ing.grid

RESEARCH ARTICLE

Searching for a vision on research data management providing guidance, decision support, tools and automation

A systematic literature review on solutions for research data management

Tobias Hamann [1]
Katja Jansen [2]
Giacomo Lanza [3]
Pia Carina Pickmann [1]
Ilona Lang [2]
Marcos Alexandre Galdino [1]
Anas Abdelrazeq [1]
Robert H. Schmitt [1, 4]

1. Chair for Intelligence in Quality Sensing (IQS), WZL, RWTH Aachen University, Aachen, Germany.
2. IT Center, RWTH Aachen University, Aachen, Germany.
3. Pyhsikalisch-Technische Bundesanstalt National Metrology Institute (PTB), Braunschweig, Germany.
4. Fraunhofer Institute for Production Technology IPT, Aachen, Germany.

Date Published:
-

Reviewer:
-

Keywords:
Engineering Research, Research Data Management, Systematic Literature Review, Toolchain, Research Data Management Process

Data availability:
-

Software availability:
-

Abstract.
Arising demands of funding organisations, universities and institutions towards research data management (RDM) force researchers to provide reusable data. Jarves provides the overall process with decision. RDMO and Coscine add DMPs, data annotation and storage respectively. Progressing digitalisation and newly arising technologies are continually increasing the demand for reusable data and the requirement for proper research data management (RDM). Despite the additional pressure applied by funding organisations, universities and institutions, researchers are still adapting to this cultural change. Their main challenges are insufficient guidelines in RDM, a lack of RDM-knowledge and the additional effort required. In the present systematic literature review we screened a total of 2.409 records and extracted data from 88 of those records. From there on, we compare features and limitations of 9 selected solutions. We then define some criteria for an all-comprehensive research data management (RDM)-solution: guidance throughout the RDM process, management of influencing factors on this process, RDM decisions support, tailored training materials and partial automation. Afterwards, we propose a self-developed toolchain which tries to address all the listed aspects, which is implemented by the connection of the three existing tools Jarves (for workflow management), RDMO (for data management plans) and Coscine (for data organisation).

1 Introduction

The growing importance of RDM is undisputed [1], [2], [3], [4], [5]. Funding organisations demand RDM in funded projects, and universities and institutions require researchers to manage their research data [6], [7], [8]. Meanwhile, not only demands of funding organisations play an important role in RDM, but new technologies such as artificial intelligence (AI) or machine learning (ML) are emerging, which, in turn, require data management [9], [10]. Additionally, good data management enhances the reusability of data, making it's collection more sustainable. In engineering sciences, data management is especially important, as huge amounts of data from various data collection methods can arise. While these amounts of data offer great potential for data reuse, they also imply a careful data management due to their heterogeneity and often complexity.

Initiatives like the National Research Data Infrastructure (NFDI) or the Research Data Alliance (RDA) are trying to close the gap between requirements to be met and needs of researchers [11], [12], [13]. New concepts, solutions and use cases are presented frequently, addressing a wide audience. These solutions include tools and services for RDM-specific problems. The National Research Data Infrastructure for the Engineering Sciences (NFDI4ING) aims to facilitate RDM by providing RDM-solutions tailored to engineering [12], [14]. NFDI4ING has been proposing many solutions, for example the Terminology Service [1], the Data Collections Explorer [2] or even this journal, ing.grid.

However, researchers are still trying to figure out how to manage their data or yet have to adapt to the current cultural change in research towards Findable, Accessible, Interoperable, Re-usable (FAIR) data management [1], [15], [16]. In a previously conducted study of Hamann et al. [17], it was shown that researchers of engineering sciences still need support in their RDM and perceive existing RDM support as lacking [17], [18]. Researchers also consider RDM as a burden and additional effort [2], [15], [16], [19] , eventually preventing them from conducting actual research across different engineering fields [17], [18]. A common starting point, a clear approach to RDM and reduction rather than increase of effort in RDM are demanded by researchers. Figure 1 summarises researchers' demands to three challenges.

On the one hand, insufficient guidelines lack explicit recommendations for action and are often not a sufficient incentivation for researchers to conduct RDM [16], [20], [21], [22]. On the other hand, researchers lack knowledge on RDM, both in regards of common or best practices and available tools or solutions [15], [19], [20], [22]. Both of these challenges force researchers to investigate their boundary conditions and applying factors on RDM, causing additional effort [2], [16], [20]. This effort itself is a difficulty, as RDM itself also requires activities to be carried out [2], [15], [19]. Additionally, the increased workload of entering the same sets of information into multiple tools is not only repetitive but further hinders the dissemination of RDM, opposing the demand for reduced effort and the need for support in RDM [2], [15], [17], [18], [21], [22].

To address the challenges shown in figure 1, we proposed a toolchain of three RDM-tools on the 2023 NFDI4ING conference. [23] In our vision, such a toolchain would enable researchers to

[1] `https://terminology.nfdi4ing.de/ts/`
[2] `https://data-collections.nfdi4ing.de/`

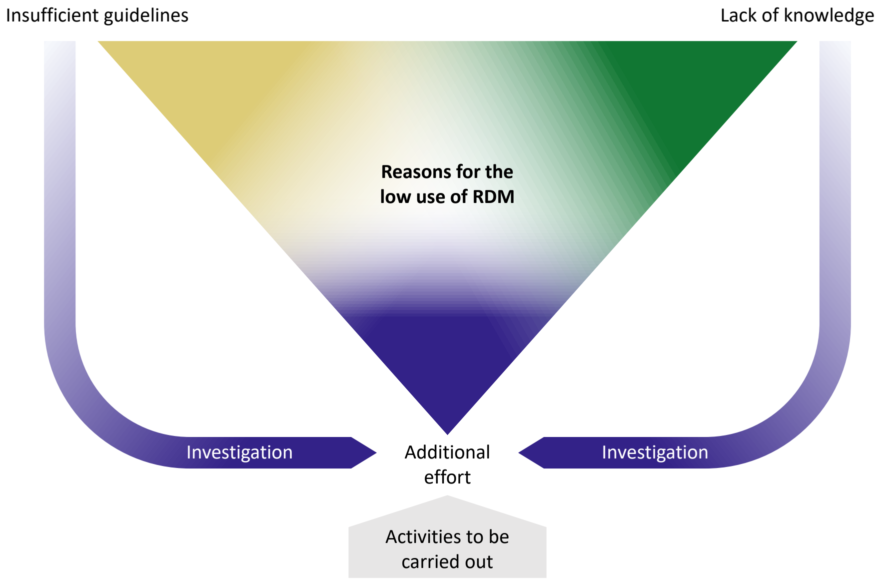

Figure 1: Challenges of the researchers in engineering sciences

conduct RDM more conveniently and more efficiently. This toolchain would be scientifically tailored to the needs and demands of researchers, not only but also in engineering sciences. It would contain a guided process for RDM that closely related to the day-to-day activities of researchers. Furthermore, it would provide decision support based on the boundary condition of the individual research project. Lastly, it would connect multiple services. The tools included as an example for this paper are (1.) the Research Data Management Organizer (RDMO) for creation of Data Management Plans (DMPs) and (2.) the RDM platform Coscine for data storage, metadata management and archiving, both connected through (3.) the Joint Assistant for Research in Versatile Engineering Sciences (Jarves). This toolchain demonstrates a concept that connects tools along a guided RDM-process that resembles the researcher's day-to-day activities for seamless integration.

Our hypothesis is, that such a toolchain does not exist yet. Hence, we investigate the state of the art on similar existing RDM-solutions. We specifically investigate solutions that address the challenges brought up by the researchers. This will clarify, why existing solutions are not able to address the challenges of researchers. Furthermore, the resulting research gap can be outlined. To address this gap, this article presents a merger of three RDM-tools into a toolchain, as outlined at the NFDI4ING conference in 2023 [23]. Following this rationale, we defined three questions to be answered by the paper:

1. What solutions are available to support RDM in engineering sciences?
2. Are there solutions that address the challenges of the researchers [cf. 17]?
3. What are the existing solutions missing to fully address the challenges of the researchers [cf. 17]?

In this paper, we present Existing RDM solutions. We do so by firstly presenting your Research methodology before discussing existing solutions in section 3, eventually pointing out the "Research gap" in literature. Afterwards, we present our concept of the "Proposed toolchain".

Three tools were chosen and included to portray a minimal working example of the toolchain as not just a bilateral data exchange but an interconnected network of tools. Lastly, we give a "Summary and Outlook".

2 Research methodology

To address our vision of an interconnected RDM toolchain, we have researched on previous work regarding the same or similar topics, ranging from solutions that are similar to our proposed toolchain to software frameworks which also provide ideas on how to approach the aforementioned challenges. This research was conducted as a systematic literature review (SLR). The aim of this review is the collection of existing solutions for the identified challenges (see figure 1) of insufficient guidance, lacking knowledge and additional effort. This additional effort is caused by the conduction of RDM itself and is aggravated both by insufficient guidelines and lack of RDM-knowledge amongst researchers.Objective is the identification of existing solutions, the investigation of the ability of the found solutions to address the challenges of the researchers and the clarification of the research gap in the context of these challenges. Furthermore, the SLR aims for the collection of knowledge from within and possibly beyond engineering. Therefore, the domain of engineering along adjacent domains like Science, Technology, Engineering and Mathematics (STEM) are considered for this review, collecting knowledge beyond the scope of the problem to utilise it, if possible, for the solving of the identified problem.

2.1 Collection of literature

The literature was conducted in February 2024 and followed the Checklist from the PRISMA approach by Page et al. [24], [25]. The questions posed for the SLR are formulated in the Introduction (see research questions 1. to 3.). In addition to the questions, inclusion and exclusion criteria have to be formulated, as shown in the appendix A. Then the search strategy was defined and literature was collected. The sources searched or consulted are also listed in appendix A. Afterwards, this literature was screened, firstly by abstract. The remaining literature was then subjected to full text screening. This process was handled by two researchers independently, while conflicts were afterwards resolved by the lead reviewer. Lastly, information was extracted for the selected literature.

Figure 2 shows the PRISMA 2020 flow diagram following Page et al. [25] but adjusted to the sources and procedures used in this literature review. As shown in this diagram, a total of 2.409 records have been identified of which one record was not retrieved, as this record was not publicly available nor retrievable by other means, e.g. by library access. The most important steps of this process are presented in the Appendix.

2.2 Coding of the found literature

The found literature was coded using the "coding reliability" [26] thematic analysis approach as presented by Braun and Clarke [26]. This approach is especially suited when "topics [...] map closely onto data collection questions" [26], as it is the case with this SLR. The utilisation of this approach implies the deduction of themes from the aforementioned questions.

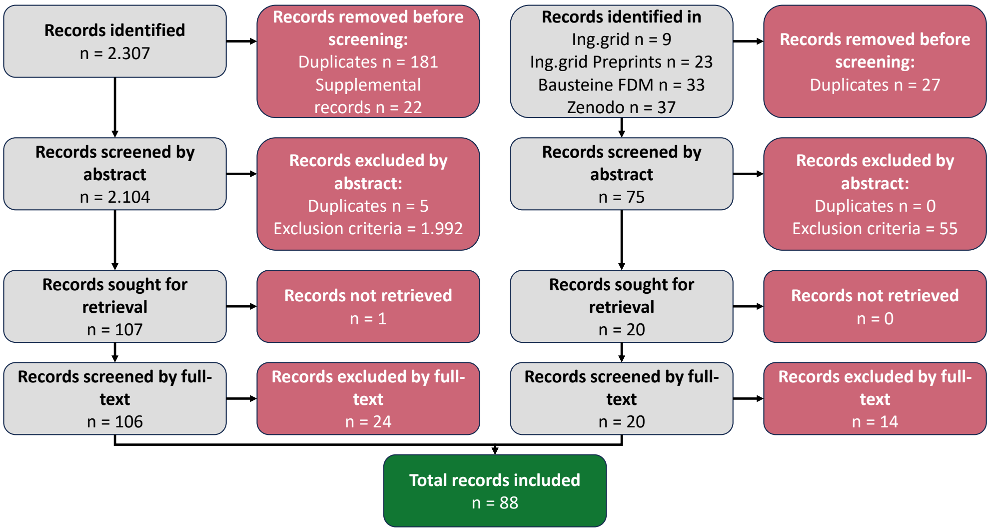

Figure 2: PRISMA Flow Diagram of the systematic literature review following [25]

As the questions revolve around solutions for RDM, the first code should represent the solutions for RDM. However, as the term “solution” is rather vague and the aim of the SLR is to investigate solutions addressing challenges brought up by the researchers, the search terms as well as the code have been specified. For the inclusion criteria, solutions have been searched for by the terms “Framework”, “Groundwork”, “Structure”, and “System” (see table 3). The term “Framework” represents all other terms as it can refer to a joint combination of multiple concepts or solutions. Besides that, “framework” can also refer to software frameworks. Hence, **Framework** is the first code for the thematic analysis.

Accordingly, every code stands for a general thematic area. For instance, anything that resembles a framework, toolchain or combination of solutions is coded as framework. Secondly, the **Methods** used for the creation of a new solution are considered of relevance, as the aim of this review is to identify scientific approaches to the researchers’ problems, which is why **Methods** is the second code. Lastly, a coding for support for researchers was included. As Hamann et al. [17] found, this support comes down to guidance on how to conduct RDM. One way to provide such guidance is by the provision of **Workflows**, which is the third code and encompasses any support in the conduction of RDM. Again, the code **Workflows** generally refers to any structured processes, workflows or procedures. Lastly, the conclusions drawn from different sources can be valuable for the vision to be developed, which is why key findings of literature are coded as **Conclusion**. These conclusions include results, findings and developed solutions of the respective papers but also open questions. With this coding, literature can be reviewed for existing frameworks, their scientific creation method, included workflows or processes and key results of the respective literature.

Beyond the thematic analysis, certain metadata was collected for each piece of literature, namely release date, type of literature, methodology or methods used, discipline and country of origin. Figure 3 depicts the percentages of records by discipline, showing that 35% of literature is based interdisciplinary research.

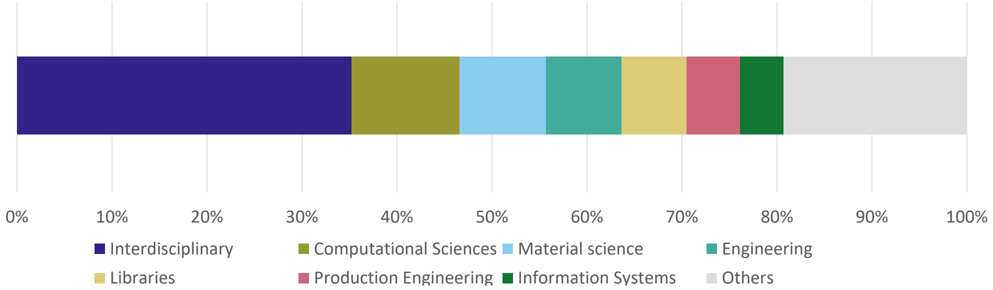

Figure 3: Included literature by discipline

The other 65% of records are specific solutions for their individual disciplines, e.g. Computational Sciences. Corresponding to the high number of specific solutions for disciplines, the majority of records do not actually present their method but solutions to the problems they emphasise. As shown in figure 4, 62% of records do not clearly specify the methods used to generate the results shown. While these solutions are mostly well motivated, and provide improvements for RDM, they only introduce new tools (36 records) or concepts (18 records). Computational sciences and material science are found slightly more frequent than other disciplines, both in numbers of records included and the number of records providing tools or concepts for RDM. Two thirds of records originating from computational sciences and 75% of material science records introduce a new tool for RDM, like HELIPORT, CAESAR or LinkAhead.

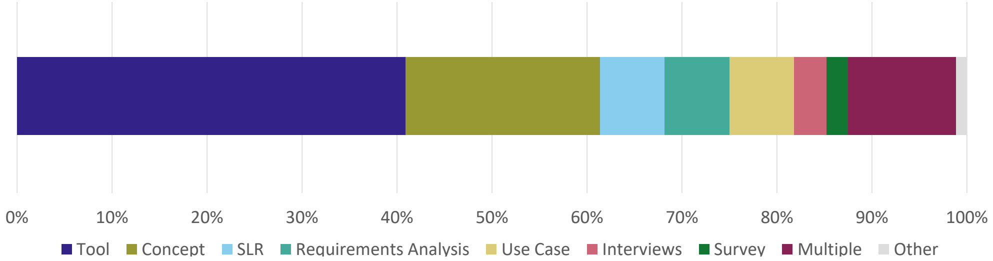

Figure 4: Included literature by methods used

3 Existing RDM solutions

Following the descriptive analysis of the literature found, this section describes existing solutions for RDM. As mentioned before, there are plenty of generic as well as specialised solutions. Solutions with a fit to the proposed questions are presented and described in this section, in decreasing order of referencing frequency. Due to page limitations, a representative sample of solutions found will be presented. Solutions not contributing to the proposed questions will not be described.

Comprehensive Knowledge Archive Network (CKAN) CKAN[3] is cited by eleven of the included records . CKAN is “a tool for making open data websites” [27], which “helps [... to] manage and publish collections of data” [27]. The CKAN website claims that this tool is used by several governments, namely the Australian, Canadian, United States and the Singapore

[3] https://ckan.org/

government [28], as well as other institutions and companies use CKAN, for example LEGO® [28]. In its core, CKAN focusses around datasets, containing metadata and resources, i.e. the data itself. Due to CKANs generic nature, it is a supporting solution to be considered when building a solution for RDM rather than a solution for RDM itself.

An RDM solution based on CKAN is described in four articles by Mozgova et al. [29], [30] and Sheveleva et al. [1], [31]. It is defined as a central research data management system (RDMS), containing a “Knowledge Management System (KMS) based on Semantic MediaWiki (SMW) and [a] Data Management System (DMS) based on the CKAN software” [30]. This RDMS is based on users’ needs collected through surveys, interviews, stakeholder analysis and workshops. Furthermore, this RDMS has been tailored on a use case study to get a full-pipeline system in contrast to CKAN as stand-alone. The aim of Mozgova and Sheveleva’s RDM system is to “support project workflows and move from fragmented project-oriented to a common data management system” [29].

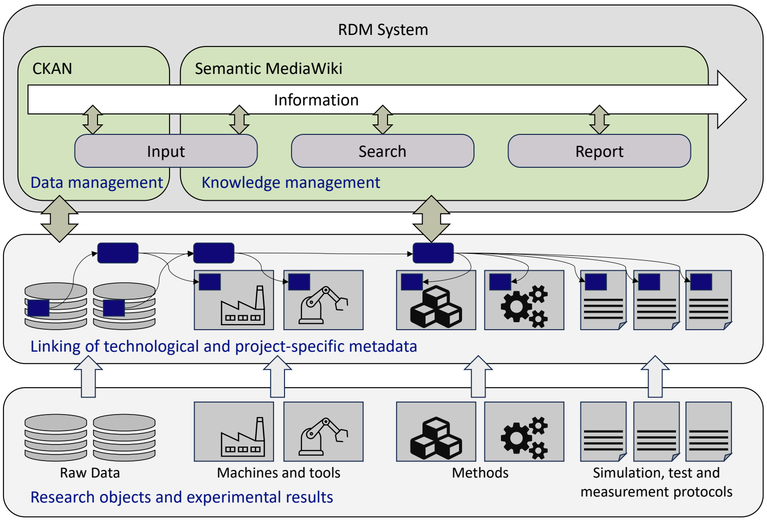

Figure 5: The concept of the RDMS as presented by [c.f. 1] and [c.f. 30]

Figure 5 shows the three layers of the concept of the RDMS proposed by Mozgova et al. [29], [30] and Sheveleva et al. [1], [31]. The lowest level represents the research objects, which are enriched by metadata and thus elevated to the second level. Enriched data and the resulting knowledge are then stored in the RDMS [1]. Mozgova et al. [29], [30] discuss several other solutions to be utilised for the RDMS. One of which is DSpace [32], a competitor to CKAN, which has the benefit over CKAN that it supports the Resource Description Framework (RDF). Yet, CKAN is adaptable via plug-ins, it is capable of supporting RDF after all and is therefore used in the solution proposed by Mozgova et al. [30].

Semantic MediaWiki The software used by Mozgova et al. [30] to create their Knowledge Management System (KMS) is another frequently mentioned solution called Semantic MediaWiki (mentioned in nine of included records). This software allows for semantic annotation of information contained in a Wiki to address challenges like inconsistencies in Wikis, findability of information and reuse of existing content by other means than reading, e.g. automated processing [33]. However, Semantic MediaWiki has not the capacity of supporting a whole RDM workflow;

it is rather a framework for semantic organisation of information, which can be also used in RDM context in combination with other tools.

Globus The Python-based "Django Globus Portal Framework" [34], also called Globus, was developed by Chard et al. [34], [35], [36] and has been referenced by seven records found in the SLR [34], [35], [36], [37], [38], [39], [40]. The technical framework Globus aims to "address important needs of both researchers and developers of scientific services, such as authentication and authorization, data transfer and sharing, and data publication" [35]. Originally designed to be "focused on reliable, high-performance, secure data transfer" [36], it became a "cloud-hosted research data management platform" [34] in recent years. As of today, Globus offers solutions for searchable metadata, authentication, data transfer and automation [34]. Nevertheless, Globus is a technical solution rather than a conceptual solution for RDM, although less distant to research in its design than CKAN.

Canonical Workflow Framework for Research (CWFR) In contrast, the CWFR is emphasising the workflow of researchers [41], [42], [43]. After Mozgova's CKAN-based RDMS , CWFR is the second solution presented focusing on RDM. In three papers Wittenburg, Peer, Betz et al. present the development of a process, a proof-of-concept (POC) study and a reflection on the CWFR itself. CWFR focusses on "self-documenting workflow scripts to automate recurring processes" [43], ensuring FAIRness via FAIR Digital Objects (FAIR-DOs), yet "without adding additional load onto the researchers who stand to benefit most from it" [43]. These processes are composed of recurring patterns, allowing also for the combination of the CWFR and other frameworks like the Reproducible Research Publication Workflow (RRPW) [42], [43]. However, in their POC study, Betz et al. conclude that, while many steps in research are similar, they often vary in their exact execution, demanding adaptability. They also stress that adaptable workflow steps are suitable for not only one domain [41].

DataFinder Maryam Idris Bugaje has written an article and a dissertation on the topic of tailoring research data repositories towards the needs of researchers [44], [45]. Due to "better engagement with the user" [45], these publications aim to "understand the different user types and groups [and] have a thorough appreciation of the tasks that each wish to accomplish through the system" [45]. The RDMS developed, called DataFinder, focuses on the reusability of research data, by making it easier to find, access and reuse while making it easier to the creators of data to provide it in a findable way. In addition, data is linked to publications [44], [45]. DataFinder aims to achieve these objectives by storing "research datasets and research publications [...] on the same platform and [connecting them] in a mutually value-adding way" [44]. The conceptualisation beneath DataFinder along with the perspective of users as "a central component of the RDM ecosystem" [44] highlights the work of Bugaje among the found records.

HELmholtz ScIentific Project WORkflow PlaTform (HELIPORT) The first solution to consider all of the research process "starting with the proposal submission, continuing with data acquisition and concluding with final publications" [46] is HELIPORT. As Knodel et al. [46] stated, "an overarching system with the potential to combine all stages of a research experiment life-cycle is to [their] knowledge still an unsolved issue" [46]. Hence, they include the Common Workflow

Language (CWL) to integrate FAIRness on the workflow level to form a RDMS containing “a full project flow with integrated computational workflows” [46], as shown in figure 6.

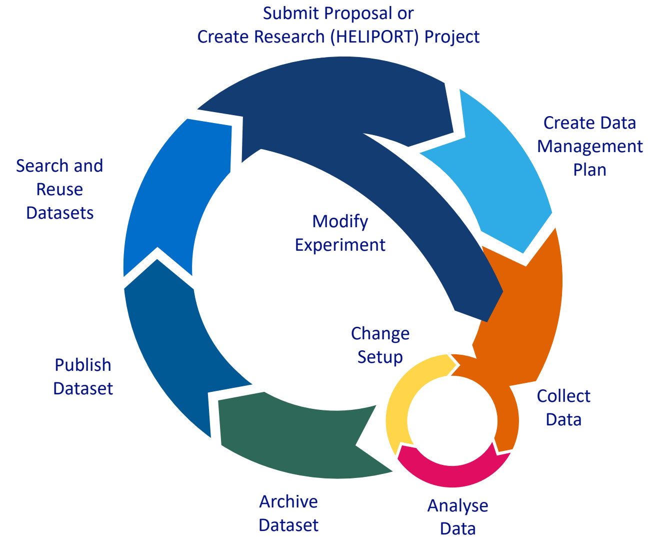

Figure 6: The life cycle of a research project as depicted by Knodel et al. [c.f. 46]

Furthermore, they design HELIPORT to be an “adaptable guidance system” [46] along with interfaces to other tools for metadata exchange [46]. The latter leads to “time savings by automating the communication between systems” [46] and allows for the “automated generation and update of a consistent project specific DMP” [46].

Open digital energy R&D platform Ferenz et al. [47] present an “open digital energy R&D platform” [47] in a similar approach as HELIPORT but for energy research. They extensively researched on the needs of researchers in the requirements analysis by Werth et al. [48] and derive a framework, which builds upon six services over the course of six phases as shown in figure 7.

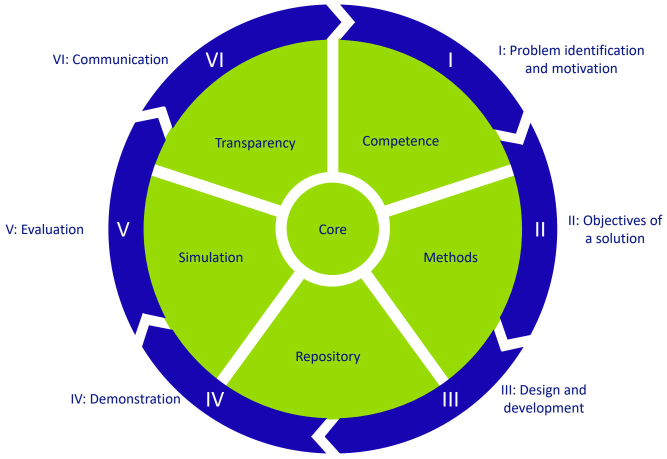

Figure 7: The open digital energy R&D platform as depicted by Ferenz et al. [c.f. 47]

The *core* of the framework contains the technical infrastructure including user management, federated search, a Persistent Identifier (PID) service and an ontology service. The *competence* service holds information on which user of the open digital energy R&D platform has which competencies in form of competence profiles [47]. “The goal of *Methods* is to provide an overview of scientific methodologies in energy research, both general and platform-specific, during the different stages of research” [47]. The *Repository* service introduces a metadata database, allowing for artefacts to be found. The fifth service is *Simulation*. It provides “an online co-simulation platform to couple different tools and models” [47]. Lastly, *Transparency* aims to “process, publish, and communicate the research and development content to promote a broader and interdisciplinary discussion among all respected types of stakeholders” [47].

Overall, the solution from Ferenz et al. [47], both as a concept as well as a tool, provides researchers with support throughout their research projects. While primarily based on energy research, it is most certainly transferable to other domains as well. Their life cycle with six phases resembles everyday research to some extent and the *Methods* service includes guidelines both for the use of the platforms services and publications.

CollAborative Environment for Scientific Analysis with Reproducibility (CAESAR) Samuel and König-Ries present the CAESAR framework, which provides an “end-to-end semantic-based provenance management platform” [49]. It aims to “support scientists to describe, preserve and visualize their experimental data by linking the datasets with the experiments along with the execution environment and images” [49]. The framework was conceptualised by conducting interviews and an exploratory study aiming at the perspective of scientific experiments and their reproducibility. On that basis, Samuel and König-Ries [49] consider both the non-computational as well as the computational steps, as “an experiment must not only be linked to its results but also to different entities, people, activities, steps, and resources” [49].

LinkAhead Hornung et al. present the RDMS LinkAhead as a “complement to Electronic lab notebookss (ELNs)” [50]. At first they derive eleven requirements for an RDMS, which they compare to existing solutions. They proceed to develop an agile solution for the processing of large amounts of automatically generated or existing data. Their “primary goal is to make searching and linking of data beneficial for its users and to allow for automation of all tasks” [50]. However, their focus is set on technical solutions rather than conceptual solutions [50].

So far, nine solutions for RDM have been presented. These have been chosen to represent the literature found both as broad as well as detailed as possible. From this overview, it can be shown, that three main categories of solutions exist. Firstly, there are meta-solutions that solve problems of RDM, such as the description of resources used, but do not directly facilitate RDM. Secondly, there are tools for RDM, that aim to reduce the effort needed to conduct RDM, but do not enable the researchers to conduct RDM in the first place. Lastly, there are conceptual frameworks that contain information on how RDM should be conducted. However, these do not offer support on a practical and applicable level.

4 Research gap

Considering the found literature, the research questions can be answered. The presented RDM solutions directly answer the first research question (see question 1), as it was shown, what solutions are available to support RDM in engineering sciences. Furthermore, three main categories of solutions have been identified. These are Conceptual Frameworks for RDM, RDM Tools and Technical Frameworks that can be utilized to support RDM but are not focussed on RDM specifically. In the following, these three categories are investigated for their capability of addressing the challenges of the researchers.

4.1 Conceptual Frameworks

HELIPORT, both viewed from a concept as well as from a tool perspective, is the solution found in the SLR, that addresses the most challenges of researchers in RDM. With its adaptable workflow management system, it guides researchers through their RDM based on the RDM-guidelines of the Helmholtz Centres. The effort for enquiring RDM requirements and conducting RDM is reduced by the defined workflow with tools integrated into it. Despite this good fit to the challenges brought up, HELIPORT is specialised on Helmholtz Centres and their specific RDM practices. While this would still allow for adaptability to non-Helmholtz organisations, the presence of existing processes for conducting RDM like the “well-defined publication process” [46] hinders its adaptability to the reality of not only one but several different policies forming the RDM of a specific project. Therefore, HELIPORT supports researchers in their everyday work at Helmholtz Centres and its ecosystem of RDM tools but lacks applicability beyond this closed community. Furthermore, while the focus of HELIPORT is on the researchers perspective, it emphasises RDM processes, aiming to facilitate the integration of the RDMS rather than adapting the RDMS to the research process. In summary, HELIPORT offers no management of other policies than their own, no decision support and no training materials.

The approach of Mozgova et al. [29], [30] and Sheveleva et al. [1], [31] also has a conceptual scientific approach. In their case however, the RDMS is focused on the procedure of enriching data with metadata and knowledge management, as shown in figure 5. Although their methodology considers the needs of researchers, specifically “the researchers’ needs for the protocol documentation” [1], the resulting concept does not provide solutions to the challenges identified by Hamann et al. [17]. Namely, it provides neither help for guidelines nor information on how to conduct RDM. The management of knowledge has the potential of reducing the effort of RDM. However, as there is a lack of guidance, this effort can also potentially increase due to uncertainty and complexity amongst researchers when utilising their concept.

The same applies to Globus as it is presented by Saint et al. [34]. The platform offers broad functionality from a data management perspective rather than the researchers point of view. Although it allows for adaptability and can be utilised by a predefined software development kit, researchers need to adapt their research to this solution.

While emphasising more on FAIR Digital Objects and the benefits of the researchers and adaptability , CWFR does not aim to facilitate RDM for researchers. Instead, its developers focus on offering a way of describing steps conduced in research and their FAIR outputs. The CWFR is a good example for the facilitation of the documentation and FAIRification of research data. However, researchers have to adapt to this approach rather than the solution adapting to the researchers needs, rendering it a hurdle once more.

Similarly, the DataFinder approach by Bugaje, while being user centred, only considers how researchers interact with the RDMS. The approach does not consider how the RDMS can be easily integrated into the day-to-day life of researchers. Additionally, it focuses solely on data FAIRness in the context of repositories, which leaves many aspects of RDM unconsidered.

4.2 Tools

The Open digital energy R&D platform follows the Design Science Research (DSR) approach by Hevner at al. and thoroughly gathers the researchers needs and demands [47], [48]. However, these requirements are strongly focused on the technical aspects required for the open digital energy R&D platform as a tool, rather than the underlying concept of support. Hence, the open digital energy R&D platform is a specialised tool for RDM, rather than a solution that allows for tools to be created from it. In that regard, the Open digital energy R&D platform forms a transition between concepts and tools.

In a similar way, CAESAR serves as an example of a well conceptualised and specialised tool. It aims to facilitate RDM for a specific scenario (microscopy research projects) in a specific community (computational biology) for a specific range of RDM steps (capture and management of scientific experiments) [49]. As such, it may offer valuable insights in the specifics of RDM in its scenario, yet does not guide researchers through their RDM.

LinkAhead is another example for a tool that aims to solve a specific problem, namely the automated processing, searching and linking of large amounts of data. While it utilises many technical frameworks like RDF, it is not a conceptual solution to a general problem but addresses the automated handling of data in the context of RDM specifically.

4.3 Technical Frameworks

Due to their adaptability, technical frameworks, especially CKAN and Semantic MediaWiki, are the most cited ones. CKAN provides a broad modular range of functionalities but is a technical framework rather than a solution. It enables other concepts, like the one by Mozgova et al. [29], [30] and Sheveleva et al. [1], [31], but does not provide RDM support on its own.

The Semantic MediaWiki approach is driven by methods collecting the researchers' needs. Yet, the concept itself revolves around the data and its enrichment, rather than the process in which the researchers collect said data. For a specific use case, the RDMS considers “manufacturing processes within the Tailored Forming Technology” [1], but in practice it does not adapt to the process, but rather offers specific places for this information to be stored (see “Machines and tools” and “Methods” in figure 5) [1].

4.4 Summary of the research gap

In the SLR it was found that Conceptual Frameworks generally provide the most advanced scientific approaches. Hence, these are to be considered the most important solutions for further research. Most of them also transfer their theoretical approach to technical solutions i.e. tools, just as we intend to. Another category are Tools for RDM. These aim at addressing a specific problem, solving it by providing a single software solution. Some records featuring tools include a requirements analysis, survey, interviews or workshops as scientific method. Lastly, Technical

Frameworks are focused on overarching technical solutions for recurring problems within and beyond RDM. While they do not provide any solutions immediately applicable to the researchers RDM problems, they are to be considered when creating a solution for RDM. It has to be mentioned that the prior division into Conceptual Frameworks, Tools and Technical Frameworks has to be seen as a spectrum rather than fixed categories. A good example for this is HELIPORT, as it is a scientifically based conceptual solution that was also implemented as a tool.

As shown, the landscape of existing solutions addressing parts of the identified problems is broad. Still, a comprehensive solution for all of the researcher's challenges could not be identified in the literature. Revising the RDM challenges of researchers in the engineering sciences as presented in Figure 1, Conceptual Frameworks, Tools and Technical Frameworks can be placed in the diagram as shown in figure 8.

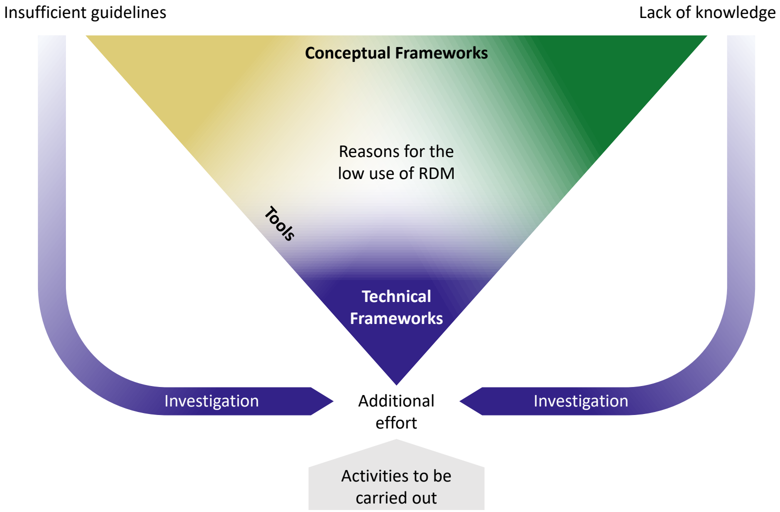

Figure 8: Reasons for the low use of RDM by researchers in the engineering sciences

As illustrated, the existing solutions address the researchers' challenges only partially. Hence, research question 2 can be answered. There are solutions, that partially address the challenges of researchers in engineering sciences in the context of RDM. However, in the found literature no solution could be identified that addresses all named challenges of the researchers.

This leads to the third and last research question (see research question 3), asking for the gap in the current literature. The gap can be summarised as missing guidance - referring both to enquiries needed to conduct RDM and lack of RDM knowledge amongst researchers - with focus on the researcher's perspective on RDM. This missing guidance is partially caused by the neglecting of the researcher's perspective on RDM in the existing solutions: researchers have to adapt their work to these solutions rather than the solutions adapting to the researcher's work.

This perspective was considered by the survey conducted by Hamann et al. [17]. Based on this survey, five characteristics of the needed support for researchers in RDM in the engineering sciences can be derived. These characteristics are listed below as they are building upon each-other, forming a concept in which they empower each-other.

1. Guidance throughout the researchers' RDM process
2. Management of external factors

3. Support on RDM-specific decisions
4. Provision of knowledge
5. Partial automation as work facilitation for researchers

A solution that encompasses these characteristics is currently missing from the found literature. Partial automation is to some extend addressed, especially by HELIPORT and some of the other Conceptual Frameworks that evolved into tools. However, these lack guidance and are either focused on a specific scenario or community, or lack alignment with the researchers work processes. Hence, we present a novel solution for RDM in engineering sciences, focused on the needs of researchers. A visualisation of the concept and its characteristics building upon each-other is shown in figure 9.

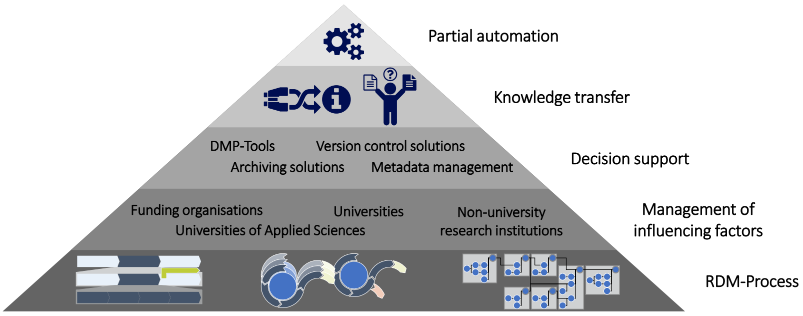

Figure 9: The layers of the proposed concept as presented in [51]

5 Proposed toolchain

As described before, on the spectrum from conceptual frameworks to technical frameworks, there are scientifically based conceptual solutions that are also implemented as a tool. The same applies for the newly proposed toolchain, as it is both a conceptual solution as well as a combination of tools. As a result, the new solution is not particularly more focused on the researchers needs as other solutions, neither is it capable of reducing the effort significantly more than specialised solutions for particular cases. In its research methodology, it also does not contrast other scientifically based solutions in its quality of research.

However, the newly proposed toolchain stands out in two ways. Firstly and most importantly, it provides **guidance** in yet unconsidered ways. This encompasses the overall guided process it is build upon, the management of external factors, the decision support and the provision of specific training materials based on decisions taken. Secondly, it combines reduction of effort needed and applicability in a new way. Researchers do no longer need to investigate their respective RDM guidelines but get the advice on the next activities right away.

5.1 Interconnected Toolchain

Additionally, the new solution combines existing tools with each other by using a new tool for the overarching process. This greatly improves the capabilities of the toolchain, as the individual solutions provide specialised solutions for e.g. DMPs or data enrichment and storage. In that way, the new toolchain includes both a guided and adaptable process. This process provides a step-by-step approach with decision-support on which solutions to used based on the external factors applying and training materials for the solution chosen.

As of now, the toolchain includes three individually conceptualised and implemented tools. The guided RDM process, the management of external factors and the decision-support are provided by Jarves. Jarves connects to Research Data Management Organiser (RDMO) and Coscine. RDMO provides support for DMPs and was recently enriched by checklists to give researchers an overview over tasks to conduct when filling a DMP. Coscine allows storage, organization and linking of data and can handle data and its metadata combined. In this way, data can directly be enriched with standard or custom information based on metadata schemes.

5.2 Jarves

While RDMO and Coscine offer solutions for specific RDM tasks, general questions arise when researchers start out with their RDM. The current lack of specific guidance or orientation [15], [19], [20] combined with missing RDM knowledge [19], [52], [53], [54] complicates the initialisation of RDM. The setup and conduction of RDM encompasses not only the clarification of tools to be used and tasks to be performed. Beforehand, rules and guidelines have to be consulted, often originating from different sources as for example funding organisations, universities or other institutions and partners. These rules and guidelines will then affect the rest of the RDM. Yet, these rules may interfere with or even contradict each other.

In addition, while some guidelines contain specifications for RDM-tools [55], others do not [56]. This leads to researchers having to first familiarise with the specific task to be conducted, e.g. the creation of a DMP, and then having to figure out suitable tools for the task. When a tool has been chosen, researchers must become acquainted with the tool, adding more effort to the process of RDM initialisation. Furthermore, researchers do not only see the effort of initialising RDM and familiarisation as a burden, but also perceive day-to-day work with RDM as a hindrance [17]. More tools to get used to and to synchronise with each other cause researchers to see RDM as “a significant burden” [2]. As “the amount of time [RDM] takes” [19] increases with the tasks involved, the “perceived increased workload”[16] directly opposes the researchers “lack of resources (time, budget, personnel etc.)” [15].

The Joint Assistant for Research in Versatile Engineering Sciences, in short Jarves, addresses these issues. By providing **guidance** in RDM throughout a research project, Jarves structures the RDM tasks along the day-to-day research. While doing so, it takes into account the environment of the specific research project. These include but are not limited to the demands of the funding organisation, the universities guidelines and the research’s discipline. Jarves **manages these guidelines** specifically for each research project. It derives RDM requirements from these and **helps researchers to decide** which RDM step to perform next and which tools to choose for the task. If needed, researchers are supported by **training materials** for their current task. By providing **partial automation**, Jarves reduces unnecessary effort of synchronising tools with each other. These five pillars are the foundation of Jarves, which are building upon each other.

Guidance in RDM Firstly, Jarves provides a clear entry point for researchers into RDM. When creating a project, Jarves queries basic information from researchers like their university, the project's current funding organisation and discipline(s) of the project. They are then guided throughout their setup of RDM by Jarves. This setup contains for example the documentation of the research idea, the assignment and documentation of roles within the research project or the decision for a DMP-tool. Afterwards, the creation, documentation and storage of data is supported along the researchers' everyday work. Finally, Jarves accompanies researchers in their project closing by, for example, making sure the right data is selected to be archived.

Management of rules and guidelines As Jarves collects the basic information from researchers about their project, the rules and guidelines, e.g. by the funding's demands, applying are derived. Jarves allows for the integration of policies on different levels, may it be by funding organisations, universities or individual institutes. In addition, each of the levels can have multiple entries, assuring the integration of all applying conditions to the project.

Decision Support for RDM Not only does Jarves provide a basis for the management of rules and guidelines, but also integrates them into the decisions to be made during the research activity and the accompanying RDM process. In that way, researchers are provided with recommendations on how to perform their RDM along with the reasoning behind them. The information on what influences their RDM on every step aims to support researchers in a comprehensible way.

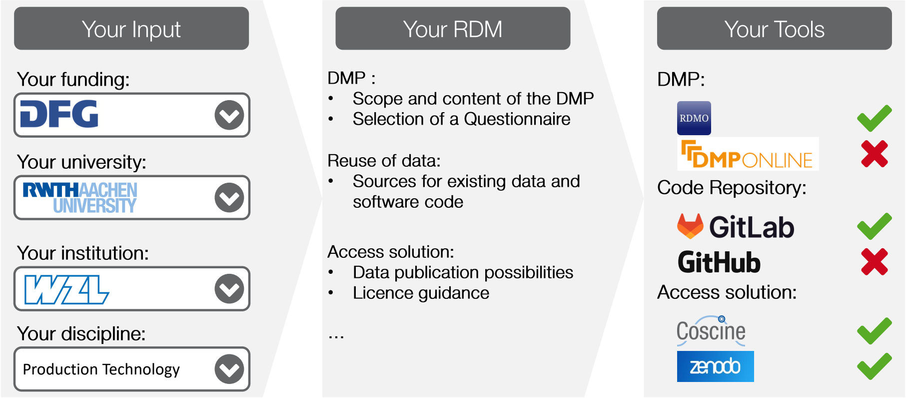

Figure 10: The Workflow integrated in Jarves: Structure with guidance and decision support for RDM solutions to be used [c.f. 23]

Figure 10 depicts the way in which the boundary conditions of a research activity are translated into RDM requirements. However, Jarves does not provide all of the information at once but weaves it into its guided process. In that way, researchers are not overwhelmed with details but receive just the needed piece of information to continue their research along with their RDM.

RDM-trainings at the point of need It might occur that researchers are uncertain what the current RDM task is about, why it is important, how to perform it and what its goals are. Jarves integrates such trainings the same way recommendations are integrated. Trainings on specific topics appear on the steps in which they become relevant. This way not only the recommendations but also further information are bundled into each of their corresponding steps.

Partial automation Lastly, as the effort of RDM not only applies for the initialisation of it but also the day-to-day work, Jarves provides partial automation to further reduce the burden of RDM. Utilising a REST Application Programming Interface (API), Jarves provides an interoperability with other RDM tools, for example RDMO and Coscine.

5.3 Research Data Management Organiser (RDMO)

RDM comprises several tasks, including data organisation and documentation, long-term preservation (archiving or publication), resource planning, legal aspects. How to perform these tasks is partly a decision of the involved actors (work group members or project partners), who must satisfy several demands from their institutions, from their disciplinary communities and from the funding organisations. All these tasks and decisions are usually reported in a DMP, which is mostly prepared at the beginning of a project activity and updated when results are being produced. The preparation and curation of a DMP is highly encouraged, and some funding organisations make it compulsory, e.g. the German Research Foundation (DFG) [6].

The DMP is usually required as a text document, where a given set of questions is answered with free text. This implementation of a DMP has some drawbacks, which enormously reduce its usefulness. Basically, answers embedded in the text flow fulfil only a documentation purpose, but they cannot be used for other applications, e.g. searching, filtering, data annotation or publication, or resource planning. That way, no connection is provided between what's declared in the plan and the concrete data-related tasks: the users are often forced to re-ingest the same information elsewhere during the data life cycle, confirming the first impression, that preparation of a DMP offers little practical value beyond compliance.

The situation may be improved by using a DMP platform. DMP platforms allow filling a DMP via a web form and saving its content as a structured data object. Compared to a free-text DMP, this approach has several advantages. For instance, all information is stored in a structured, machine-readable form and is available for further processing by the project coordination or an infrastructure provider. A "document-like DMP" with a given layout can always be exported from the platform.

There are several such platforms with different strengths, localisation and focus, i.e., optimised for specific countries or funding programs. Institutions have the possibility to install an "instance" of a DMP platform for their own, or to share one within one scientific cooperation/consortium, or to use a central service offered by an academic service provider with a regional/national/disciplinary coverage. Some of them, notably RDMO, allow a cooperative writing of a DMP.

Among other DMP platforms, RDMO is peculiar because it brings machine interoperability to the next level. The information stored in a DMP can be fetched from other applications via an API and thus be used to trigger or provide parameters to a task in the data life cycle. It is also possible to set reminders or alerts (tasks) within the software itself.

Customisable Modules RDMO's data model contains different modules (see Figure 11), all customisable by an instance manager for different use cases. What the end user sees are questions contained in a question catalogue. They are usually more fine-grained than the funding organisation's original questions, and answering them is facilitated thanks to help texts and controlled

Table 1: A selection of DMP tools and their specifications

Tool	Hosting	Focus
ARGOS	local	Funding: EU (H2020/OpenAIRE)
DataWiz	central	Discipline: Psychology
DMPonline	central	Funding: British Research Councils
DMPRoadmap	central	Funding : USA, British Research Councils
DMPTool	central	Funding: USA
DSW	local or central	General
GFBIO DMP	central	Funding: DFG & Discipline: Biodiversity
RDMO	local	Different combinations of Funding, Discipline, Geographical area, Infrastructure
TUB-DMP	local	Own repository DepositOnce

answer lists (option sets), which can be provided either locally or from external authoritative sources via dedicated import plug-ins. The answers are saved internally as values of internal variables (attributes), which can be then downloaded, or retrieved via the API, or exported in form of export templates (views).

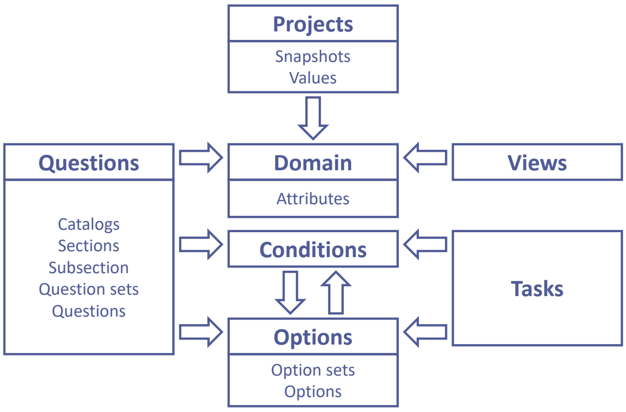

Figure 11: RDMO's data model

Broad and Open Community The RDMO software is open source and based on Python and Django. There is a strong community supporting the curation and further development of the software on one side, and of the content (catalogues, attributes, option sets, views) on the other side. RDMO has reached the stage of a mature software and has been elected as a favourite DMP platform within the NFDI by the most consortia, as well as a central DMP service (DMP4NFDI) [57].

5.4 Coscine

As stressed out above, researchers (especially from the engineering sciences) have to deal with huge amounts of data from various data collection methods. In addition to the challenge of obtaining enough storage space for the resulting data volumes, the collection methods often result in various project-related data types (e.g., managing code, collaborative work, multiple large files). The different file types and the required storage space often force researchers to distribute the data to different service providers with varying (if any) levels of maturity in terms of FAIR [5] RDM. Thus, a software solution is needed to provide access to storage space for different kind of research data under one roof by adding a FAIR [5] layer to already established services. To create such a data management system, the RDM platform Coscine was developed at the IT Center of the RWTH Aachen University [58].

Coscine enables researchers to integrate different service providers, gain access to storage space, link metadata with project data, and work collaboratively with defined roles. All researchers can use the platform via the ORCID login and, depending on their organizational affiliation, storage financed by the MKW (Research Data Storage and DataStorage.nrw) if offered. In addition, data can be best described using appropriate metadata based on individually created metadata profiles. This helps researchers comply with the FAIR principles. The data is stored in a uniform location, access rights are regulated, and reusability is enabled by archiving the data for 10 years after the end of the project. For identification of data, PIDs are assigned at both project and resource level and work is currently underway to integrate FAIR-DO. As part of the Coscine.nrw project, Coscine is available to all DH.NRW universities as a software as a service model.

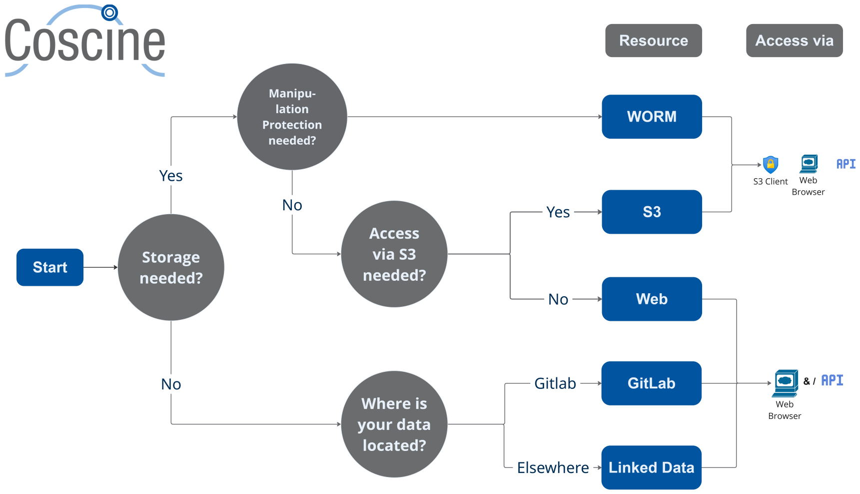

Figure 12: Coscine Workflow: 5 main Steps from Login to Archiving.

Description of the Workflow Figure 12 shows the main workflow for researchers using Coscine. Various login options are available to enable researchers a low-threshold login. In addition to ORCID, researchers can also select between two Authentication and Authorization Infrastructure (AAI) login options (SSO and NFDI Login) provided by the German National Research and Education Network (DFN) and link their accounts in the user profile. After login, researchers can create their projects in Coscine using the web interface or via the Coscine API, also utilised by Jarves. Created projects and sub-projects are mapped in Coscine as a tree structure, enabling simple and hierarchical visualization. Project owners can invite project members easily via e-mail and regulate their corresponding access via the project-related rights and roles concept. Based on the low-threshold login, external project partners can be invited to Coscine project, fostering collaborative co-operation across institutes and universities. Research data is managed in Coscine at the resource level, where different resource types (see Figure 13) are available in each project. This availability is depending on access rights and the selected login method of the respective project owner. Inside the resources, research data can be uploaded, described with metadata and shared with other project members. To share data outside Coscine, project owners can create download links. If data and metadata from Coscine should be published, project owners can send a request with the project and resource metadata collected in Coscine to an external publication service for further support. Coscine also offers a ten-year archiving period for the files after the end of the project. During this period, the uploaded files are in read-only mode and can no longer be changed.

Resource Types GitLab and LinkedData resources are available to all Coscine users. Storage space resources are available based on the default access rights of the respective home organization of the project creator and/or based on successful application for storage space via the JARDS Application Platform [59]. The available storage space resource types are web resources (accessible only via the browser/API), S3 resources (accessible via browser/API/S3) and WORM (write once, read many) resources (accessible via browser/API/S3).

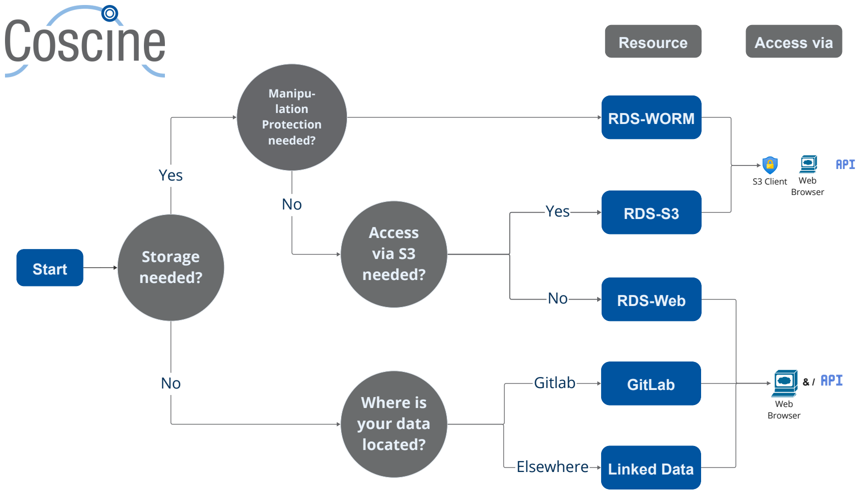

Figure 13: Decision tree for available resources in Coscine.

Metadata and FAIR Digital Objects Metadata plays a central role in Coscine and is mandatory at multiple levels. When initiating a project or creating a resource, users are required to provide essential generic metadata, including the title, brief descriptions, creation date, and principal investigator. Discipline-specific metadata is captured during the upload of files or folders to a resource through either the browser or API interface. This process involves completing so-called metadata profiles, which can be created via the Applying interoperable metadata standards (AIMS) platform and reused by other Coscine users. To enhance reusability, all entered metadata is searchable and can be made accessible to other Coscine users.

Coscine embraces the concept of FAIR-DO as a vital means to align digital objects with the FAIR principles [60], [61]. An FAIR-DO is generally defined as a machine-readable and machine-actionable unit of data that is uniquely identified by a PID and described through a comprehensive record. The Coscine approach ensures that the FAIR-DO record includes essential information regarding the object's context and the potential operations associated with it. Coscine prioritize providing a minimal set of metadata within this context, enabling effective discovery and reusability of the data. The FAIR-DOs created by Coscine support researchers in maximizing the value and accessibility of their digital objects [62].

6 Summary and Outlook

In this article, the three main challenges of researchers in the context of RDM are summarised. Firstly, sufficient guidelines in RDM are missing. Secondly, researchers lack knowledge on how to conduct RDM. Thirdly, additional effort is caused both by the application of RDM itself as well as by the enquiries caused by insufficient guidelines and lacking knowledge.

These problems were used as motivation to search for existing solutions. In a systematic literature review, a total of 2.409 records were screened of which 88 were assessed as relevant. This answers research question 1, asking for the RDM support available. While many records gather requirements from researchers, they still provide RDM solutions to be adapted by the researcher rather than solutions that adapt to the researcher's work. This results in solutions that lack needed guidance throughout the RDM process. Hence, the existing RDM solutions are not addressing the challenges of the researcher as asked for in research question 2. The research gap, that there is currently no RDM solution that focuses on the challenges and needs of researchers from their perspective, could be identified. This gap leads to the answer of research question 3, as the gap shows what current RDM solutions are missing to fully address the researchers' challenges. The results of the SLR could not falsify our hypothesis, that such a solution does not exist yet.

To address this gap, we propose a novel toolchain, which combines existing tools with a five-layered approach on supporting researchers, as shown in figure 9. Within a guided RDM process, researchers can conduct their RDM along their everyday work processes. Within this process, the influencing factors on their RDM are managed for them. Based on these factors, RDM decisions are supported, e.g. which tools and solutions to choose for specific RDM tasks. In addition, researchers are supported with training materials if and when needed. Lastly, partial automation supports researchers as it reduces effort.

In this paper, the automation in this toolchain is emphasised by a combination of three tools, which exchange data with each other. Jarves provides the overall process with the decision support and directs researchers to RDMO and Coscine for their DMPs and data annotation and storage respectively. Thus, a toolchain is formed, which encompasses a single entry point into RDM, general support tailored towards the needs of researchers and specialised RDM tools. In future work, this toolchain has to be validated and expanded.

A Appendix

In the following, the procedure of the literature is thoroughly explained.

Inclusion and exclusion criteria for the review [25] At first, our literature search selected possibly relevant articles using the inclusion criteria listed in table 2. For the inclusion criteria, search terms were gathered, shown in table 3. Only English and German literature was considered, starting from the publication year (2016) of the FAIR Principles [5]. Secondly, the list was reduced in size by the application of the exclusion criteria listed in table 4, allowing exceptions for ”particularly relevant” contributions fitting to the topic in other regards.

Furthermore, only English and German literature was included.

Criteria-No.	Criteria
1.	The publication contains any combination of the defined search terms.
2.	The publication was published in the same year (2016) or after the FAIR principles of Wilkinson et al. [5] was published.
3.	The language of publication is German or English.

Table 2: Inclusion criteria for the literature review

ENG	AND	AND	AND
	Framework	Research data management	Engineering
OR	Groundwork	RDM [2]	STEM
OR	Structure	Research data administration	
OR	System	Scientific data management	
GER			
	Rahmenwerk	Forschungsdatenmanagement	Ingenieur*[4]
OR	Grundgerüst	~~FDM~~[3]	MINT
OR	Struktur	Forschungsdatenverwaltung	
OR	System	Wissenschaftliches Datenmanagement	

Table 3: Search strings used for the inclusion criteria for the literature review. Languages were queried independently from one another

However, if one or more of the exclusion criteria (see table 4) were met it led to an immediate or subsequent exclusion. Still, exceptions were allowed if a publication contained an important contribution fitting to the topic in other regards.

Sources searched or consulted to identify studies [25] The platforms used for the literature review are listed in table 5 along with the dates of the search and the number of results. The results of the search engines were filtered as far as possible by date, language, topics and coverage, to intercept only relevant literature (see column “Filters Used” in table 5).

It has to be mentioned that the results of ProQuest have been limited to the top 20% of the most relevant results. This is due to the fact, that after 300 screened articles, the review showed less than 5% of the articles being relevant. Therefore the results of ProQuest were filtered by

---

2. Not included in the Search in Science Direct as it only supports nine boolean operators.
3. Because the German abbreviation for RDM “FDM” is commonly used for Fused Deposition Modeling in English, causing it to be excluded from the search.
4. Not included in the Search in Science Direct as it does not include wildcards. Instead, “Ingenieurwissenschaften” was used.
5. Already included in the English search results.

Criteria No.	Criteria	Records excluded in abstract review	Records excluded in full-text review
4.	The publication does not fit to engineering, MINT or STEM subjects.	From search engines: 487 From other sources: 8	From search engines: 4 From other sources: 1
5.	The publication does not contain any information on RDM.	From search engines: 1236 From other sources: 12	From search engines: 3 From other sources: 0
6.	The publication does not contain any information on frameworks.	From search engines: 269 From other sources: 35	From search engines: 17 From other sources: 13
	Sum from search engines:	1.992	24
	Sum from other sources:	55	14
	Total sums:	2047	38
	Total sum:		2085

Table 4: Exclusion criteria for the literature review with number of records excluded both for the search engines (see table 5) an other sources (see table 6)

Search Engine	Last Searched	Filters Used	Results EN	Results GER
ACM Digital Library	21.02.2024	Publication date: 2016 - 2024	749	0
ProQuest	06.05.2024	Publication date: 2016-01-01 - 2025-01-01 Language: EN or GER Filtered by topics Results: In subscription of RWTH Aachen.	574	1
Engineering Village	21.02.2024	Year: 2016 - 2024	430	0
Scopus	06.05.2024	Year: 2016 - 2024 Subject area: Engineering Filtered by keywords Language: EN or GER	315	1
Web of Science	20.02.2024	Year: 2016 - 2024 Research Areas: Engineering Language: EN or GER	164	0
ScienceDirect	21.02.2024	Year: 2016 - 2024 Subject area: Engineering	40	0
IEEE Xplore	20.02.2024	Year: 2016 - 2024	29	0
RWTH Publications	21.02.2024	Year: 2016 - 2024	4	0
		Sum:	2.305	2
		Sum without duplicates:	2.102	2
		Total sum without duplicates:		2.104

Table 5: Used search engines, filters and results for the literature review

Search Engine	Last Searched	Filters Used	Results EN	Results GER
Ing.grid	02.01.2025	Manuscript	9	0
Ing.grid Preprints	02.01.2025	-	23	0
BausteineFDM	02.01.2025	-	4	29
Zenodo	02.01.2025	Year: 2016 or newer Subjects: Engineering	21	16[5]
		Sum:	57	48
		Sum without duplicates:	45	27
		Total sum without duplicates:		75

Table 6: Used search engines, filters and results for the literature review

relevant topics, namely *research*, *decision making*, *collaboration*, *research data management*, *data management*, *reproducibility* and *information sharing*. Every entry with the topic *Datasets* was excluded from the search. This reduces the initial 3.034 results of ProQuest to 574. As this is about 19% of the initial results, there is a safety factor to the initial screening of 3,8 (19% of results to 5% relevant results).

The same procedure was applied to Scopus. Here, the results were filtered by relevant keywords, namely *data curation*, *data management*, *data quality*, *database systems*, *decision making*, *decision support systems*, *FAIR*, *FAIR data*, *FAIR principles*, *information management*, *knowledge based systems*, *knowledge management*, *research data*, *research data management*, *research data managements*, *reproducibility* and *scientific data*. Every entry with the keyword *data set* was excluded from the search. This reduces the initial 1.238 results to 315. As this is about 25% of the initial results, there is a safety factor to the initial screening of 5 (25% of results to 5% relevant results). The same problem again occurs for the ACM Digital Library, however there is no option to filter the results further. As a result, all 749 results were be screened.

In addition to typical search engines, other sources have been consulted as well. As the field of RDM is a dynamic one, specific journals for RDM have been consulted. Articles in ing.grid both published and on the preprint server have been scouted. Furthermore, with BausteineFDM another journal for RDM has been consulted. BausteineFDM is active since 2018 and is a German journal for RDM . As ing.grid and BausteineFDM are active since 2023 and 2018, no filters for publication date needed to be applied for the search [63], [64].

Selection and data collection process [25] Afterwards, the resulting papers were exported in the .ris format along with their abstracts into the PICO-Portal[8]. There, a independent double abstract screening took place. Selected publications were screened by full text by both reviewers independently. Adjudications were decided by the lead reviewer and when in doubt, included in the review rather than excluded.

---

[8] https://picoportal.net/

B 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 [14] for NFDI4ING and project number 432233186 [65] for AIMS.

C Roles and contributions

Tobias Hamann: Conceptualization, Methodology, Writing

Katja Jansen: Conceptualization, Methodology, Writing

Giacomo Lanza: Conceptualization, Writing

Pia Carina Pickmann: Conceptualization, Methodology, Writing

Ilona Lang: Conceptualization, Methodology, Writing

Marcos Alexandre Galdino: Writing - Review

Anas Abdelrazeq: Writing - Review

Robert H. Schmitt: Writing - Review, Supervision, Funding acquisition

References

[1] T. Sheveleva et al., "Creation of a knowledge space by semantically linking data repository and knowledge management system - a use case from production engineering," *24058963*, vol. 55, no. 1, pp. 2030–2035, 2022, Conference proceeding, ISSN: 24058963. DOI: 10.1016/j.ifacol.2022.10.006

[2] T. Austin, K. Bei, T. Efthymiadis, and E. Koumoulos, "Lessons learnt from engineering science projects participating in the horizon 2020 open research data pilot," *Data*, vol. 6, no. 9, p. 96, 2021, PII: data6090096. DOI: 10.3390/data6090096

[3] J.-M. Rodriguez, "Aktuelle softwareplattformen für forschungsdatenrepositorien auf dem prüfstand," *ABI Technik*, vol. 39, no. 1, pp. 23–25, 2019, ISSN: 0720-6763. DOI: 10.1515/abitech-2019-1004

[4] K. Gregory, S. J. Khalsa, W. K. Michener, F. E. Psomopoulos, A. de Waard, and M. Wu, "Eleven quick tips for finding research data," *PLoS computational biology*, vol. 14, no. 4, 2018. DOI: 10.1371/journal.pcbi.1006038

[5] M. D. Wilkinson et al., "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160 018, 2016, Journal Article, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18

[6] DFG. “Umgang mit forschungsdaten, Checkliste für antragstellende zur planung und zur beschreibung des umgangs mit forschungsdaten in forschungsvorhaben.” version 21.12.2021, Accessed: Jul. 28, 2023. [Online]. Available: `https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_de.pdf`

[7] Bundesministerium für Bildung und Forschung. “BMBF-Aktionsplan Forschungsdaten, Impulse für eine Kultur der Datenbereitstellung und Weiterverwendung in Bildung, Wissenschaft und Forschung,” Accessed: Jun. 2, 2023. [Online]. Available: `https://www.bildung-forschung.digital/digitalezukunft/shareddocs/Downloads/files/163_20_faktenblatt_aktionsplan_3_final.pdf`

[8] K. L. Wilms, S. Stieglitz, A. Buchholz, R. Vogl, and D. Rudolph, “Do researchers dream of research data management?,” 2018, Conference proceeding, ISSN: 15301605. DOI: `10.24251/HICSS.2018.556`

[9] D. Williams and H. Tang, *Data quality management for industry 4.0: A survey*. 2020. Accessed: Mar. 15, 2023. [Online]. Available: `https://search.proquest.com/openview/f7aaf15f64ccb032852e4edbd1ad16c9/`

[10] Y. Roh, G. Heo, and S. E. Whang, “A survey on data collection for machine learning: A big data - ai integration perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2021, ISSN: 1041-4347. DOI: `10.1109/TKDE.2019.2946162`

[11] N. Hartl, E. Wössner, and Y. Sure-Vetter, “Nationale forschungsdateninfrastruktur (nfdi),” *Informatik Spektrum*, vol. 44, no. 5, pp. 370–373, 2021, PII: 1392, ISSN: 0170-6012. DOI: `10.1007/s00287-021-01392-6`

[12] R. H. Schmitt et al., *Nfdi4ing - the national research data infrastructure for engineering sciences*, 2020. DOI: `10.5281/ZENODO.4015201`

[13] L. Laaksonen, “Research Data Alliance - Research Data Sharing without barriers: Big Data & Open Data Workshop 2014,” vol. 2014, 2014. [Online]. Available: `https://www.elettra.eu/Conferences/2014/BDOD/uploads/Main/Research%20Data%20Alliance.pdf`

[14] DFG. “Dfg - gepris - nfdi4ing – national research data infrastructure for engineering services,” Accessed: Nov. 29, 2022. [Online]. Available: `https://gepris.dfg.de/gepris/projekt/442146713`

[15] L. Costanzo and A. Cooper, “Setting the foundations for stronger partnerships and collaborations for developing institutional rdm strategies in canada,” DOI: `10.5281/ZENODO.8010869`

[16] K. L. Wilms, S. Stieglitz, B. Ross, and C. Meske, “A value-based perspective on supporting and hindering factors for research data management,” *International Journal of Information Management*, vol. 54, 2020, PII: S0268401219305006, ISSN: 02684012. DOI: `10.1016/j.ijinfomgt.2020.102174`

[17] T. Hamann et al., “A survey on the dissemination and usage of research data management and related tools in german engineering sciences,” Ing.grid Volume 2 Issue 1 2024, *ing.grid*, 2024, ISSN: 2941-1300. DOI: `10.48694/inggrid.4073` Accessed: Mar. 9, 2024.

[18] T. Hamann, A. I. Metzmacher, and P. Mund, *Engineering rdm survey - dataset*, 2023. DOI: 10.5281/ZENODO.7645547

[19] J. A. Borghi and A. E. van Gulick, "Data management and sharing in neuroimaging: Practices and perceptions of mri researchers," *PloS one*, vol. 13, no. 7, 2018. DOI: 10.1371/journal.pone.0200562

[20] M. Björnmalm et al., *Advancing research data management in universities of science and technology*, 2020. DOI: 10.5281/ZENODO.3665372

[21] A. Wuchner, M. Robrecht, P. Kehl, and H. R. Schmitt, "Challenges in publishing research data – a fraunhofer case study," Ing.grid PrePrint, 2023. Accessed: Oct. 31, 2023. [Online]. Available: https://preprints.inggrid.org/repository/view/20/

[22] A. Palsdottir, "Data literacy and management of research data – a prerequisite for the sharing of research data," *Aslib Journal of Information Management*, vol. 73, no. 2, pp. 322–341, 2021, ISSN: 2050-3806. DOI: 10.1108/AJIM-04-2020-0110

[23] T. Hamann and J. Werheid, *Jarves: The Digital Data Steward for Engineering Science Research*, 2023. DOI: 10.5281/ZENODO.8378834

[24] M. J. Page et al., "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *BMJ (Clinical research ed.)*, vol. 372, n71, 2021. DOI: 10.1136/bmj.n71

[25] M. J. Page et al., "Prisma 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," *BMJ (Clinical research ed.)*, vol. 372, n160, 2021. DOI: 10.1136/bmj.n160

[26] V. Braun and V. Clarke, "Conceptual and design thinking for thematic analysis," *Qualitative Psychology*, vol. 9, no. 1, pp. 3–26, 2022, ISSN: 2326-3601. DOI: 10.1037/qup0000196

[27] Open Knowledge Foundation. "User guide — ckan 2.11.1 documentation," Accessed: Dec. 20, 2024. [Online]. Available: https://docs.ckan.org/en/2.11/user-guide.html#what-is-ckan

[28] ckanorg. "Showcase," Accessed: Dec. 20, 2024. [Online]. Available: https://ckan.org/showcase

[29] I. Mozgova, O. Koepler, A. Kraft, R. Lachmayer, and S. Auer, "Research data management system for a large collaborative project," 2020, Conference proceeding. DOI: 10.35199/NORDDESIGN2020.48

[30] I. Mozgova et al., "Knowledge annotation within research data management system for oxygen-free production technologies," *Proceedings of the Design Society*, vol. 2, pp. 525–532, 2022. DOI: 10.1017/pds.2022.54

[31] T. Sheveleva, O. Koepler, I. Mozgova, R. Lachmayer, and S. Auer, "Development of a domain-specific ontology to support research data management for the tailored forming technology," *2351-9789*, vol. 52, pp. 107–112, 2020, CPAPER, ISSN: 2351-9789. DOI: 10.1016/j.promfg.2020.11.020

[32] M. Smith et al., "Dspace," *D-Lib Magazine*, vol. 9, no. 1, 2003. DOI: 10.1045/january2003-smith

[33] M. Krötzsch and D. Vrandecic, “Semantic wikipedia,” in *Social Semantic Web*, ser. X.media.press, A. Blumauer and T. Pellegrini, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 393–421, ISBN: 978-3-540-72215-1. DOI: `10.1007/978-3-540-72216-8_19`

[34] N. Saint et al., “Active research data management with the django globus portal framework,” pp. 43–51, 2023. DOI: `10.1145/3569951.3593597`

[35] K. Chard, Foster, Ian, and S. Tuecke, “Globus: Research data management as service and platform,” 2017. DOI: `10.1145/3093338.3093367`

[36] K. Chard, S. Tuecke, Foster, and Ian, “Globus: Recent enhancements and future plans,” *ACM International Conference Proceeding Series*, vol. 17, 2016. DOI: `10.1145/2949550.2949554`

[37] K. Czajkowski, Kesselman, Carl, R. E. Schuler, and H. Tangmunarunkit, “ERMrest: a web service for collaborative data management,” 2018. DOI: `10.1145/3221269.3222333`

[38] A. Ellerm, M. Gahegan, and B. Adams, “LivePublication: The Science Workflow Creates and Updates the Publication,” 2023. DOI: `10.1109/e-Science58273.2023.10254857` [Online]. Available: `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174243942&doi=10.1109%2fe-Science58273.2023.10254857&partnerID=40&md5=38e71996748cafda467de1e05de64314`

[39] D. Nikolov and E. Tuna, “A Lightweight Framework for Research Data Management,” 2019. DOI: `10.1145/3332186.3333157`

[40] D. Stansberry, S. Somnath, J. Breet, G. Shutt, and M. Shankar, “DataFed: Towards reproducible research via federated data management,” *arXiv*, 2020, ISSN: 23318422.

[41] D. Betz et al., “Canonical workflow for experimental research,” *Data Intelligence*, vol. 4, no. 2, pp. 155–172, 2022, Journal. DOI: `10.1162/dint_a_00123`

[42] L. Peer, C. Biniossek, D. Betz, and C. Thu-Mai, “Reproducible research publication workflow: A canonical workflow framework and fair digital object approach to quality research output,” *Data Intelligence*, vol. 4, no. 2, pp. 306–319, 2022, Journal. DOI: `10.1162/dint_a_00133`

[43] P. Wittenburg et al., “Canonical workflows to make data fair,” *Data Intelligence*, vol. 4, no. 2, pp. 286–305, 2022, Journal. DOI: `10.1162/dint_a_00132`

[44] M. I. Bugaje, “A novel framework for user-centered research data management,” *ProQuest Dissertations and Theses Global*, 2019. [Online]. Available: `https://www.proquest.com/openview/dfa2890a9c50390e6ca5b93f4f5790ac`

[45] M. Bugaje and G. Chowdhury, “Identifying design requirements of a user-centered research data management system,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11279, pp. 335–347, 2018, ISSN: 03029743. DOI: `10.1007/978-3-030-04257-8_35`

[46] O. Knodel, R. P. Martin Ufer, M. David Lokamani, and T. Stefan E.Gruber, “Heliport: A portable platform for fair workflow | metadata | scientific project lifecycle management and everything,” pp. 9–14, 2021. DOI: `10.1145/3456287.3465477`

[47] S. Ferenz et al., “An open digital platform to support interdisciplinary energy research and practice—conceptualization,” *Energies*, vol. 15, no. 1, p. 6417, 2022, Journal. DOI: 10.3390/en15176417

[48] O. Werth, S. Ferenz, and A. Nieße, *Requirements for an Open Digital Platform for Interdisciplinary Energy Research and Practice: Wirtschaftsinformatik* (Proceedings). 2022, vol. 2. [Online]. Available: https://aisel.aisnet.org/wi2022/sustainable_it/sustainable_it/2

[49] S. Samuel and B. König-Ries, “A collaborative semantic-based provenance management platform for reproducibility,” *PeerJ Computer Science*, 2022, Journal. DOI: 10.7717/peerj-cs.921

[50] D. Hornung, F. Spreckelsen, and T. Weiß, “Agile research data management with open source: Linkahead,” Ing.grid, 1(2), *ing.grid*, vol. 1, 2024, ISSN: 2941-1300. DOI: 10.48694/inggrid.3866

[51] T. Hamann, C. Florides, A. Abdelrazeq, and H. R. Schmitt, “Proof-of-concept-studie für das fdm in den ingenieur:innenwissenschaften, Ein forschungsdatenmanagement-rahmenwerk,” *wt Werkstattstechnik online*, 2025, ISSN: 1436-4980. DOI: 10.37544/1436-4980-2024-11-12-57

[52] L. Hofeditz et al., “How to design a research data management platform? technical, organizational and individual perspectives and their relations,” in *Human Interface and the Management of Information. Interacting with Information*, ser. Lecture Notes in Computer Science, S. Yamamoto and H. Mori, Eds., vol. 12185, Cham: Springer International Publishing, 2020, pp. 324–337, ISBN: 978-3-030-50016-0. DOI: 10.1007/978-3-030-50017-7_23

[53] W. D. Chawinga and S. Zinn, “Global perspectives of research data sharing: A systematic literature review,” *Library & Information Science Research*, vol. 41, no. 2, pp. 109–122, 2019, PII: S074081881830330X, ISSN: 07408188. DOI: 10.1016/j.lisr.2019.04.004

[54] A. M. Elsayed and E. I. Saleh, “Research data management and sharing among researchers in arab universities: An exploratory study,” *IFLA Journal*, vol. 44, no. 4, pp. 281–299, 2018, ISSN: 0340-0352. DOI: 10.1177/0340035218785196

[55] TU Darmstadt, *Leitlinien zum Umgang mit digitalen Forschungsdaten an der TU Darmstadt*, 2015. Accessed: Apr. 26, 2025. [Online]. Available: https://tuprints.ulb.tu-darmstadt.de/23200/1/Leitlinien_Forschungsdaten_2022_de.pdf

[56] RWTH Aachen University, *Leitlinie zum Forschungsdaten-management an der RWTH Aachen - RWTH AACHEN UNIVERSITY - Deutsch*, 2024. Accessed: Oct. 21, 2024. [Online]. Available: https://www.rwth-aachen.de/cms/root/Forschung/Forschungsdatenmanagement/~ncfw/Leitlinie-zum-Forschungsdatenmanagement/

[57] K. Diederichs et al., *DMP4NFDI - NFDI Basic Service for Data Management Plans*, 2024. DOI: 10.5281/zenodo.13445355

[58] I. Lang and M. Politze. “Coscine,” Accessed: Aug. 9, 2023. [Online]. Available: https://coscine.rwth-aachen.de/

[59] Coscine-JARDS. “Coscine-JARDS | electronic project application system for coscine data projects,” Accessed: Dec. 18, 2025. [Online]. Available: `https://jards.coscine.de/`

[60] M. Politze, B. Heinrichs, S. Hunke, I. Lang, and T. Eifert, “Fair digital objects: Fairtilizer for the digital harvest,” in *Proceedings of EUNIS 2024 annual congress in Athens*, R. Vogl et al., Eds., ser. EPiC Series in Computing, vol. 105, EasyChair, 2025, pp. 284–292. DOI: `10.29007/hfzk`

[61] I. Lang, K. Jansen, and M. Politze, *Implementing the FAIR Principles with Coscine*, 2025. DOI: `10.5281/zenodo.14886334`

[62] M. Grönewald et al., “Mit AIMS zu einem Metadatenmanagement 4.0: FAIRe Forschungsdaten benötigen interoperable Metadaten,” *E-Science-Tage 2021*, 2022. DOI: `10.11588/heibooks.979.c13721`

[63] ing.grid. “Focus and scope,” Accessed: Dec. 20, 2023. [Online]. Available: `https://www.inggrid.org/site/focusandscope/`

[64] BausteineFDM. “Über die zeitschrift | bausteine forschungsdatenmanagement,” Accessed: Dec. 20, 2023. [Online]. Available: `https://bausteine-fdm.de/about`

[65] DFG, *DFG - GEPRIS - Applying Interoperable Metadata Standards (AIMS) 2 - A Platform for Creating and Sharing Metadata Standards and their Integration into Scientific Workflows*, 2025. Accessed: Oct. 23, 2025. [Online]. Available: `https://gepris.dfg.de/gepris/projekt/432233186`