A Metadata Model for Harmonising Engineering Research Data Across Process and Laboratory Boundaries

Martin Zinner*, Felix Conrad*, Kim Feldhoff*, Hajo Wiemer*, Jens Weller†, and Steffen Ihlenfeldt*‡
*Chair of Machine Tools Development and Adaptive Controls (LWM)
Institute of Mechatronic Engineering (IMD)
Technische Universität Dresden
Dresden, Germany
Email: {martin.zinner1, felix.conrad, kim.feldhoff, hajo.wiemer, steffen.ihlenfeldt}@tu-dresden.de
†Symate GmbH
Dresden, Germany
Email: jens.weller@symate.de
‡Fraunhofer Institute for Machine Tools and Forming Technology (IWU)
Dresden, Germany

***Abstract*—The availability of precise and comprehensive experimental data in science and technology is crucial for the usability of Artificial Intelligence (AI) models. A digitally analysable, system-independent representation of datasets is essential for enabling the deployment of data-driven applications across different platforms. We propose a metadata model based on domain-specific languages and terminologies, that allows researchers to focus on data provision by reducing routine activities rather than attempting to align with other research groups. Furthermore, it enables fast and efficient integration of new partners from different laboratories and disciplines. To conclude, our approach supports a paradigm shift away from more or less subjectively designed individualistic conceptions in handling research data towards objectively established harmonised solutions. The approach is illustrated for an Interdisciplinary Research Training Group, in which researchers from more than 10 different departments are involved with the main research profiles, such as textile and polymer technology and material sciences.**

***Index Terms*—Metadata Model; FAIR Principles; Research Data Management; Ontology; Machine Learning; Domain-Specific Technical Languages.**

I. Introduction

Metadata is data about data [1], i.e., metadata provides information about one or more aspects of the data. This layered structure enhances the ability to capture subtle data relationships, thereby improving data management and analysis. To work effectively with metadata, organisations can use several key tools and technologies, including taxonomies, ontologies and semantics.

Ontologies and taxonomies are key tools used by researchers to understand and retrieve large quantities of scientific and engineering data. However, the management and application of ontologies themselves can prove to be a daunting task. Although similar in function, ontologies and taxonomies differ in complexity. Taxonomies have a hierarchical structure and use only parent-child relationships, while ontologies are considerably more complicated [2] [3]. In simple terms, an ontology represents the structured and formal knowledge related to a specific domain. The semantic system uses clear and understandable representations of concepts, relationships, and rules to develop knowledge. It is not possible to rely entirely on database programmers or data engineers to build a system, that considers target applications, such as materials or production technologies. They lack domain-specific knowledge, which is fundamental for characterising the associations between concepts. Therefore, to acquire domain knowledge, it is necessary to seek guidance from various domain experts [4].

Over the past decade, Machine Learning (ML) has gained significance in the fields of materials engineering. ML is a subset of the broader category of Artificial Intelligence (AI) that involves the development of algorithms and models that enable systems to learn and improve from data without explicit programming. AI encompasses a wider range of technologies integrated into a system that aims to facilitate reasoning, learning, and problem solving to address complex problems.

ML algorithms analyse vast amounts of data, extract insights, and use them to inform decision-making [5] to detect and extrapolate patterns. ML is becoming increasingly popular worldwide owing to the growing demand for data analysis solutions [6]. However, they also require large amounts of data, which may not be meaningful in many areas, partly due to the need for elaborate large-scale laboratory tests. There has been an increase in the utilisation of ML methodologies in materials science research [7]. Research suggests that the limited practicality of AI in certain domain-specific contexts, partially because of the need for elaborate laboratory tests on a large scale, is a significant obstacle to its application.

Focusing on industrial requirements, we developed a novel approach for the applicability of AI techniques, termed Usable Artificial Intelligence (UAI) [8]. At present, data-driven, machine learning, and artificial intelligence methods are not fully utilised to solve the associated technical challenges, especially in industrial applications, despite versatile development progress. This is mainly due to the limited practicability of AI solutions. Technical practitioners frequently depend on interdisciplinary collaboration with data science specialists to fully exploit the capabilities of AI methods [9]. In our work, a flexible, tractable, scalable, and adaptable technique

 ISBN: 978-1-68558-157-2

for constructing anticipatory models has been introduced [8] and it is demonstrated on a use-cases [9].

Multi-Task Learning (MTL) methodology, which is novel in materials informatics, can be utilised for example to learn and forecast various polymer characteristics simultaneously, efficiently and effectively [10]. MTL is a machine learning approach, in which multiple tasks are trained simultaneously, optimising multiple loss functions simultaneously. Rather than training independent models for each task, we allow a single model to learn to complete all tasks at once. In this process, the model uses all available data across different tasks to learn generalised representations of data that are useful in multiple contexts [11]. For example, multitask models can be utilised to overcome the data scarcity in polymer datasets. This approach is expected to become the preferred technique for training materials data [10].

Additionally, in other fields, existing predictive models struggle to capture the complex relationships between mechanical characteristics and behaviour. These studies used ML to predict the mechanical properties of carbon nanotube-reinforced cement composites [12]. Successful training, validation, and testing of ML and Deep Learning (DL) models require significant amounts of relevant data [13].

According to a survey, data scientists spend most of their time cleaning and organising data (60 %), collecting datasets (19 %), and mining data for patterns (9 %). Messy data are by far the most time-consuming aspect of typical data scientist's workflow [14].

There is an urgent need to enhance the infrastructure that facilitates the reuse of educational data [15]. In addition, it is necessary to consider that data governance is fundamental for other activities besides data within any Information Technology (IT) establishment.

Through analysis, it can be concluded that the difficulty of identifying, collecting, retaining, and granting access to all relevant data for organisations at an acceptable cost is significant. Data integration is a long-standing issue in data management, and the above observations attest to its continuing importance [16]. It is important to tap into the full potential of data to create added value. This provides new insights and justifies the costly initially data collection.

A. Motivation

In recent years, data-driven methods have significantly improved various engineering tasks by providing valuable insights, pattern recognition, and identification of the underlying relationships in complex datasets. This has led to remarkable progress and numerous potential data-driven applications, including production engineering [17] and materials science [18]. However, the availability and usability of underlying data are fundamental to the application of these methods.

In engineering, proper documentation of research data is highly significant as experiments are often complex, intricate and elaborate. Inadequate data documentation can lead to the misinterpretation of experiments by other researchers and/or unnecessary repetition of already completed experiments, with data that are publicly accessible in repositories. High-quality data documentation is crucial for researchers seeking to understand the relationships among the processes, structures, and properties of manufactured components. This is sought and increasingly demanded by public project sponsors, such as the German Research Foundation (DFG).

Multi-stage manufacturing in the process chains is common for many products. Cross-process data analysis can be used to identify relationships in process chains. A prerequisite for this is that an evaluable, comprehensive and well-documented global dataset is available [19]–[23]. Nevertheless, acquiring such a dataset across process boundaries presents a formidable obstacle, owing to the distinct handling of individual process steps by different partners.

To facilitate cross-platform implementation of AI models, a digitally analysable, system-independent dataset representation is necessary. These datasets can be combined to form a unique dataset that represents different system properties, ultimately enabling holistic data-driven modelling, for example, through MTL or transfer learning. This will enable the harmonisation of workflows across diverse domains, thereby facilitating communication between areas of expertise or specialists themselves. An overarching strategy is key to aligning different approaches ensuring tha the experimental data are reusable without modification.

We propose a strategy that allows specialists to focus on data provision by reducing routine activities rather than aligning with similar groups. This strategy enables researchers to focus on their experiments and research questions. The objective was to document research data across process boundaries, thereby enabling researchers to maintain their perspectives during data preparation and documentation. Metadata schemas with synonyms grounded on ontologies or taxonomies guarantee that research data that is understandable, usable for further analyses, interoperable across laboratory boundaries, replicable at a qualitative level, complete, and of superior quality.

In conclusion, the main motive of our research is to support data-driven analysis and modelling, including comparisons across laboratory boundaries. Datasets from different laboratories are to be merged, for example, for round robin-tests, etc. For multi-stage process chains this reveals overarching correlations in the overall dataset. Taking into account the FAIR principles [15], i.e., Findability (F), Accessibility (A), Interoperability (I), and Reusability (R), third party researchers will be able to understand and analyse datasets from disciplines that are unfamiliar to them for their own research questions.

B. Challenges

Effectively documenting data across processes and laboratory boundaries presents a key challenge. At the heart of these challenges is the need for data to comply with the guidelines of good scientific practice and the FAIR principles. Each domain possesses its own technical language and unique working culture that needs to be integrated, while allowing researchers to retain their languages. A clear and concise example of the objective can be illustrated by the symbols and units of

 ISBN: 978-1-68558-157-2

measurement used for the tensile strength in tensile tests. The standards differ in the symbols for the tensile strength, which are used for different materials. The ISO 1920-4 standard for the tensile strength of concrete specifies f_{ct} as a symbol for tensile strength, ISO 6892-1 for metals specifies R_m, ISO 527-1 for plastics specifies σ_m, and the RILEM TC 232-TDT [24] technical guideline for textile-reinforced plastics specifies σ_{cu} as a symbol for tensile strength. In addition, the frequently used units of measurement differ, which are MPa (Megapascal) and GPa (Gigapascal).

A common technical language that allows researchers from different domains to communicate effectively is relevant, without the necessity for a uniform overarching technical language across all laboratories. Local technical terminology should be compatible without the need for a uniform overarching technical language across all laboratories. This can improve recognition and reduce expenses, furthermore, it is necessary for interoperability. We are not aware of any other method for integrating data records. The completeness of reporting is also critical. Researchers from various disciplines consider – due to their specific research questions – different quantities to be significant, leading to incomplete and inconsistent data documentation across process and laboratory boundaries. Therefore, complete data documentation is relevant for the subsequent use of data by third parties and adherence to the principles of good scientific practice is indispensable to ensure accuracy. Sometimes, if experiments were carried out a long time ago, it is not always possible to remember the details, especially because staff turnover in the research sector is very high.

C. Aim

The main objective of this paper is to provide a workable approach towards synchronised documentation of research data within the engineering sector across various phases while meeting all the requirements regarding the FAIR principles. It is specifically geared towards enabling researchers to maintain their domain-specific perspective during the data preparation and documentation phases. The goal is to develop a methodology that is achievable, extensible, and effective for promoting cross-platform functionality. The deployment of AI models is facilitated through a digitally analysable, system-independent presentation of training datasets that enable cross-process data analysis.

D. Contribution

We present a solution concept in which research data can be documented based on a subject-specific ontology. The feasibility of the concept is illustrated using an example of the documentation of compression tests as part of the joint GRK2250 project [25]. The concept is largely independent of the above project and can therefore be easily transferred to other collaborative projects. Subsequent data-driven modelling is outside the scope of this study.

To conclude, our strategy supports a paradigm shift from more or less subjectively designed individualistic conceptions to the handling of research data towards objectively established harmonised solutions. The motivation for this work is the importance of harmonised data preparation and subsequent documentation in the engineering field. The impetus for this work comes from recognising the fundamental significance of standardising data preparation and subsequent documentation in the engineering domain.

E. Paper organisation

The remainder of the paper is structured as follows: Section II provides an overview of existing work related to the described problem. A description of our strategy is presented in Section III. In Section IV, the strategy is illustrated for an Interdisciplinary Research Training Group, in which researchers from more than 10 different departments are investigating mineral-bonded composites for improved structural impact safety.

The presentation of the main results and discussions based on these results constitute the content of Section V. Section VI summarises our contributions and draws perspectives for future work.

II. Related work

This section offers an overview of the existing approaches to metadata schemes for research software. Whereas some publications focus specifically on metadata, others introduce software ontologies that can serve as a vocabulary for research software. A recent summary [26] of existing approaches to metadata schemes for research software includes DataCite [27], CodeMeta [28], and EngMeta [29], etc. The international consortium DataCite was founded in late 2009, to address the ever-increasing amount of digital research data. The objectives of the consortium include promoting the acceptance of research data to facilitate data archiving and enabling future studies to verify and repurpose the results.

CodeMeta is a community driven metadata standard for research software based on the schema.org. Various crosswalks to other metadata schemes already exist. CodeMeta contains multiple elements, some focusing on technical details, such as file size or supported operating systems and others including administrative information, such as licenses. The metadata standard does not have mandatory elements. It supports the use of uniform research identifiers for authors and contributors as well as licenses. Content-specific metadata are limited to the application categories and keywords.

EngMeta is an XML-based formal definition of the information required to find, understand, reproduce, and reuse data from engineering disciplines [29]. It uses a metadata schema for the description of engineering research data and the documentation of the entire research process, including the people, software, instruments and computing environment involved, as well as the methods used and their parameters [30] [31].

In general, the more precise the data documentation that can model a specialist area, the more suitable it is. This means that general ontologies, on which knowledge databases, such as WikiData [32] or DBpedia [33] are based, are only suitable

 ISBN: 978-1-68558-157-2

to a limited extent for use in highly specialised fields of application, such as additive manufacturing. The European Materials Modelling Ontology (EMMO) [34] is an approach for standardising technical terms in applied sciences, particularly in materials science. It can be used to model experiments and simulations.

OntoSoft [35] captured scientific software metadata, and expanded them using machine-readable descriptions of the expected content of the inputs and outputs of the software. The EDAM ontology contributes to open science by enabling the semantic annotation of processed data, thus making the data more understandable, findable, and comparable [36]. Software Ontology (SWO) [37] has been developed to extend the EDAM ontology to describe software in this research area [38]. SWO includes licences, programming languages, and data formats as taxonomies. In contrast to OntoSoft, the use of taxonomies improves the usability of semantic web applications and links [39].

Several universal metadata standards are available in the literature, and metadata schemes have been used in online retail for over a decade. More than 100 metadata standards were visualised in [40]. Additionally, metadata standards related to engineering domains are available, such as EngMeta [31]. However, metadata templates for specific experiments are lacking. Even in experiments standardised according to the German industry standard (DIN), there is no guidance on what metadata should be stored. The standards focus on the execution of experiments rather than on managing the data collected during the experiments.

There are already a number of different Research Data Infrastructures (RDIs) for collaborative projects in the fields of engineering sciences, such as the Karlsruhe Digital Infrastructure for Materials Science (Kadi4Mat) [41]. The software includes many features for data management and collaborative work in joint projects (including web-based access, fine-grained role management, creation of reproducible workflows, and publication of research data). This basic functionality can be easily extended using plug-ins. Metadata schemas with key-value pairs are commonly used to document data in a machine-readable form [42], usually stored in XML or JSON format.

III. STRATEGY

We begin by examining some basic concepts. There is a continuous need to enhance infrastructure that supports the reuse of research data. To this end, a concise and measurable set of principles has been developed to govern the reusability of research data, known as FAIR Data Principles [15]. These foundational principles, namely Findability (F), Accessibility (A), Interoperability (I), and Reusability (R) are guidelines for those wishing to improve the quality of their data. However, they also have wider applicability, as researchers who wish to share and reuse their data can benefit from them. They can also be used by professional data publishers, who offer services and expertise in this area.

It is important to note that these values did not end. Data sharing and collaboration are important elements of scientific research. Researchers must share and collaborate in order to broaden their knowledge and perspectives. They must rely on each other's data and interpretations without bias. However, researchers must always maintain objectivity and balance when using technical terminology and adhering to conventional academic structures. It should be applied not only to data in the traditional sense, but also to the algorithms, tools and workflows that produce it. The emphasis on fairness, which applies to both human and machine activities is the focus of the FAIR Guiding Principles. Good data management is not an end in itself, but rather the key to knowledge discovery and innovation, and to the subsequent integration and reuse of data and knowledge by the community after data publication [15].

An ontology describes the structure of data, including classes, properties, and relationships within a particular field of knowledge, ensuring consistency and understanding of the data model. Description Logics (DLs) provides fundamental concepts and information about this family of logic, which has become increasingly important in recent years as the formal basis for most contemporary applications. The Web Ontology Language (OWL) family includes expressive ontology languages [43] [44] An ontology is expressed using OWL 2 QL (query logic). A query is expressed using SPARQL, a mapping is expressed using R2RML, whilst SPARQL is the standard query language for RDF data. Ontologies offer several advantages over the relational and object models. They allow a strict definition of conceptual schemas and enable systems to understand the semantics of the data [45].

The proposed approach is based on an information structure that includes keys (term classes) and values (concrete expressions of terms). The keys are derived from ontology, whereas the values reflect the potential forms of metadata, such as the value ranges for numerical properties, etc. Examples of process-specific metadata include characteristics such as the ultimate tensile strength in tensile tests.

- Schema Structure: Metadata schemas are created for all investigations, which serve as a template for recording the metadata. These schemas are divided into chapters, whereby some chapters are the same for all investigations and other chapters are adapted to the respective investigation.
- Ontologies: Key names are generated based on domain-specific ontology. Key names are used in process-specific metadata schemas to document the research data. Keys and values are filled with specific terms from the taxonomy and ontology to obtain concrete values. Specific ontologies and taxonomies exist for each domain or process.
- Thesaurus: The creation of a common language that encompasses all processes of the involved domains is necessary. A thesaurus is a domain-specific dictionary of synonyms that lists technical terms that have the same meaning or are similar to the technical terms. This helps ensure that an individual researcher can maintain his familiar terminology, even in an interdisciplinary

 ISBN: 978-1-68558-157-2

environment. Constant readjustment of new partners is unnecessary. Nevertheless, it is ensured that the data remain comprehensible and, therefore, usable for other researchers, both internally and externally. The researcher's own language is linked to the master language through synonyms.

The metadata schema has a clearly defined structure. The static metadata chapters have the same keys for all investigations and are mandatory. Dynamic metadata chapters must be redefined for each investigation along with domain experts to ensure the reproducibility of the investigation.

Key names are created using a domain-specific ontology and then used in metadata schemas that are specific to the research process. To ensure specific, concrete values, specific terms obtained from the ontology are used to populate the keys and values of metadata schemas. A universal language that covers all relevant domains is essential. Each domain has its own specific ontologies that must be merged into a global ontology. Domain-specific dictionaries of synonyms (thesauri) enable researchers to use familiar terminology even in interdisciplinary settings, as they list technical terms that have the same or similar meanings. This eliminates the need to constantly adapt to new partners.

The resulting data can easily be shared with other researchers. Thesauri were included to allow the laboratories to retain their preferred terminology. In addition, researchers complete metadata schemas in their native languages. Nevertheless, the information is presented objectively to facilitate understanding by researchers inside and outside the laboratory. The researcher's own language is linked to the master language using synonyms.

IV. Use cases: An excerpt

In order to exemplify our concept as outlined in Section III, we present a research case, i.e., the Interdisciplinary Research Training Group "Graduiertenkolleg 2250" (GRK 2250) [9], [25], which is dedicated to the investigation of mineral-bonded composites for improved structural impact safety [8]. This project provides an overview of common procedures, such as experiments, tests, numerical simulations, and manufacturing. Another special feature of the project is that three cohorts are planned to be established, with each cohort being worked on by a different team. This makes data transfer very important.

The proposed solution is based on an objective information structure known as metadata schema. This schema comprises keys representing term classes and values that express the meaning of the term more concretely.

GRK 2250 was established in 2017 and is currently funded by the German Research Foundation (DFG). Researchers from nine different departments and four faculties at TU Dresden and the Leibniz Institute of Polymer Research, IPF Dresden, were involved in the program organised in three consecutive three-year periods. Each cohort comprised 12 researchers representing the main research profiles, such as textile technology, polymer and material sciences, construction materials, structural engineering, continuum mechanics, numerical modelling, 3D optical monitoring techniques, sustainability, resilience, and machine learning. The scope of research ranges from the microscale to the structural scale and includes experimental, numerical and data-based investigations. Examples of investigations at the microscale include fibre pull-out tests and corresponding simulations. At the structural level, for example, drop tower tests are carried out in a 10-metre drop tower with plates measuring 1.5 metres by 1.5 metres by 30 centimetres and accompanied by corresponding simulations.

The current status of research data management within GRK 2250 shows considerable variation in the amount of data, ranging from a few megabytes to several hundred megabytes per experiment. The cumulative data volume of 3 terabytes was stored from 15 test systems in six different laboratories. Each test system was conducted for 20 to 300 experiments. To support this diverse dataset, the research data infrastructure consisted of a shared drive that could be accessed by all project partners and Excel spreadsheets dedicated to the documentation of the research data. The data documentation workflow involves manually storing the research data in appropriate folders within the group drive. Researchers manually entered the metadata into an Excel spreadsheet, which was automatically named according to a specific scheme. The completed Excel file was then saved to the corresponding research data folder on a shared drive.

The availability of appropriate data in materials science has a major impact on the performance of the applied AI models [6] [7]. Therefore, data management is particularly important to the usability of AI models. To consider and analyse cross-process relationships, a global view of the dataset in an analysable form is required. This requires well-documented data, which can be combined into a global dataset.

Interdisciplinary research networks bring researchers together with different specialised ontologies. To work together effectively in this case, i.e., in order to be able to understand the data of the research partners, a common ontology is required. For example, a measure for the amount of textile per unit for textile reinforcement of concrete is typical weight per unit area (unit: $\mathrm{g/m^2}$) in textile engineering, whereas, in civil engineering, it is the cross-sectional area per linear metre (unit: $\mathrm{m^2/m}$).

The current solution envisages a top-down approach, i.e., there is a group/person responsible for metadata management within the research network. This defined the standard ontology used in the research network. Enforcing the use of a specified standard ontology may lead to poor overall acceptance by participants. This is because the ontology may overwrite terms that have been established in their domain for years. It is important to consider the difficulty of pushing changes through.

The participants/working groups may be involved in other projects or networks in addition to the specific research network. These projects/networks may have agreed upon a different standard ontology. Therefore, conflicts may arise, since researchers must constantly switch between those ontologies.

As already mentioned, the acceptance of specialists of new ontologies is poor, as acceptance decreases the risk of errors increases. Our response to this wrong-headed development are summarised in Figure 1. This example illustrates the basic

 ISBN: 978-1-68558-157-2

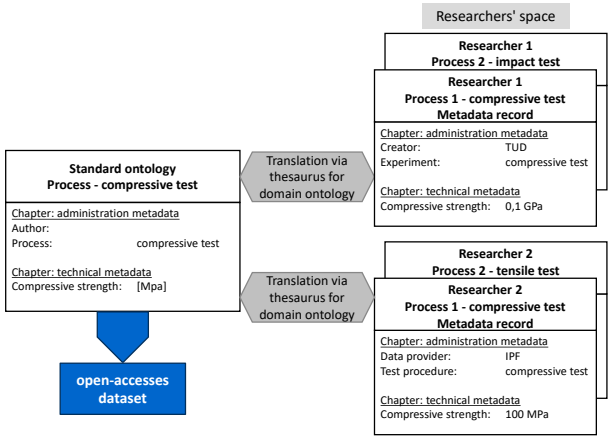

Figure 1. Symbolic representation of the metadata management workflow accessing a mapping thesaurus.

difficulty to overcome the different representation of metadata. As depicted, researchers from different institutes use different terminology for "Author" and different units (MPa/GPa) to record the compression strength, hence the values are also different. The basic idea is to introduce an intermediate layer (thesaurus) as the translation layer. This allows each researcher to use his own "laboratory ontology" or "researcher ontology". This is then used for the overall project, and translated into the "standard ontology" of the research network.

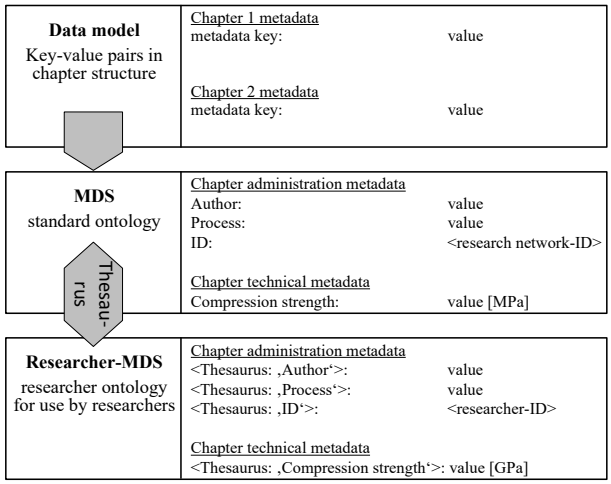

Figure 2. Extended symbolic representation of the metadata management workflow

The data model illustrated in Figure 2 as the general structure of a Metadata Schema (MDS) is divided into several chapters. These chapters contain metadata-key/metadata-value pairs that represent the metadata itself, a method that can generally be used To investigate specific chapters, the keys must be redefined for each investigation. As exemplified, MDS is divided into administrative and technical metadata.

The technical metadata schema (investigation specific) has to be adapted to the specific use case of data generation, whereas the administrative metadata (investigation-independent) has general validity.

There should be a standard ontology agreed upon in the research network. Accordingly, this ontology is also used when data are made publicly available. Standards already in use should be used to facilitate communication with external parties, for example, EngMeta [31] or the European Materials Modelling Ontology (EMMO) [46].

The Metadata Schema is a general structure that is then applied to the specific experiments/data. It can be partitioned into "static" and "dynamic" metadata, the static metadata is similar for all data (i.e., general data, e.g., author, date, etc.), whereas, the dynamic metadata is experiment-specific (i.e., specifically adapted to the experiment, e.g., temperature, test speed, etc.). Adjustment of dynamic metadata should be performed together with domain experts. The MDS is linked to the standard ontology via a translation layer. This allows the researcher to view the schema in their usual domain-specific language. To share the data, the metadata are translated back into a standard ontology and can therefore be understood by everyone.

An initial solution proposal and translation of the ontology has already been published [9]. Furthermore, it explains in more detail what data and metadata management must be able to do in order to support data-driven applications. Figure 3 presents an overview of the FAIR Data Principles proposed by Wilkinson. These principles were expanded to include the principle of "usability" to ensure their practical implementation.

For the extension of the FAIR principles of Wilkinson et al. [15] to "usable FAIR" as depicted in Figure 3, the GRK is working with the company Symate [47] to extend their software Detact. Detact is a cloud-based software for collecting data from various sources along the process chains for subsequent automated data analysis. As stated on the home page from Detact: "Now we are able to merge the material data across disciplines and from different data sources. This leads us to completely new industrial planning and control systems in the sense of the Fourth Industrial Revolution ('Industry 4.0')" (Leibniz IPF) [48]. A major aspect of this study was the development of a process modeller.

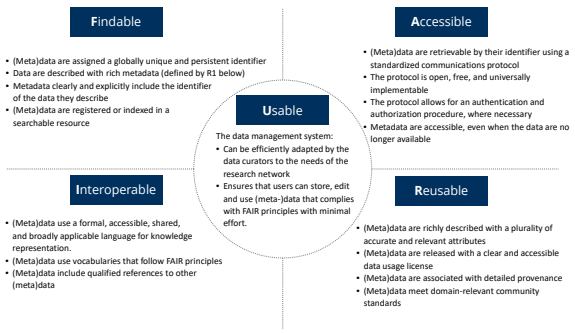

Figure 3. Extension of the FAIR principles of Wilkinson et al. [15] to "usable FAIR".

Figure 4 shows the planned process modeller used to generate the metadata. In addition to conducting the experiments, the researcher created a process data model for the entire experimental process chain. This model consists of four blocks

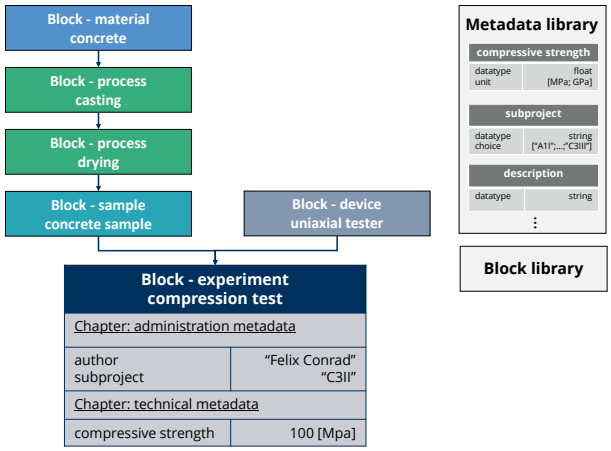

Figure 4. Schematic representation of the process modeller diagram for the use case “uniaxial tensile test of a textile-reinforced concrete test specimen”.

(Material, Processes, Devices, and Experiment). The metadata associated with each block can be recorded within that block. All the blocks had the same metadata structure, as depicted in our example, and both administrative and technical metadata were included.

In this particular instance (depicted in Figure 4), the compression test of a concrete test specimen was modelled. The material used is precisely defined in the first block “material concrete”, in which the exact composition is recorded. This is followed by the casting of the material defined in the block “process casting” block and the subsequent drying of the material defined in the block “process drying”. These steps lead to the final test specimen. The test is then carried out in a testing machine represented by block “device uniaxial tester". The test procedure can then be recorded in the block “experiment compression test”.

Figure 4 shows only the metadata of the Block ’Experiment - compressive test’ for the purpose of simplification. However, all other blocks also have metadata in the same structure. The corresponding metadata keys can be added to the blocks from a predefined library, known as the metadata library. The experimental process chain for the concrete compression tests, as shown in Figure 4, is relatively simple and short. In this case, the metadata could be included in a single scheme for the concrete compression test experiment, without the process data model.

In real-life applications, the process chains are often much longer and more complex, as shown, for example, by the work in the research network Research Training Group GRK 2250 “Mineral-bonded composites for enhanced structural impact safety” [25], where the process chain begins with mixing the concrete and fibre production and extends to the reinforcement of an existing component with textile-reinforced concrete. Another example is the Collaborative Research Centre SFB 639 “Textile-reinforced composite components for function-integrating multi-material design in complex lightweight applications” [49], where the process chain ranges from the fibre and plastic to the complex function-integrated component. In addition, as shown in the example, the complexity increases very quickly if we use a composite material, such as textile-reinforced concrete (TRC). The block “material concrete” would be accompanied by the block “material - textile”, which together would then form the block “material - TRC”. The block “material textile” itself can also have upstream blocks, like “process textile manufacturing”, “process fibre formation”, etc.

This information is required to manage trial data according to FAIR principles. Attempting to fit all of these metadata into a single schema without a process data model will most likely result in important parameters being omitted, hence the need for the process data model. According to the example shown in Figure 4, the drying time of concrete has a significant influence on its strength. If the drying time was not recorded, the data would be essentially unusable for further analysis as a major influence is not recorded.

All individual blocks and overall process data models can be collected and shared in a library, to unify the workflows of the cooperating researchers. Using the process data model, the workflows of other researchers can be easily understood and adopted. Here, the process data model makes communication between the researchers easier. The process data model provides the instruction on how the experiment/entire workflow is to be carried out. Standardisation of workflows is a fundamental step towards sustainable data management and research. Standardised workflows are significant for the comparability of experiments and, thus, for the reusability of the series of experiments. Currently, researchers create their own workflows during their research. Many experiments have not yet been standardised because they require a new type of experimental setup or a new type of material, the production of which is not yet standardised. Even small variations in the manufacturing process can significantly impact the target properties. This makes it difficult to compare/reuse the data. The standardisation of workflows also reduces unnecessary duplicate developments in common process steps.

With the implementation of the process data model in Detact, researchers do not have to scroll through “metadata lists”, but can use the graphical version of the process model. Access to Detact can occur via a web browser and metadata can be recorded directly in the laboratory. The comment function can be used to mark deviations that would otherwise not be recorded (e.g., “Concrete stickier than usual today”). Using the process model, metadata can be captured quickly and easily and subsequently recorded. A new test can be initialised with default values and can be easily adapted. Experience has shown that metadata are currently insufficiently (not completely) captured. The perceived costs for a complete recording are too high (see point “Usability”) and can be significantly reduced with the “process model”.

Based on this approach, it is possible to merge different experiments with the corresponding data sets into a global data set. The merging of the three different experiments into one dataset is sketched in Figure 5. The example above was adapted from the research project GRK 2250 [25]. The purpose

 ISBN: 978-1-68558-157-2

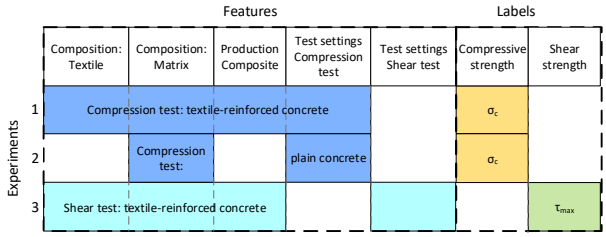

Figure 5. Schematic illustration of a section of the overall dataset of the research network.

of creating a combined dataset is to gain a comprehensive understanding of the material from a data perspective, such that in order to map the material behaviour, data-driven models can be trained on the basis of this broad dataset. To achieve this, the data from various material tests and, thus, properties are to be combined. If the same material is analysed in different tests, a corresponding dataset can be created, as outlined in Figure 5.

The benefit of using such a dataset is the wider range of information available for modelling, and the ability to determine the interactions between multiple influencing parameters. This allows the creation of models that can simultaneously map several material properties.

The importance of the metadata management methods outlined in this section is obvious, as experiments in research networks are frequently conducted by various researchers in different laboratories. Therefore, merging experimental data into a single dataset is feasible using a collaborative data management approach. The following three experiments were conducted:

1) Determination of compressive strength of textile reinforced concrete.
2) Determination of compressive strength of unreinforced concrete.
3) Determination of the tensile strength of textile reinforced concrete.

Each test had its own data space, which sometimes overlapped. The data available for each experiment are indicated by the coloured boxes. The absence of colour markings indicates the absence of the specific data. For example, unreinforced concrete does not have data on textile reinforcement, as this is not present and is outside the scope of the investigation. In this context, features refer to descriptive elements of the experiment, whereas labels denote the outcome of the experiment. Terms features and labels were chosen because they are typical terms in the field of machine learning. In detail, the representation for the experiments is as follows:

- Experiments 1 and 3 share the same features to describe textile-reinforced concrete, but have different features to describe the experiment itself, as there are different tests to determine the compressive and shear strength. In addition, different properties were determined, which are labelled here.
- Experiments 1 and 2 have the same features to describe the experiment and also have the same label (the same material property that is determined). However, in experiment 2 the features that are related to textile reinforcement are missing, since they do not apply to the experiment.
- Experiments 2 and 3 shared the same features for describing the concrete matrix. They have different features to describe different tests and, accordingly, different labels. In experiment 2, the features related to the textile reinforcement were missing.

V. Outline of the Results

The proposed strategy helps bridge different domain-specific languages and working cultures by providing a common language that all researchers and engineers from different domains can understand. This is achieved by using metadata. In particular, the metadata model helps unify physical units and terms. Data are stored and documented such that data from different processes along a process chain can be merged, resulting in a single overall dataset. Therefore, cross-process data analysis methods can be applied. The solution approach allows merging of research data in the following ways:

- Merging data from similar processes provided by different institutions or fields.
- Merging data from different processes along a process chain.

A multi-level metadata model connects different domain-specific languages by defining a common set of concepts and relationships that can be used in different domains. The model provides a method to manage metadata by defining a set of rules on how metadata should be structured and stored.

The drawbacks of the metadata model are as follows:

- The creation of a metadata template relies on the assistance of metadata experts, whereas the goal should be for researchers or domain experts to use/create it independently.
- The outcome depends on the domain experts. In the end, the solution approach allows for the creation of global datasets in an analysable manner. In this way, the interoperability and collaboration among different research groups in the engineering domain will be improved.
- The common language was created by surveying data providers to establish a shared technical vocabulary.

The proposed model can be used as a framework for managing digital objects in other research domains, such as the social sciences and natural sciences. Additionally, further research can be conducted to explore how this metamodel can be integrated into existing research data management systems or how it can be improved to better meet the needs of different users. Overall, the model provides a promising approach for addressing the challenges of research data management and improving collaboration among researchers and engineers from different domains. This solution ensures that the data are captured in a comprehensible manner through clear documentation, thus, it can be understood by other researchers too. The interoperability of data across laboratory boundaries is ensured by the proper identity management of components, processes, and machines across these boundaries.

 ISBN: 978-1-68558-157-2

This makes the data available for subsequent data-driven analysis across the laboratory and process boundaries. The analysis results based on the documented research data can be reproduced at a qualitatively high level owing to the detailed data documentation.

VI. Conclusion and future research perspective

The proposed strategy helps bypass different working cultures by providing a harmonised approach that all researchers and engineers from different domains can understand. This is achieved by using metadata enhanced by the development of adequate ontologies. In particular, the metadata model helps to store and document data, such that data from different processes along a process chain can be merged, enabling cross-process data analysis methods.

As a use case approach, this article also summarises the existing requirements of the GRK 2250 joint project for practicable research data management and presents a solution concept for the investigation of mineral-bonded composites within the GRK 2250 project. The concept is largely independent of GRK 2250 and can therefore be easily transferred to other collaborative projects.

Experience has shown that researchers need support in setting up a structured process data model. It is difficult for them to identify all the metadata that needs to be recorded so that the experiment is documented in a repeatable manner. As a result, it can happen that important influencing factors in the experiments were not recorded and the generated data is hardly reusable. The structured process data model (as exemplified in Figure 4) is intended to help identify all necessary steps and influences.

To consider and analyse cross-process relationships, a global view of the dataset in an analysable form is required. This requires well documented data that can be combined into global datasets [6] [7], as subsequent data-driven modelling is not part of this study.

The aspect of "usability" is not fully covered in this paper. It is still an open question as to whether and how it can be satisfactorily solved, since setting up a metadata management system that covers the FAIR principles is one thing, but in the end it only works if all participants are on the board. According to the experience gained through the use cases presented, it often fails in the end because there is an overhead due to the metadata management system for the individual researcher.

The aim of the Industrial Ontologies Foundry (IOF) initiative [50], is similar to that proposed for the OBO Foundry (for biomedicine) [51]. In both cases, commitment to a standard upper-level ontology plays a key role in supporting harmonisation. This upper-level ontology is termed Basic Formal Ontology (BFO) [52]. Consideration of how the current effort relates to this wider effort to curate and facilitate access to industrial ontologies might be a useful focus area for future research.

Acknowledgement

This research was partially funded by the German Research Foundation within the Research Training Group GRK2250/2 - Project C3 (grant number 287321140), by the Sächsische Aufbaubank (SAB), through the European Regional Development Fund (ERDF) and co-financed with tax revenue based on the budget approved by the parliament of the Free State of Saxony, Germany within the research project "AMTwin" (grant number 100373343), by the German Federal Ministry for Economic Affairs and Climate Protection (BMWK) based on the decisions made by the German Bundestag within the joint research projects "SWaT" (grant number 20M2112F) and "LaSt" (grant number 20M2118F), and by the BMWK in the funding guideline "Digitization of the vehicle manufacturers and supplier industry" in the funding framework "Future investments in vehicle manufacturers and the supplier industry", financed by the European Union, and supervised by the project sponsor VDI Technologiezentrum GmbH within the joint research project "Werk 4.0" (grant number 13IK022K). We thank the anonymous reviewers for their valuable feedback, which has significantly enhanced the quality of this paper.

References

[1] J. Riley, "Understanding metadata," *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf)*, vol. 23, pp. 7–10, 2017, retrieved: April 2024.

[2] M. Tainter. (2020) What is the difference between a taxonomy and an ontology? retrieved: April 2024. [Online]. Available: https://www.copyright.com/blog/taxonomy-vs-ontology/

[3] JANZZ.technology. (2019) Ontologies and the Semantic Web. retrieved: April 2024. [Online]. Available: https://janzz.technology/ontology-and-taxonomy-stop-comparing-things-that-are-incomparable/

[4] I. Horrocks. (2008) Ontologies and the Semantic Web. retrieved: April 2024. [Online]. Available: http://www.cs.ox.ac.uk/ian.horrocks/Publications/download/2008/Horr08a.pdf

[5] Google Cloud. (2023) Artificial intelligence (AI) vs. machine learning (ML). retrieved: April 2024. [Online]. Available: https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning

[6] F. Conrad, M. Mälzer, M. Schwarzenberger, H. Wiemer, and S. Ihlenfeldt, "Benchmarking AutoML for regression tasks on small tabular data in materials design," *Scientific Reports*, vol. 12, no. 1, p. 19350, 2022, number: 1 Publisher: Nature Publishing Group, retrieved: April 2024. [Online]. Available: https://www.nature.com/articles/s41598-022-23327-1

[7] Y. Zhang and C. Ling, "A strategy to apply machine learning to small datasets in materials science," *npj Computational Materials*, vol. 4, no. 1, pp. 1–8, 2018, number: 1 Publisher: Nature Publishing Group, retrieved: April 2024. [Online]. Available: https://www.nature.com/articles/s41524-018-0081-z

[8] H. Wiemer *et al.*, "Need for uai–anatomy of the paradigm of usable artificial intelligence for domain-specific ai applicability," *Multimodal Technologies and Interaction*, vol. 7, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2414-4088/7/3/27

[9] ——, "Illustration of the usable AI paradigm in production-engineering implementation settings," in *Artificial Intelligence in HCI*, ser. Lecture Notes in Computer Science, H. Degen and S. Ntoa, Eds. Springer Nature Switzerland, 2023, pp. 640–661, retrieved: April 2024.

[10] C. Kuenneth, A. C. Rajan, H. Tran, L. Chen, C. Kim, and R. Ramprasad, "Polymer informatics with multi-task learning," *Patterns*, vol. 2, no. 4, p. 100238, 2021, retrieved: April 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389921000581

[11] D. Soni. (2021) Multi-task learning in Machine Learning. retrieved: April 2024. [Online]. Available: https://towardsdatascience.com/multi-task-learning-in-machine-learning-20a37c796c9c

[12] J. Huang, J. Liew, and K. Liew, "Data-driven machine learning approach for exploring and assessing mechanical properties of carbon nanotube-reinforced cement composites," *Composite Structures*, vol. 267, p. 113917, 2021, retrieved: April 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0263822321003779

 ISBN: 978-1-68558-157-2

[13] F. Conrad, E. Boos, M. Mälzer, H. Wiemer, and S. Ihlenfeldt, "Impact of data sampling on performance and robustness of machine learning models in production engineering," in *Production at the Leading Edge of Technology*, ser. Lecture Notes in Production Engineering, M. Liewald, A. Verl, T. Bauernhansl, and H.-C. Möhring, Eds. Springer International Publishing, 2023, pp. 463–472, retrieved: April 2024.

[14] CrowdFlower. (2016) Data Science Report. retrieved: April 2024. [Online]. Available: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

[15] M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, 2016, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/sdata201618

[16] G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, and R. Rosati, "Using ontologies for semantic data integration," *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, pp. 187–202, 2018.

[17] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23–45, 2016.

[18] D. Morgan and R. Jacobs, "Opportunities and challenges for machine learning in materials science," *Annual Review of Materials Research*, vol. 50, pp. 71–103, 2020.

[19] N. Angrist, H. A. Patrinos, and M. Schlotter, "An expansion of a global data set on educational quality: a focus on achievement in developing countries," *World Bank Policy Research Working Paper*, no. 6536, 2013, retrieved: April 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2295861

[20] S. Lindersson, L. Brandimarte, J. Mård, and G. Di Baldassarre, "A review of freely accessible global datasets for the study of floods, droughts and their interactions with human societies," *Wiley Interdisciplinary Reviews: Water*, vol. 7, no. 3, p. e1424, 2020, retrieved: April 2024. [Online]. Available: https://doi.org/10.1002/wat2.1424

[21] J. T. Abatzoglou, S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch, "Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018, retrieved: April 2024. [Online]. Available: https://www.nature.com/articles/sdata2017191

[22] K. Großmann, H. Wiemer, and K. K. Großmann, "Methods for modelling and analysing process chains for supporting the development of new technologies," *Procedia Materials Science*, vol. 2, pp. 34–42, 2013, retrieved: April 2024. [Online]. Available: https://doi.org/10.1016/j.mspro.2013.02.005

[23] B. Awiszus *et al.*, "A holistic methodology to evaluate process chains." *Springer Berlin Heidelberg, part of Springer Nature 2022, L. Kroll (Ed.), Multifunctional Lightweight Structures - Resource Efficiency by MERGE of Key Enabling Technologies*, pp. 236–260, 2022, retrieved: April 2024. [Online]. Available: https://doi.org/10.1007/978-3-662-62217-9

[24] RILEM Technical Committee 232-TDT (Wolfgang Brameshuber), "Recommendation of RILEM TC 232-TDT: Test methods and design of textile reinforced concrete," *Materials and Structures*, vol. 49, no. 12, pp. 4923–4927, 2016. [Online]. Available: https://doi.org/10.1617/s11527-016-0839-z

[25] I. Curosu, V. Mechtcherine, M. Hering, and M. Curbach, "Mineral-bonded composites for enhanced structural impact safety–overview of the format, goals and achievements of the research training group grk 2250," *Bayonne, Frankreich*, 2019, retrieved: April 2024.

[26] S. Ferenz and A. Nieße, "Towards improved findability of energy research software by introducing a metadata-based registry," *ing.grid*, 2023.

[27] DataCite Metadata Working Group. (2016) DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. retrieved: April 2024. [Online]. Available: http://doi.org/10.5438/0012

[28] T. Habermann, "Mapping iso 19115-1 geographic metadata standards to codemeta," *PeerJ Computer Science*, vol. 5, p. e174, 2019, retrieved: April 2024. [Online]. Available: https://doi.org/10.7717/peerj-cs.174

[29] Metadata Standards Catalog. (2019) EngMeta. retrieved: April 2024. [Online]. Available: https://rdamsc.bath.ac.uk/msc/m100

[30] Informations- und Kommunikationszentrum der Universität Stuttgart (IZUS). (2019) EngMeta - Beschreibung von Forschungsdaten; Eng: EngMeta - Description of research data. retrieved: April 2024. [Online]. Available: https://www.izus.uni-stuttgart.de/fokus/engmeta/

[31] B. Schembera and D. Iglezakis, "EngMeta: Metadata for Computational Engineering," *International Journal of Metadata, Semantics and Ontologies*, vol. 14, no. 1, pp. 26–38, 2020, retrieved: April 2024. [Online]. Available: https://arxiv.org/pdf/2005.01637

[32] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014, retrieved: April 2024. [Online]. Available: https://doi.org/10.1145/2629489

[33] J. Lehmann *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic web*, vol. 6, no. 2, pp. 167–195, 2015.

[34] C. to emmo repo. (2024) Elementary multiperspective material ontology (emmo): A top-level ontology for applied sciences. retrieved: April 2024. [Online]. Available: https://github.com/emmo-repo/

[35] Y. Gil, V. Ratnakar, and D. Garijo, "Ontosoft: Capturing scientific software metadata," in *Proceedings of the 8th International Conference on Knowledge Capture*, 2015, pp. 1–4, retrieved: April 2024. [Online]. Available: https://doi.org/10.1145/2815833.2816955

[36] M. Black *et al.* (2021) Edam: The bioscientific data analysis ontology (update 2021)[version 1; not peer reviewed]. retrieved: April 2024. [Online]. Available: https://bora.uib.no/bora-xmlui/handle/11250/2988255

[37] M. Copeland, A. Brown, H. E. Parkinson, R. Stevens, and J. Malone, "The swo project: A case study for applying agile ontology engineering methods for community driven ontologies." *ICBO*, vol. 7, p. 2012, 2012, retrieved: April 2024. [Online]. Available: http://ceur-ws.org/Vol-897/session4-paper20.pdf

[38] J. Malone *et al.*, "The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation," *Journal of biomedical semantics*, vol. 5, no. 1, pp. 1–13, 2014, retrieved: April 2024. [Online]. Available: https://doi.org/10.1186/2041-1480-5-25

[39] A.-L. Lamprecht *et al.*, "Towards fair principles for research software," *Data Science*, vol. 3, no. 1, pp. 37–59, 2020, retrieved: April 2024. [Online]. Available: https://doi.org/10.3233/DS-190026

[40] J. Riley, "Seeing Standards: A Visualization of the Metadata Universe," 2018, retrieved: April 2024. [Online]. Available: https://doi.org/10.5683/SP2/UOHPVH

[41] N. Brandt *et al.*, "Kadi4mat: A research data infrastructure for materials science," *Data Science Journal*, vol. 20, pp. 8–8, 2021, retrieved: April 2024. [Online]. Available: https://doi.org/10.5334/dsj-2021-008

[42] S. Büttner, H.-C. Hobohm, and L. Müller, "Research data management," in *Handbuch Forschungsdatenmanagement.-Hrsg. von Stephan Büttner, Hans-Christoph Hobohm, Lars Müller.-Bad Honnef: Bock u. Herchen, 2011.-ISBN 978-3-88347-283-6*, 2011, retrieved: April 2024. [Online]. Available: https://opus4.kobv.de/opus4-fhpotsdam/files/192/1.1_Research_Data_Management.pdf

[43] S. Rudolph, "Foundations of description logics," in *Reasoning Web International Summer School*. Springer, 2011, pp. 76–136.

[44] F. Baader, *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[45] H. Kamal and B. Fouzia, "From relational databases to ontology-based databases," in *International Conference on Enterprise Information Systems*, vol. 2. SCITEPRESS, 2013, pp. 289–297.

[46] G. Goldbeck, E. Ghedini, A. Hashibon, G. Schmitz, and J. Friis, "A reference language and ontology for materials modelling and interoperability," 2019, retrieved: April 2024. [Online]. Available: https://publica.fraunhofer.de/handle/publica/406693

[47] (2024) Home Page Symate GmbH. retrieved: April 2024. [Online]. Available: https://www.symate.de/

[48] A. of detact.com. (2024) Home Page Detact. retrieved: April 2024. [Online]. Available: https://www.detact.com/en/

[49] N. Modler *et al.*, "Novel hybrid yarn textile thermoplastic composites for function-integrating multi-material lightweight design," *Advanced Engineering Materials*, vol. 18, no. 3, pp. 361–368, 2016, retrieved: April 2024. [Online]. Available: https://doi.org/10.1002/adem.201600028

[50] Industrial Ontologies Foundry. retrieved: April 2024. [Online]. Available: https://oagi.org/pages/industrial-ontologies

[51] Open Biological and Biomedical Ontology Foundry Community development of interoperable ontologies for the biological sciences. retrieved: April 2024. [Online]. Available: https://obofoundry.org/

[52] M. Drobnjakovic, B. Kulvatunyou, F. Ameri, C. Will, B. Smith, and A. Jones, "The industrial ontologies foundry (IOF) core ontology," *FOMI 2022: 12th International Workshop on Formal Ontologies meet Industry, September 12-15, 2022, Tarbes, France*, 2022, retrieved: April 2024. [Online]. Available: https://ceur-ws.org/Vol-3240/paper3.pdf