which translates sonar output into images that appear camera-like, thus combining the best aspects of both modalities. Table 3 illustrates the sonar-to-image translation.

To gather training data, divers performing typical work tasks were recorded underwater with sonar and camera simultaneously. Figure 1 shows the training data collection setup schematically.

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



(a) real camera (b) real sonar (c) predicted from sonar

Table 3: 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

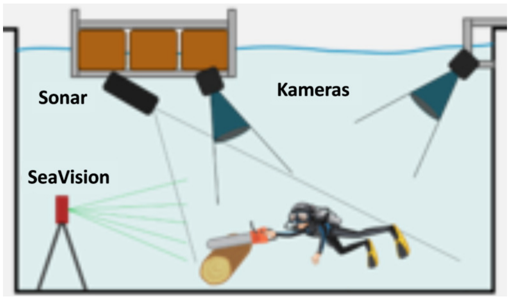


Figure 1: DeeperSense training data collection setup. © DFKI, Bilal Wehbe. License: CC BY 4.0 International

Selected parts of the sensor data were published on the Zenodo platform [8] [9]. Since it is currently impractical, due to its size, to make the entire sensor data corpus available online, the metadata was published as a standalone database [10], allowing researchers to select portions relevant for their use cases, which are made available on demand. This is an effort to comply with the FAIR principle A2 to make metadata accessible independently of the base data.

2.3 Comparison

Table 5 summarizes and constrasts the data-related properties of RoBival and DeeperSense as presented in Sections 2.1 and 2.2. An immediate takeaway is that the scope and form of the base data, its purpose and handling can be quite different between projects even at a single institute. A data management solution should be flexible enough to accommodate such variance.

3 Content dimension: Executive metadata and rich reusable metadata

This section discusses the content dimension of metadata creation and management in RoBivaL and DeeperSense from the perspectives of data producers on the one hand, and potential reusers as characterized by the FAIR principles on the other.

Subsection 3.1 introduces necessary background about high-level purposes of metadata, metadata semantics in the context of robotics and engineering in general, and the concept of metadata ”richness” according to the FAIR principles. We propose a mathematical model of richness designed to maximize the global FAIRness of metadata under economic and social constraints.

Subsection 3.2 lays out a collection of metadata topics from RoBivaL and DeeperSense for different purposes, and divides it into executive metadata relevant for producers, and reusable

Table 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

	RoBivaL	**DeeperSense**
Objective and method	• Compare robot performances using statistics and visualization	• Train neural network using multimodal machine learning
Base data	• Ten distributed sensor outputs and manual measurements observing robot behavior and field characteristics before, during, and after an experiment	• Indefinite stream of synchronized, co-located camera and sonar snapshots showing divers working underwater
Data acquisition in the field	• Multiple sessions at one location • Deliberate field preparation	• Multiple sessions and locations • Adapt to given field conditions

Table 5: Summary of data-related project properties of RoBival and DeeperSense

metadata for public consumers, based on the different motives of both parties. This analysis foreshadows the discussion of social aspects of metadata management in Section 4.

Subsection 3.3 attempts to model the process of metadata creation abstractly and at the micro level for production purposes. We illustrate our model with examples from DeeperSense and RoBivaL, and compare it to the communication-oriented ”processing step” class from the Metadata4Ing (M4I) ontology.

3.1 Metadata purposes, semantics, and richness

Virtually every general metadata definition starts with the assertion that metadata is ”data about data” [11] [12] [13] [14] [15]. A common purpose-based classification distiguishes at a high level between descriptive, administrative, and structural metadata [16] [17] [18]: Descriptive metadata ”enables discovery, identification, and selection of resources”. Administrative metadata ”facilitates the management of resources”. Structural metadata ”describes relationships among various parts of a resource”, and is ”generally used in machine processing”.

There is a broad spectrum of domain-specific approaches to model metadata semantics. In robotics, recent ontological efforts include [19] [20] [21] [22] [23]. For the communication of metadata in engineering disciplines including robotics, the NFDI4Ing community has developed the Metadata4Ing (M4I) ontology [24] [25] [26] [27]. It features a generalized process model, centered around the ”processing step” class. This is an attempt to communicate multi-stage data processing, to satisfy the FAIR principle R1.2 of detailed provenance tracking.

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

3.2 Executive metadata and reusable metadata

While the FAIR principles promote the development of metadata for data reusers, data producers already create and manage metadata routinely for their own purposes. What is the relationship between the *executive* metadata necessary for data production and the *rich* FAIR metadata supporting, enabling, or facilitating data reuse? This question has a content aspect and a form aspect: Which metadata topics are relevant for producers or reusers? Which formal requirements are demanded by either group? Since these questions address two different stakeholders, they foreshadow the discussion of social aspects of FAIR RDM in Section 4.

	Both projects	**RoBivaL**	**DeeperSense**
Producer	• Logistics (DAS) • Production standards, tools, workflows (A) • Team coordination (A) • Errors (DAS)	• Robot maintenance and development (A)	• High performance computing (A)
Producer and Reuser	• Field spec. (DAS) • Sensor spec. (DS) • Software spec. (DA) • Data format (DS) • Data statistics (A) • Related work (D)	• Experiment spec. (D) • Robot spec. (D) • Key robot properties, perf. metrics (D) • Measuring methods (D) • Source categorization (DAS)	• Machine learning methodology (DAS) • Scene description (D) • Sensor configuration (DS)
Reuser	• DOI, URL (A) • Usage license (A) • Provenance (A) • Public ontology (DS) • Extra use cases (DA)	• Tag unused data: by-catch, invalid runs (D)	• Typical examples (D)

Table 6: Metadata topics relevant for producers or reusers in RoBivaL and DeeperSense. Predominant purposes: descriptive (D), administrative (A), or structural (S).

The metadata topics of RoBivaL and DeeperSense are presented in Table 6, categorized by project and by relevance for producers or reusers. On both dimensions, intersections are possible. Each topic is labeled with its predominant purpose(s), i.e., descriptive (D), administrative (A), or structural (S).

The assignment of topics to producers or reusers is guided by the assumption that either group has a different primary motive: Producers want a *correct execution* of their project plan to achieve their primary research goal. Reusers want a *sufficient understanding* of the base data to assess its utility, and to integrate it into their own work flow.

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

3.3 Base elements of the metadata creation process

This section analyzes the process of metadata creation and derives some process-related metadata categories. The matter is treated abstractly and at the micro level, i.e., with regard to individual data elements; the big picture of the data lifecycle is discussed in Section 5. The analysis yields elements for the design of metadata production workflows. This is useful in a collaborative setting with a division of labor, where responsibilities must be communicated effectively.

The data flow diagram in Figure 2 illustrates the first order of metadata creation on a single data processing stage. The Output represents a piece of base data, which is generated by some Procedure. Output and Procedure are the subjects of metadata. For both, metadata creation has two phases: *Before* the subject exists, it is designed; *after* it exists, it may be documented. The design is metadata that is *injected* into the Procedure; the documentation is metadata that is *extracted* either from the Procedure or from the Output.

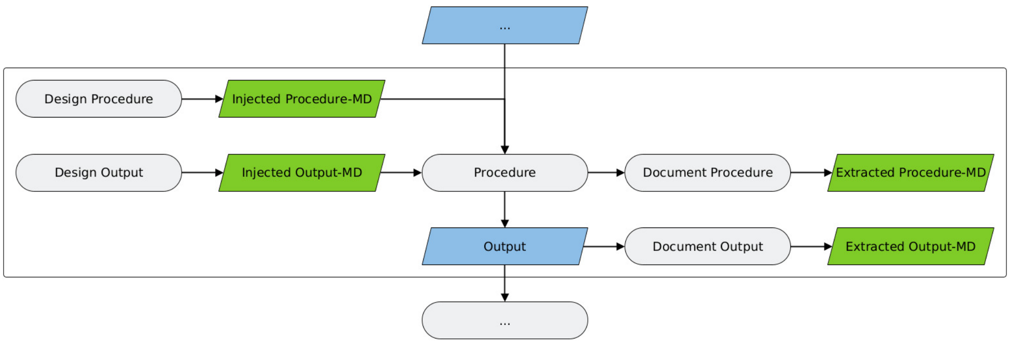


Figure 2: Data flow on a single stage of first order metadata creation. Rounded box: Process. Angled box: (Meta)data. Blue: Data. Green: Metadata. © DFKI, Christian Backe. License: CC BY 4.0 International

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

Table 7 lists examples for each of the four first order metadata categories, taken from the DeeperSense project. They are related to the same Procedure (”Capture camera and sonar images of a diver”) and corresponding Output (”Logfiles with raw camera and sonar data”).

	Procedure-related	**Output-related**
Injected	• Middleware (ROS 2.0) • Start time	• Raw data structure (”topic”) • File name (Identifier)
Extracted	• Scene description • Event documentation	• Number of recorded samples • File size

Table 7: Examples of the four first order metadata categories from the DeeperSense project

The assertion that metadata is data, as mentionend in Section 3.1, implies that metadata creation is recursive: Higher orders of metadata treat metadata of lower orders as their base data. Visually, this means we can stack the first order metadata creation model not just vertically, but also horizontally. This is illustrated in Figure 3. It contains a condensed version of Figure 2: The process ”Create MD” represents all four metadata creation processes of the first order, which are applied to the base Procedure and Output. ”First order MD” represents all four first order metadata types. The recursion recognizes that ”Create MD” and ”First order MD” themselves are a procedure-and-output pair, hence they become subjects of meta-metadata creation.

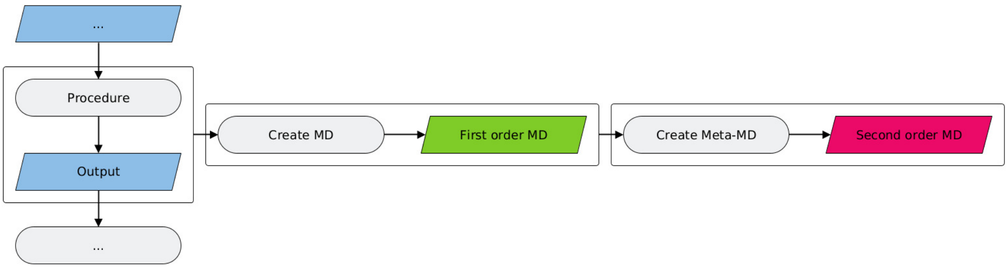


Figure 3: Recursive creation of higher order metadata. © DFKI, Christian Backe. License: CC BY 4.0 International

Table 8 gives two sets of generic examples for higher order metadata on multiple levels. The first example features pieces of literal base data and metadata: A speed measurement is taken at a certain time; the time stamp formatting is expressed in C string format notation; syntax and semantics of this formatting are governed by an ISO standard. The second example has a similar 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 8: Generic examples of higher order metadata

How does our metadata process model compare to the processing step class of the Metadata4Ing (M4I) ontology, depicted in Figure 4? The M4I model acknowledges that each data output is generated by a process, and that data processing may be chained, which corresponds to the vertical direction of our model. But the M4I model does not appear to cover the process of metadata creation itself, i.e., our model’s horizontal direction (injected vs. extracted metadata, higher order metadata). We assume this absence is at least partly a result of the purpose of the M4I model, which is *communication* of metadata to data consumers *after* the base data and metadata have been created. As mentioned above, the modularization of metadata in our model serves to design workflows for metadata *creation* by a data production team *during* project execution. A further difference between the M4I processing step and our model is that the former specifies a fixed set of attributes, while the latter is agnostic in this regard. Finally, the M4I processing step model provides the opportunity to encapsulate multipe substeps into a single step of larger scale. So far, our model does not feature a similar means of abstraction.

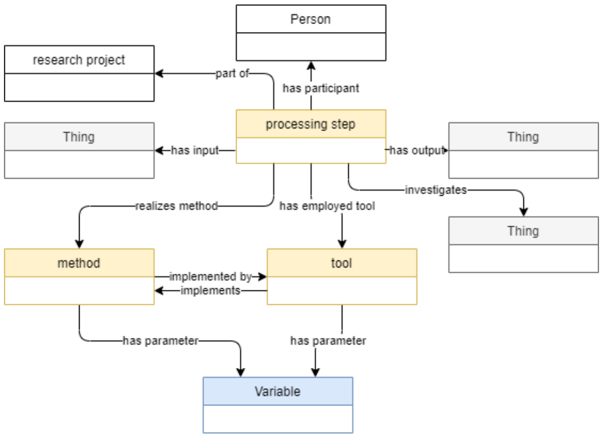


Figure 4: Processing step class of the Metadata4Ing ontology [24]. © Metadata4Ing Workgroup. License: CC BY 4.0 International

4 Social dimension: Collaborative FAIR data management in field research

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

4.1 Collaboration with the data production team

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

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

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

Figure 5 gives an impression of the atmosphere in the context of DeeperSense. In the final hours of the last day of data collection, the team is working against the clock to gather a critical piece of data necessary for the final demostration event. They are on a boat on a lake, wearing life vests, trying, against the glaring sun, to analyze their monitoring information on a computer screen, to figure out why the underwater system keeps failing.

Figure 5: Field team work in DeeperSense. © DFKI, Christian Backe. License: CC BY 4.0 International

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

4.2 Mediation between producers and reusers

The previous section dealt with RDM in general. For a FAIR data manager in particular, there is a second social domain, namely the larger research domain. Here, their primary task is *mediation* between conflicting requirements of their data production team on the one hand, and potential data reusers (as characterized by the FAIR principles) on the other.

The root of this conflict was sketched in Section 3.2 and can be extended in light of the discussion from Section 4.1: Reusers require explicit, formal, rich metadata to thoroughly understand the data that is foreign to them, easily interface with it using machines, and have it serve a broad spectrum of potential use cases. But this demands extra effort from the producers, who not only have the privilege of being more implicit, informal, and brief in their internal communication, but who may actually be forced to cut corners, especially under field conditions, in order to reach their primary research goal.

Table 9 summarizes this proposition and adds two aspects derived from experience in RoBivaL and DeeperSense: In a collaborative setting with division of labor, executive metadata may be distributed over many places convient for different contributors; to become reusable, it must be consolidated. While research is ongoing, the executive metadata design may need to evolve to adapt to changing circumstances; reusers prefer reliable APIs.

The conflicting priorities and requirements of data producers and reusers have two implications for a FAIR research data manager: First, they must motivate their team to apply the extra effort.

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 9: Different priorities and requirements of data producers and reusers

One possible incentive may be that today's producers are their own reusers tomorrow, so the investment in more elaborate metadata will pay off directly towards themselves. There is an indirect version of this: By creating metadata they would be happy to receive if they were reusers, producers influence the standards of their community to their own benefit. Another incentive may be increased impact of their research if the underlying data is broadly adopted in the community. A second implication is that FAIR research data managers must design the workflow of their team such that the extra effort necessary to satisfy reuser requirements does not coincide with peak effort towards the primary research goal, because the latter will always have precedence.

4.3 Standardization in the FAIR RDM community

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

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

5 Time dimension: A self-improving data lifecycle

This section divides FAIR research data management into different tasks and organizes them across time. Subsection 5.1 discusses the concept of a data lifecycle, and proposes some modifications to the type of data lifecycle currently used by NFDI4Ing and similar parties. The two main modifications are (1) the introduction of an *internal* data provision phase necessary for collaborative research, and (2) the introduction of an *evaluation* phase to drive an iterative improvement of the RDM system. Subsection 5.2 presents some lessons learned from RoBivaL and DeeperSense in each phase.

5.1 Model of a self-improving data lifecycle

There is no consensus which phases constitute a data lifecycle, and how the phases shall be ordered. In their recent survey of 76 data lifecycles, Shah et al. identify at least 14 distinct phases [28]. NFDI4Ing uses a model with six phases, named (1) Planning, (2) Production, (3) Analysis, (4) Storage, (5) Access, and (6) Re-Use [29]. It is similar to other six-phase models prevalent in the FAIR RDM community [30] [31] [32] [33] [34], but there are still differences about the naming and ordering of the phases.

In our view, these models have at least two shortcomings if they shall be applicable to collaborative and iterative research.

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

Second, while the remaining phases are all actions that apply *to data* (data is produced, is analyzed, ...), Planning is the only phase that applies *to other actions* (production is planned, analysis is planned, ...). Further, it is a missed opportunity to only have a phase dedicated to looking into the future (Planning), but none into the past (*Evaluation*). In iterative research, reflecting on the difference between how things were planned and how they actually turned out, would enable an *iterative improvement* of the data management system and of the research system in general. Since our research is in fact iterative, within a single project and over multiple projects, such a self-improving data lifecycle would be welcome.

Our proposed model has six data-related phases: (1) Creation, (2) Provision, (3) Processing, (4) Publication, (5) Reuse, and (6) Archiving. Each of these is divided into three process-related phases: (1) Planning, (2) Execution, and (3) Evaluation. The model is illustrated in Figure 6.

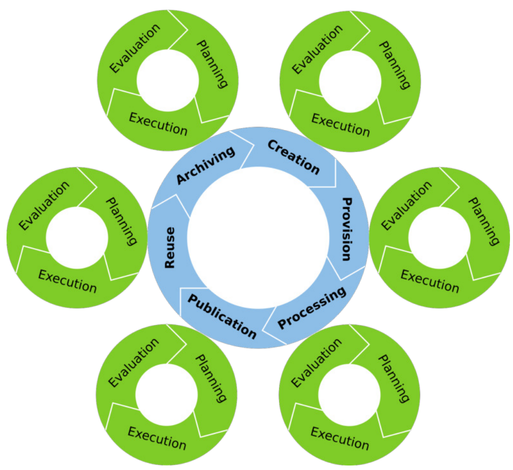


Figure 6: Self-improving nested data lifecycle. © DFKI, Christian Backe. License: CC BY 4.0 International

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

5.2 Lessons learned from RoBivaL and DeeperSense

This section applies the data lifecycle model discussed in Section 5.1 to RoBivaL and DeeperSense, by presenting selected lessons learned from both projects in the different lifecycle phases.

5.2.1 Creation

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

Data management needs to specify the scope and form of the metadata set (see Section 3), and provide tools and procedures for metadata creation. In field research, the environmental conditions and the data creation process need to be documented more extensively than in the lab, because there are fewer means of control and more chances of surprise.

Terms coined during planning will propagate through a growing corpus of communication, documentation, and implementation. To avoid costly changes later, it is advisable to stabilize the terminology early on. *FAIR* terminologies must reflect community practices in their domains. This can either facilitate planning, if a communal terminology already exists; or it can complicate planning, if a terminology first needs to be compiled from scholarly sources.

Terminology requirements may be different for humans and machines. Machines need more consistency, less ambiguity, and may accept only restricted token sets. Since humans are more flexible, a machine-consumable version is preferable for co-processed information, e.g., file and directory names. In this case, human collaborators must be educated on machine requirements.

Execution In a collaborative setting, different people may observe different features of an object. Having a single person responsible to record all observations can help avoid misalignment, ensure that values are consistent and complete, and that standards e.g., related to accuracy or measuring units are followed uniformly. The information relay requires structured communication, and routines to avoid, detect, and correct miscommunication.

A designated record keeper can also take note of problems and unforeseen events, which may help improve the planning of future data creation sessions. This may be performed proactively, by looking out for and trying to prevent errors in the first place. As discussed in Section 4.1, field data creation can be cognitively very taxing, so it is easy to miss e.g., a critical failure of a single component. Therefore, having someone specifically focussed on error detection is useful.

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

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

5.2.2 Provision

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

Logically, the repository can be specified with different resolutions, on multiple layers and domains. Aspects to consider may be file trees, database schemas, and request APIs; encodings, types, and formats; sources and processing stages; separation of base data and metadata; auxiliary assets (e.g., documentation, specification, schemas, logs, errors). The terminology should be consistent between layers and domains, and be compatible with the data creation terminology.

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

Execution An explicit specification of the physical and logical layout can improve team alignment. User onboarding is an opportunity to check if the specification is properly understood and reflects the actual requirements. The layout and its specification may need to be updated to account for e.g., larger volume, changing pipelines, different data formats, etc.

DeeperSense and RoBivaL developed dedicated metadatabases to facilitate reporting (e.g., volume per data layer, sample count per sensor type and session, runs per experiment and robot). They provide information for decision-making, both by the executing researchers (e.g., are there critical data gaps?) and reusers (e.g., is this dataset suitable for my use case?). As standalone items, they can be transmitted separately from the large base data corpora.

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

But there is tension: Research data processing must be flexible enough to adjust to new findings and changing views. In interdisciplinary research, practices from different domains must be accommodated. Too much specification too early or too rigidly may lower the acceptance and adoption of a layout. Further, writing a comprehensive, accurate, and understandable specification may be difficult and time-consuming, thus conflicting with other priorities. On the other hand, working with undocumented, inconsistent layouts that need to be reverse engineered and might change without notice, lowers productivity and risks producing bad results.

5.2.3 Processing

Planning Processing resources must be supplied for different tasks and stages. This includes individual workstations for all team members, and high performance servers that are used as a shared resource.

If there are multiple contributors, it is important to specify who is responsible for which processing job, and what are the interfaces between consecutive steps in a processing pipeline.

Execution One core responsibility of the data manager is metadata processing. In RoBivaL and DeeperSense, this was done in the context of developing and maintaining a metadatabase, involving schema design, metadata extraction, fusion, and aggregation. The data manager may also be tasked with quality assurance, which affects all other processing jobs. This involves the conception of error cases, error logging, escalation of errors, and resolution management.

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

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

5.2.4 Publication

Planning The data repository where the data and metadata are to be published should match the given content. Where will the intended reusers be likely to look for data to match a certain use case? The journal Scientific Data recommends various data repositories geared towards particular natural and social sciences, as well as some generalist repositories [35]. For large datasets, space constraints by different repositories may have to be considered.

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

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

5.2.5 Reuse

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

5.2.6 Archiving

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

6 Conclusion

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

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

One risk to consider here is the possibility of scope explosion in multiple directions: Firstly, since *executive* metadata covers many areas, metadata management for internal purposes might soon turn into general knowledge management. Secondly, since *rich* metadata lacks a comprehensive definition and is grounded in potential needs of data reusers, it is difficult to judge what must be included and what may be omitted. Thirdly, *higher order* metadata implies an infinite recursion, which must be capped at a level that is reasonable for different stakeholders.

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

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

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

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

7 Acknowledgements

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

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

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

8 Roles and contributions

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

Veit Briken: Conceptualization, Writing - review & editing

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

Rayen Hamlaoui: Investigation, Writing - review & editing

Malte Wirkus: Data curation, Funding acquisition, Investigation, Project administration, Software, Writing - review & editing

Bilal Wehbe: Data curation, Funding acquisition, Investigation, Project administration, Visualization, Writing - review & editing

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

References

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

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

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

[4] M. Wirkus, S. Hinck, C. Backe, *et al.*, "Comparative study of soil interaction and driving characteristics of different agricultural and space robots in an agricultural environment," *Journal of Field Robotics*, pp. 1–34, DOI: https://doi.org/10.1002/rob.22347. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.22347. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.22347.

[5] C. Backe, M. Wirkus, S. Hinck, *et al.*, *RoBivaL data corpus*, https://zenodo.org/record/8424933, Experiment Data, Mar. 2024. DOI: 10.5281/zenodo.8424933.

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

[7] B. Wehbe, N. Shah, M. Bande, and C. Backe, "Sonar-to-RGB Image Translation for Diver Monitoring in Poor Visibility Environments," in *OCEANS 2022, Hampton Roads*, ISSN: 0197-7385, Oct. 2022, pp. 1–9. DOI: 10.1109/OCEANS47191.2022.9977024. [Online]. Available: https://ieeexplore.ieee.org/document/9977024 (visited on 12/10/2023).

[8] B. Wehbe, N. Shah, M. Bande, and C. Backe, *Sonar-to-RGB Image Translation for Diver Monitoring in Poor Visibility Environments*, Mar. 2023. DOI: 10.5281/zenodo.7728089. [Online]. Available: https://zenodo.org/records/7728089 (visited on 12/10/2023).

[9] O. Köken and B. Wehbe, *Fusion of Underwater Camera and Multibeam Sonar for Diver Detection and Tracking*, version 1, Zenodo, Dec. 2023. DOI: 10.5281/zenodo.10220989. [Online]. Available: https://doi.org/10.5281/zenodo.10220989.

[10] C. Backe, B. Wehbe, M. Bande, N. Shah, D. Cesar, and M. Pribbernow, *Metadata of a Large Sonar and Stereo Camera Dataset Suitable for Sonar-to-RGB Image Translation*, Zenodo, Dec. 2023. DOI: 10.5281/zenodo.10373154. [Online]. Available: https://doi.org/10.5281/zenodo.10373154.

[11] J. Furner, “Definitions of ”Metadata”: A Brief Survey of International Standards,” en, *Journal of the Association for Information Science and Technology*, vol. 71, no. 6, Jun. 2020, ISSN: 2330-1635, 2330-1643. DOI: 10.1002/asi.24295. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24295 (visited on 10/11/2023).

[12] K. G. Jefferey and R. Koskela, “RDA Metadata Principles and their Use,” en, Research Data Alliance, Tech. Rep., Nov. 2014, p. 6. [Online]. Available: https://rd-alliance.org/metadata-principles-and-their-use.html (visited on 06/05/2023).

[13] RFII - Rat für Informationsinfrastrukturen, *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*, de. Göttingen, May 2016.

[14] OECD, *Data and Metadata Reporting and Presentation Handbook*, en. OECD, Jun. 2007, ISBN: 978-92-64-03032-9 978-92-64-03033-6. DOI: 10.1787/9789264030336-en. [Online]. Available: https://www.oecd-ilibrary.org/economics/data-and-metadata-reporting-and-presentation-handbook_9789264030336-en (visited on 10/11/2023).

[15] National Institute of Statistical Sciences, “Metadata and Paradata: Information Collection and Potential Initiatives,” en, National Institute of Statistical Sciences, Expert Panel Report, Nov. 2010, p. 36. [Online]. Available: https://www.niss.org/research/metadata-and-paradata-information-collection-and-potential-initiatives (visited on 10/11/2023).

[16] W3C, *Data on the Web Best Practices*, en, Jan. 2017. [Online]. Available: https://www.w3.org/TR/dwbp/ (visited on 02/20/2024).

[17] J. Riley, *Understanding Metadata: What is Metadata, and What is it For?* (NISO Primer Series), en-US. National Information Standards Organization (NISO), 2017, ISBN: 978-1-937522-72-8. [Online]. Available: https://www.niso.org/publications/understanding-metadata-2017 (visited on 02/20/2024).

[18] M. Cofield, *Metadata Basics*, en, Feb. 2024. [Online]. Available: https://guides.lib.utexas.edu/metadata-basics (visited on 02/20/2024).

[19] M. Yüksel, “An expert knowledge representation based on real world experiences in proof-of-concept robot system design,” Available at https://doi.org/10.26092/elib/2337, PhD thesis, Universität Bremen, Jun. 2023.

[20] Open Robotics, *URDF XML Specifications*, Sep. 2022. [Online]. Available: http://wiki.ros.org/urdf/XML (visited on 10/16/2023).

[21] S. R. Fiorini and V. Fortes Rey, *OWL specification of the Core Ontology for Robotics and Automation (CORA) and other IEEE 1872-2015 ontologies*, original-date: 2017-10-20T08:23:09Z, Dec. 2023. [Online]. Available: https://github.com/srfiorini/IEEE1872-owl (visited on 02/20/2024).

[22] J. Malec, *The ROSETTA ontology of industrial robotics devices and robotic skills*, original-date: 2013-10-11T13:52:05Z, Feb. 2023. [Online]. Available: https://github.com/jacekmalec/Rosetta_ontology (visited on 02/20/2024).

[23] IEEE, "IEEE Standard Ontologies for Robotics and Automation," *IEEE Std 1872-2015*, pp. 1–60, 2015. DOI: 10.1109/IEEESTD.2015.7084073. [Online]. Available: https://ieeexplore.ieee.org/document/7084073 (visited on 02/20/2024).

[24] S. Arndt, B. Farnbacher, M. Fuhrmans, *et al.*, "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity.," en, Sep. 2023, Publisher: Zenodo Version Number: 1.2.0. DOI: 10.5281/ZENODO.5957103. [Online]. Available: https://zenodo.org/record/5957103 (visited on 10/11/2023).

[25] D. Iglezakis and S. Arndt, *How to Use Metadata4Ing - First Steps Tutorial*, en, Mar. 2023. [Online]. Available: https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing/-/blob/1.1.0/training/first-steps-guide.md (visited on 10/11/2023).

[26] M. Fuhrmans and D. Iglezakis, "Metadata4Ing - Ansatz zur Modellierung interoperabler Metadaten für die Ingenieurwissenschaften," en, Aug. 2020, Publisher: Zenodo. DOI: 10.5281/ZENODO.3982367. [Online]. Available: https://zenodo.org/record/3982367 (visited on 10/11/2023).

[27] B. Schembera and D. Iglezakis, "EngMeta: Metadata for computational engineering," en, *International Journal of Metadata, Semantics and Ontologies*, vol. 14, no. 1, p. 26, 2020, ISSN: 1744-2621, 1744-263X. DOI: 10.1504/IJMSO.2020.107792. [Online]. Available: http://www.inderscience.com/link.php?id=107792 (visited on 10/11/2023).

[28] S. I. H. Shah, V. Peristeras, and I. Magnisalis, "DaLiF: A data lifecycle framework for data-driven governments," en, *Journal of Big Data*, vol. 8, no. 1, p. 89, Dec. 2021, ISSN: 2196-1115. DOI: 10.1186/s40537-021-00481-3. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00481-3 (visited on 10/13/2023).

[29] D. Schmitz and M. Politze, "Forschungsdaten managen – Bausteine für eine dezentrale, forschungsnahe Unterstützung," de, *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, vol. 5, no. 3, pp. 76–91, Sep. 2018, Number: 3, ISSN: 2363-9814. DOI: 10.5282/o-bib/2018H3S76-91. [Online]. Available: https://www.o-bib.de/bib/article/view/5339 (visited on 02/23/2024).

[30] NFDI4Ing, *Trainings zum Datenlebenszyklus (Data Life Cycle, DLC)*. [Online]. Available: https://nfdi4ing.pages.rwth-aachen.de/education/education-pages/main/html_slides/startpage_dlc.html#/ (visited on 02/22/2024).

[31] NFDI4Ing, *NFDI4Ing-Schulungsplattform*. [Online]. Available: https://nfdi4ing.pages.rwth-aachen.de/education/education-pages/main/html_slides/startpage.html#/ (visited on 02/22/2024).

[32] NFDI4Chem, *Research Data Life Cycle*, en-GB. [Online]. Available: https://knowledgebase.nfdi4chem.de/knowledge_base/docs/data_life_cycle/ (visited on 02/22/2024).

[33] UK Data Service, *Research data management*, en-US. [Online]. Available: https://ukdataservice.ac.uk/learning-hub/research-data-management/ (visited on 02/22/2024).

[34] UK Data Service, *Research Data Lifecycle*, Aug. 2019. [Online]. Available: https://www.youtube.com/watch?v=-wjFMMQD3UA (visited on 02/22/2024).

[35] Scientific Data. "Data repository guidance." ISSN: 2052-4463. (2021), [Online]. Available: https://www.nature.com/sdata/policies/repositories (visited on 04/01/2024).