A Metadata Model for Data-Driven Applications in Engineering Sciences: a Use Case Approach

Kim Feldhoff*, Martin Zinner*, Felix Conrad*, Hajo Wiemer*, Jens Weller†, and Steffen Ihlenfeldt*‡
*Technische Universität Dresden
Institute of Mechatronic Engineering (IMD)
Chair of Machine Tools Development and Adaptive Controls (LWM)
01062 Dresden, Germany
Email: {kim.feldhoff, martin.zinner1, felix.conrad, 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 is crucial for the usability of Artificial Intelligence (AI) models. To enable the deployment of machine learning models across different platforms, a digitally analysable, system-independent representation of datasets is essential. The overall objective of this contribution is to document research data across process boundaries, as well as across laboratory boundaries and interdisciplinary fields of expertise, empowering researchers to maintain their usual domain specific perspective throughout the data preparation and documentation process. A strategy is proposed in this regard, whereby specialists can focus on data provision by reducing routine activities, rather than attempting to align with other groups. Metadata schemas with synonyms based on ontologies guarantee that research data is understandable, reproducible on a qualitative level, interoperable across laboratory boundaries, and useful for future analysis. The proposed meta-metadata model is formulated in a mathematical setting and its feasibility has been proven. The applicability of the strategy is demonstrated by integrating the model to the research data management of two joint research projects in the engineering domain. To conclude, the proposed strategy supports a paradigm shift away from more or less subjectively designed individualistic conceptions in handling research data towards objectively established harmonised solutions.**

***Keywords*-Metadata Model; FAIR Principles; Research Data Management; Domain-Specific Technical Languages; Ontology; Engineering Research Project; Artificial Intelligence; Machine Learning.**

I. Introduction

Firstly, it will be described, how the preceding research conducted by the authors on the topic of data alignment will be extended in this contribution. The paper is then motivated by the importance of comprehensive and accurate data documentation in engineering. This is followed by an enumeration of the main challenges associated with the harmonisation of research data across different organisations. In light of these challenges, a number of requirements are derived in Subsection I-C, including the need for comprehensive documentation and interoperability. The main objective of the paper is then presented in Subsection I-D. The section concludes with a structured overview of the contents of the paper in Subsection I-E.

A. Preliminary remarks

An extensive study on the alignment and harmonisation of engineering and research data across process and laboratory boundaries has been carried out by the authors in [1]. This paper extends that study by modelling the proposed metadata model in a mathematical setting and by demonstrating the overall validity of the approach.

B. Motivation

In recent years, data-driven methods have significantly improved various engineering tasks by providing valuable insights, pattern recognition and identification of underlying relationships in complex data sets. This has led to remarkable progress and numerous potential data-driven applications, including production engineering [2] and materials science [3]. However, the availability and usability of the underlying data is critical to the application of these methods.

In engineering, proper documentation of research data is important because experiments are often complex, intricate and elaborate. Inadequate data documentation can lead to misinterpretation of experiments by other researchers and/or unnecessary repetition of experiments that have already been completed, with the data being publicly available in repositories. High quality data documentation is essential for researchers seeking to understand the relationships between processes, structures and properties of manufactured components. This is sought and increasingly required by public project sponsors such as the German Research Foundation (Deutsche Forschungsgemeinschaft).

Multi-stage manufacturing in process chains is common for many products. Cross-process data analysis can be used to identify relationships in process chains. This requires the availability of an evaluable, comprehensive and well-documented global dataset [4]–[8]. However, the acquisition of such a dataset across process boundaries is a formidable obstacle due to the different treatment of individual process steps by different partners.

To facilitate cross-platform implementation of AI models, a digitally analysable, system-independent representation of data sets is essential. These datasets can be combined to form a unique dataset representing different system properties, ultimately enabling holistic data-driven modelling, e.g., through multi-task learning or transfer learning. This will enable the harmonisation of workflows across different domains, facilitating communication between domains or between specialists themselves. An overarching strategy is essential to align the different approaches and ensure that experimental data can be reused without modification.

The Multi-Task Learning (MTL) methodology, which is new to materials informatics, can be used, for example, to learn and predict different polymer properties simultaneously, efficiently and effectively [9]. MTL is a machine learning approach in which multiple tasks are trained simultaneously to optimise multiple loss functions simultaneously. Instead of training independent models for each task, a single model is allowed to learn to perform all tasks at once. In the process, the model uses all the available data from the different tasks to learn generalised representations of the data that are useful in multiple contexts [10]. For example, multitask models can be used to overcome data scarcity in polymer datasets. This approach is expected to become the preferred technique for training materials data [9].

Furthermore, existing predictive models are unable to adequately capture the intricate relationships between mechanical characteristics and behaviour in other fields. These studies employed machine learning to predict the mechanical properties of carbon nanotube-reinforced cement composites [11]. The successful training, validation, and testing of machine learning (ML) and deep learning (DL) models necessitate the availability of a substantial amount of relevant data [12].

A survey of data scientists revealed that the majority of their time is spent on data cleaning and organisation (60 %), data collection (19 %), and data mining for patterns (9 %). The process of cleaning and organising data is by far the most time-consuming aspect of the typical data scientist's workflow [13].

The availability of suitable data in materials science has a significant influence on the performance of applied AI models [14], [15]. It thus follows that data management is of particular importance with regard to the usability of AI models. In order to consider and analyse cross-process relationships, it is necessary to have a global view of the dataset in an analysable form. This necessitates the availability of comprehensive documentation that can facilitate the aggregation of data into a unified global dataset.

In conclusion, the principal objective is to facilitate the utilisation of data-driven analysis and modelling, encompassing comparisons across disparate laboratory domains.

C. Challenges

In the following, the main challenges from the researcher's point of view are summarised and presented, based on the conventional setup in which the above-mentioned data analyses are typically carried out.

Chall:Organisations: A number of organisations with diverse scientific and industrial backgrounds, including research institutes and companies, are engaged in collaborative endeavours, typically in the form of joint research projects.

Chall:WorkingCultures: Each domain, and more specifically each organisation, is characterised by a distinctive and unique working culture, encompassing its own technical languages and terminologies. A clear and concise illustration of the objective can be provided by reference to the symbols and units of measurement employed in tensile tests for the purpose of determining tensile strength. The standards diverge with regard to the symbols employed for the tensile strength, with different materials requiring different symbols. The ISO 1920-4 standard for the tensile strength of concrete employs the symbol f_{ct} for tensile strength, while the ISO 6892-1 standard for metals utilises the symbol R_m, the ISO 527-1 standard for plastics employs the symbol σ_m, and the RILEM TC 232-TDT technical guideline for textile-reinforced plastics employs the symbol σ_{cu} for tensile strength. Furthermore, there is considerable variation in the units of measurement that are most frequently employed. These include the megapascal (MPa) and the gigapascal (GPa). The introduction of new terminology is typically met with reluctance by the research community. Nevertheless, as acceptance declines, the potential for errors to occur rises in tandem.

Chall:Acceptance: Existing solution approaches are characterised by a top-down structure, whereby a designated individual or entity is responsible for metadata management within the research network. This results in the creation of a specialised vocabulary, which is then used within the research network. The imposition of a particular domain vocabulary may result in a lack of overall acceptance among the researchers. This is because the domain vocabulary may result in the overwriting of terms that have been established in their respective domains for a considerable period of time.

Chall:MultipleVocabularies: Researchers may be involved in other research projects or networks in addition to the specific research network. It is possible that the aforementioned projects or networks have agreed upon a different standard vocabulary. Consequently, discrepancies may emerge due to the necessity for researchers to frequently transition between these vocabularies.

Chall:TestStandards: The lack of uniform standards within companies or laboratories regarding the testing of material properties or quality characteristics results in the employment of varying testing methods by organisations. It is not uncommon for partners to have diverse backgrounds, encompassing different domains, cultures, and linguistic traditions.

Chall:ProcessChains: A variety of processes are involved in engineering, including planning (which encompasses design, simulated validation, and process programming), production (which includes goods receipt, storage, and manufacturing), and test (which comprises component tests and quality assurance). These processes typically comprise a number of sub-steps and are conducted across a range of stations and organisations.

Chall:TaggingSystems: It is common practice for organisations and research partners to utilise distinct, location-specific tagging systems for the design and identification of physical objects. Subsequently, the research data are stored within decentralised storage systems that are owned by the respective partners.

Chall:DataDocumentation: It is not uncommon for research data to lack adequate documentation, which may be attributed to a lack of structured documentation practices. The use of different notations may present a challenge for researchers from other fields when attempting to comprehend the documentation. Researchers frequently lack clarity regarding the type and manner of information that should be recorded.

Chall:DataManagement: The management of research data is frequently not systematic. To illustrate, there are no established guidelines delineating the requisite organisational measures.

This setup gives rise to a number of requirements, including fundamental requirements pertaining to the Research Data Management (RDM) such as the FAIR data principles [16] (Findable, Accessible, Interoperable, Reusable).

Req:Findability: The data must be readily retrievable and searchable in order to facilitate efficient collaboration on the project (Chall:Organisations).

Req:Accessibility: It is essential to ensure that the data is readily accessible, both internally and externally, in order to facilitate the completion of tasks from various locations (Chall:Organisations).

Req:DataSecurity: It is imperative that robust data security measures be implemented to safeguard research data from unauthorised access. In particular, this necessitates the implementation of a granular role-based access control system and the utilisation of encrypted communication channels for data in transit and at rest within the RDI.

Req:Interoperabibilty: It is essential to ensure the seamless integration of data from disparate sources to establish a unified data repository. This necessitates the establishment of robust data interoperability standards to facilitate the seamless integration of data from diverse processes, involving the participation of multiple organisations and project partners (Chall:Organisations, Chall:ProcessChains).

Req:Reusability: The reusability of data is a crucial aspect to consider, as it allows for the continued utilisation of data throughout the project's lifespan, even beyond its conclusion. One effective approach to achieve this is through data archiving.

Req:WorkingCultures: It is essential that the working cultures of the participating organisations are preserved or integrated in order to guarantee the acceptance of the RDM concept and to prevent the emergence of parallel worlds or laboratory-specific solutions.

Req:Citation: It is necessary to ensure that the data is citable in order to facilitate referencing of published research data. One way of doing this is by using persistent identifiers such as Digital Object Identifiers (DOI).

Moreover, additional requirements pertaining to process chains can be discerned.

Req:Labelling: It is essential that samples are clearly labelled in order to facilitate the merging of data from disparate processes and to enable the tracing of samples throughout the process chain (Chall:ProcessChains, Chall:TaggingSystems).

Req:DataLinking: In order to facilitate the exploration of cross-process interactions, it is necessary to establish a data linkage along the process chain (Chall:ProcessChains).

Req:Workflows: The execution of data-intensive compute workflows is essential for the automation of pre-processing and post-processing tasks related to research data.

One of the most significant challenges in the field of data management is the effective documentation of data across process and laboratory boundaries. The necessity for data that adhere to the tenets of good scientific practice and the FAIR data principles is at the core of these challenges. In the following, a more detailed description of the requirements related to data documentation is provided.

Req:TechnicalLanguage: It is essential that the descriptions are based on a common technical language, as this will ensure that all researchers in the collaborative project, who often come from different specialist disciplines, have the same understanding of the terms used. It is essential to provide an initial explanation of any technical abbreviations and to maintain a clear and objective language throughout. Moreover, the descriptions created should be compatible with existing descriptions from other disciplines so that they can be reused in the long term. It is essential to integrate the technical language and distinctive working culture of each domain while enabling researchers to retain their own linguistic conventions. It is crucial to establish a common technical language that enables researchers from diverse domains to communicate effectively, without the necessity for a uniform, overarching technical language across all organisations. It is not necessary for there to be a uniform overarching technical language across all organisations; local technical terminologies should be compatible. This may lead to an improvement in recognition and a reduction in expenditure. Moreover, it is crucial for interoperability. The authors are unaware of any alternative methods for integrating data records.

Req:ComprehensiveDatadoc It is essential to provide comprehensive documentation and clear explanations of the data, ensuring that technical experts involved in the project and third parties have a comprehensive understanding of the data's meaning, its objectives, and the conditions under which it was generated. In order to document experiments in a comprehensible manner, it is necessary to provide a detailed description of the processes, machines, and materials employed, so that correlations can be established. For example, one might note that material X with property a was produced on machine Y with setting b using process Z.

Req:Completeness: It is similarly vital to guarantee comprehensive reporting. Researchers from a range of disciplines tend to prioritise different quantities according to their specific research questions. This can result in incomplete and inconsistent data documentation across process and laboratory boundaries. It is therefore imperative that complete data documentation is provided in order to facilitate subsequent use of the data by third parties. Adherence to the principles of good scientific practice is also of paramount importance in order to ensure accuracy. In some cases, the passage of time may make it challenging to recall specific details from experiments conducted in the past, particularly given the high turnover rate in the research sector.

Req:Correctness: Errors of a typographical nature in documentation may result in misinterpretations. It is therefore essential that, in addition to the accuracy of the data itself, the documentation pertaining to it is also meticulous and rigorous.

Req:Applicability: The information structure of the data documentation should be designed in such a way that it can be widely applied in the engineering field, rather than being limited to specific processes. The information structure must accommodate the various data types, including text and tabular data, which give rise to different parameters. It is essential that data documentation includes chapters that cover each phase of the data life cycle. Furthermore, the documentation should include details on the lockout and retention periods for both digital and physical items. Furthermore, it is essential to consider the various levels of release and publication, including internal/private, group internal/protected, and worldwide/public, along with the roles involved.

Req:ResearchDataInfrastructure: It is essential that the model employed for the documentation of data is compatible with the majority of RDM systems, particularly those that adhere to the widely acknowledged FAIR data principles. A systematic approach to research data management, with a particular focus on the clarity and longevity of the documentation produced, is an essential component of the process of creating high-quality documentation.

Req:Compatibility: It is imperative to guarantee compatibility and interoperability with data repositories for the purposes of archiving and publishing. Furthermore, the model should facilitate the integration of data from disparate processes within a process chain. It is necessary to expand the resulting dataset with sections based on the process chain stations. It is imperative that the documentation clearly delineates the requisite licensing for the publication of the data.

Req:Usability: The data documentation processing should be designed in a manner that is supportive and practical for researchers, without imposing an additional burden on them. It is crucial to determine the extent of data publication, encompassing the scope (e.g., whether public or restricted to project collaborators), the degree of data publication (whether partial or comprehensive), and the necessity for data anonymisation. Furthermore, additional integrability requirements must be taken into account.

Req:Integrability: It is essential that the researcher be able to integrate their local data documentation into the overall data documentation with minimal effort. This integration must accommodate different forms of local documentation.

In conclusion, it is recommended that datasets from different organisations be merged, for example, for the purpose of conducting round-robin tests. In the case of multi-stage process chains, this approach allows for the identification of overarching correlations within the overall dataset. In accordance with the FAIR data principles, third-party researchers will be able to comprehend and examine datasets from disciplines with which they are unfamiliar, with a view to answering their own research questions.

D. Aim

The objective of this study is to develop a practical methodology for the synchronised documentation of research data within the engineering domain across various phases. This approach will enable researchers to maintain their perspectives during data preparation and documentation while ensuring compliance with the FAIR data principles. It is anticipated that the methodology will be achievable, extensible and effective in promoting cross-platform functionality. The deployment of AI models is then facilitated through the presentation of training datasets in a digitally analysable, system-independent format, thereby enabling cross-process data analysis.

E. Outline

The remainder of the contribution is organised as follows: Section II provides an overview of existing work related to the described problem. A description of the proposed strategy is provided in Section III. In Section IV, the feasibility of the proposed strategy is demonstrated by applying it to two joint research projects in the engineering domain. The main results are presented and discussed in Section V. Section VI summarises the contribution and draws perspectives for future work.

II. State of the Art and Related Work

This section provides an overview of the current state of the art and related work regarding the documentation of research data within the engineering domain across various phases. In order to achieve this, an overview of existing infrastructures for research data management is provided in Subsection II-A, while an overview of existing metadata schemes for research data and research software is given in Subsection II-B. The current state of knowledge representation is outlined in Subsection II-C. Finally, an overview of ML methods is provided, with a particular focus on their usability and the availability of data in an analysable format. This is presented in Subsection II-D.

A. Research Data Infrastructures

It is imperative that the infrastructure that supports the reuse of research data is continuously enhanced. In order to address this issue, the FAIR data principles have been developed [16]. These foundational principles serve as a set of guidelines for those seeking to enhance 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. Furthermore, these principles can be utilised by professional data publishers who offer their services and expertise in this domain.

A number of research data infrastructures (RDI) have been developed for use in collaborative engineering science projects. One such example is the Karlsruhe Digital Infrastructure for Materials Science (Kadi4Mat), which has been the subject of a recent review [17]. The software boasts a plethora of features designed to facilitate data management and collaborative work in joint projects. These include web-based access, fine-grained role management, the creation of reproducible workflows, and the publication of research data. The aforementioned basic functionality can be readily augmented through the use of plug-ins.

It is, however, important to note that their value extends beyond this. The sharing and collaboration of data are fundamental aspects of scientific research. It is incumbent upon researchers to engage in the processes of data sharing and collaboration in order to expand their collective knowledge and perspectives. It is imperative that researchers rely on each other without bias in their data and interpretations. Nevertheless, researchers are obliged to maintain objectivity and balance when utilising technical terminology and adhering to the established conventions of academic discourse. This should be applied not only to data in the traditional sense but also to the algorithms, tools, and workflows that produce it. The FAIR data principles place particular emphasis on the importance of fairness, which applies to both human and machine activities. Effective data management is not an end in itself; rather, it is a means to an end, facilitating knowledge discovery and innovation. Furthermore, it enables the subsequent integration and reuse of data and knowledge by the research community after data publication [16].

B. Metadata schemes

A systematic documentation of research data is a fundamental requirement, as outlined in “Req:Technical Language”. The documentation of research data is typically accomplished through the use of semantic annotations, as exemplified by Fensel [18]. This process entails the enrichment of the data through the incorporation of a set of metadata derived from a predefined vocabulary. In order to facilitate the collection of data in a machine-readable format, metadata schemas are frequently employed [19]. It is therefore necessary to reiterate the definitions of key terms such as ’metadata’ and ’metadata schema’ as presented in the existing literature.

Definition 1 (Metadata [20]) *“Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource, especially in a distributed network environment like for example the internet or an organization.”*

Definition 2 (Static and dynamic metadata) *Metadata that are not dependent on the underlying process but are of a general nature are referred to as* static metadata*. Metadata that are dependent on the underlying process and thus process-specific are referred to as* dynamic metadata*.*

Example 3 (Static and dynamic metadata) *Example of static metadata include the name of the data provider and the date of creation of the research data. Examples of dynamic metadata include setup parameters, such as the laser velocity during additive manufacturing processes, or characteristics, such as the ultimate tensile strength in tensile tests.*

Definition 4 (Data model [21]) *A model which organises the constituent data elements and establishes standardised relationships between them is a* data model (DM)*.*

Definition 5 (Metadata model) *A data model for managing the metadata of research data originated from a given process is a* metadata model (MDM)*.*

Definition 6 (Metadata schema) *“Metadata elements grouped into sets designed for a specific purpose [...] are called* metadata schemas (MDS)*.”*

A plethora of universal metadata schemes are available for consultation in the literature. Metadata schemes have been employed in the context of online retail for a period exceeding ten years. As illustrated in [22], there are in excess of one hundred metadata schemes. A recent summary of existing approaches to metadata schemes for research software is provided by [23]. The summary includes a number of commonly used metadata schemas related to the engineering domain, including DataCite [24], CodeMeta [25], EngMeta [26], [27].

In order to address the ever-increasing amount of digital research data, the DataCite international consortium was established in late 2009. The consortium’s objectives include promoting the acceptance of research data, with a view to facilitating 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 ontology. A number of crosswalk solutions for integrating with alternative metadata schemas are currently available. CodeMeta comprises a number of elements, some of which concentrate on technical aspects, such as file size or the operating systems that the used research software is compatible with, while others include administrative information, such as the licence under which the data is distributed. The metadata standard does not comprise any mandatory elements. It enables using uniform resource identities (URIs) to identify authors, contributors, and licenses. The content-specific metadata is restricted to the categories of application and keywords.

EngMeta is an XML-based formal definition of the information required to locate, comprehend, reproduce, and reuse data from engineering disciplines [26]. It employs a metadata schema for the description of engineering research data and the documentation of the entire research process, including the individuals involved, software, instruments, and computing environment, as well as the methods used and their parameters [27], [28].

Nevertheless, the number of metadata templates for particular experiments is insufficient. Even in experiments that are standardised in accordance with the German industry standard (DIN), there is no guidance available regarding the type of metadata that should be stored. The standards concentrate on the methodology of the experiments, rather than on the management of the data collected during the course of the experiments. This emphasises the necessity for the expansion of metadata schemes within the engineering domain.

C. Technical languages

In the following, an overview of the various technical languages and their representations within the engineering domain is presented.

Generally, it can be postulated that knowledge can be represented as a generic network, comprising nodes and links. In most cases, nodes are used to define concepts, which are sets or classes of individual objects, while links between the nodes are used to define relationships between them. In certain instances, more complex relationships are themselves represented as nodes, and are distinguished from nodes representing concepts with great care. Furthermore, concepts may possess elementary properties, frequently designated as attributes, which are typically associated with the corresponding nodes [29].

The representation of knowledge is illustrated by a simple example in Figure 1. Please refer to reference [30] for further details. The network represents knowledge concerning destructive testing methods, tensile tests, fracture limits, and other related subjects. The relationship between tensile tests and destructive testing methods can be defined as follows: "tensile tests are destructive testing methods." This is sometimes referred to as an "IS-A" relationship, whereby tensile tests are classified as "IS-A" destructive testing methods. The IS-A relationship establishes a hierarchy among the concepts and provides the foundation for the "inheritance of properties." In the event that a given concept is more specific than any other, it will inherit the properties of the more general concept. Moreover, concepts may possess elementary properties, frequently designated as attributes, which are customarily linked to the corresponding nodes. To illustrate, a tensile test is defined in terms of its test speed [29].

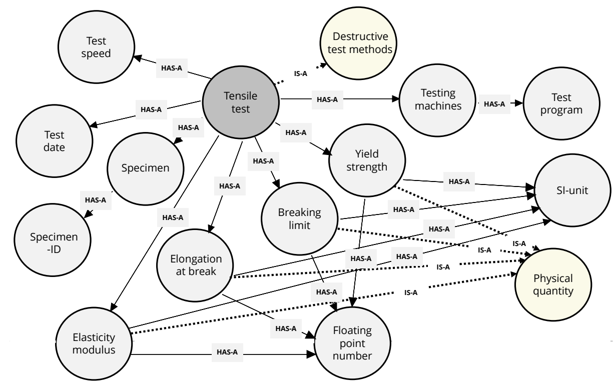

Figure 1. Symbolic representation of an excerpt from an ontology concerning destructive testing methods, exemplifying the “IS-A” and “HAS-A” relationship [30].

Description logics (DLs) are regarded as the fundamental building blocks of knowledge representation systems, constituting subsets of first-order logic. They concentrate on a language comprising unary and binary predicates. Description logics (DLs) constitute a family of languages designed to represent conceptual knowledge in a formal way as a set of ontological axioms. DLs provide a formal foundation for the ontology language OWL, which is a W3C-standardised language for representing information in web applications [31].

A number of abstract definitions of the terms "taxonomy" and "ontology" exist. The data model proposed in the underlying contribution is centred on the terminology employed during the documentation process, with a particular focus on the technical terms associated with these structures. Thus, the definitions provided in 7 and 8 are sufficient for describing the solution concept.

Definition 7 (Taxonomy) *A systematic classification of terms, organised in a hierarchical structure reflecting their interrelationships, is referred to as a* taxonomy.

Definition 8 (Ontology) *A formally organised representation of sets of terms, properties, and relationships between terms is an* ontology *[32]. If an ontology contains technical terms that can be used across different domains, then the ontology is an* upper ontology. *If an ontology contains technical terms that are domain-specific and thus, can be only used within this domain, then the ontology is a* domain ontology.

Example 9 (Domain ontologies) *Examples of domain ontologies include EMMO [33] and Materials Design Ontology (MDO) [34]. Further examples can be found in the BIG-MAP Project [35].*

Ontologies possess a number of advantages over relational and object models. They permit the establishment of precise definitions of conceptual schemas and facilitate the interpretation of data semantics by systems [36].

An ontology is a formal description of the structure of data, including the classes, properties, and relationships that are characteristic of a particular field of knowledge. It serves to ensure the consistency and understanding of the data model. Description logics offer a fundamental understanding of this family of logics, which has become a crucial formal basis for contemporary applications in recent years. An established ontology description language for modelling ontologies is OWL [37], [38].

In general, the greater the precision with which data documentation models the specialist area, the more suitable it is. This indicates that general ontologies, on which knowledge databases such as WikiData [39] or DBpedia [40] are based, are only applicable to a limited extent in highly specialised fields of application such as additive manufacturing.

The European Materials Modelling Ontology (EMMO) [33] is a standardisation approach for technical terms in the applied sciences, with particular relevance to materials science. It can be employed for the modelling of experiments and simulations. EMMO serves to establish a connection between the physical world, the domain of material characterisation, and the realm of material modelling. The ontology is founded upon the principles of physics, analytical philosophy, and information and communication technologies. The field of materials science gave rise to the development of EMMO with the objective of establishing a coherent framework for the capture and organisation of knowledge, one that is aligned with the established scientific principles and methodologies [33], [41].

The ontology OntoSoft [42] is designed to capture scientific software metadata and expand it with machine-readable descriptions of the expected content of the inputs and outputs of software. The ontology EDAM contributes to the advancement of open science by facilitating the semantic annotation of processed data, thereby enhancing its intelligibility, discoverability, and comparability [43]. The Software Ontology (SWO) [44] has been developed to extend the EDAM ontology in order to describe software within the context of this research area [45]. SWO incorporates licensing, programming languages, and data format taxonomies. In contrast to OntoSoft, the utilisation of taxonomies enhances the usability of semantic web applications and facilitates interlinking, as evidenced by Lamprecht [46].

The manufacturing industry is undergoing a period of rapid evolution, characterised by increasing complexity, interconnectivity and geographical dispersion. The intensifying competitive pressures and the growing diversity of consumer demand are compelling manufacturing companies to place an increasing emphasis on the implementation of enhanced knowledge management practices. In response to this challenge, the Additive Manufacturing Ontology (AMO) [47] has been developed. This ontology is designed to represent the additive manufacturing product life cycle.

In Mayerhofer [48], a knowledge-based framework is presented that can be used to automatically analyse the geometric properties of components and compare them with the guidelines for additive manufacturing. The knowledge base is founded upon an ontology that delineates processes, printers, and materials pertinent to the domain of additive manufacturing. In Ali [47], an ontology for the description of manufacturing processes utilising additive manufacturing in dentistry is presented.

The heterogeneity of energy ontologies presents a significant challenge to the interoperability of ontology-based energy management applications for large-scale energy management. One potential solution to this challenge is a global energy ontology, namely the Domain Analysis-Based Global Energy Ontology (DABGEO) [49]. This ontology provides a balance between reusability and usability, with the aim of reducing the effort required for reuse in different applications. Conversely, the ontology "Ontology for Energy Management Applications" (OEMA) [50] represents an endeavour to unify existing heterogeneous ontologies that represent energy performance and contextual data. Moreover, the "Open Energy Ontology" (OEO) has been introduced as an ontology for energy systems analysis (Neuhaus [51], Li [52]). OKG-Soft [53] is a framework designed to facilitate the capture and publication of machine-readable software metadata. It builds upon the OntoSoft platform.

There are information systems which facilitate data exchange and retrieval based on an appropriate ontology and given data sources, see Definition 10.

Definition 10 (Ontology-based data integr. system [54])
An information management system comprising an ontology, a set of data sources, and the mapping between the two as components is an ontology-based data integration (OBDI) system.

The reuse of a common vocabulary and incorporation of mappings between the ontology and data sources facilitate data exchange and retrieval. In the event that the organisation in question possesses an appropriate ontology, the OBDI system is capable of facilitating the integration and sharing of data that was previously stored in disparate, heterogeneous databases.

Taxonomies and ontologies are indispensable instruments for researchers seeking to comprehend and retrieve extensive collections of scientific and engineering data. Nevertheless, the management and application of ontologies themselves can prove challenging. While both ontologies and taxonomies serve similar functions, they differ in terms of their complexity. Taxonomies are characterised by a hierarchical structure and the exclusive use of parent-child relationships, whereas ontologies are considerably more complex [55], [56]. In essence, an ontology represents a structured and formalised body of knowledge pertaining to a specific domain. The semantic system employs transparent and intelligible representations of concepts, relationships, and rules to cultivate this knowledge. It is not feasible to rely exclusively on the expertise of database programmers or data engineers to develop a system that considers target applications, such as materials or production technologies. Such tools lack the domain-specific knowledge that is essential for characterising associations between concepts. It is therefore essential to seek guidance from a number of domain experts in order to acquire the requisite domain knowledge [57].

D. Usability in Machine Learning

Over the past decade, machine learning (ML) has emerged as a prominent field of study within the discipline of materials engineering. ML constitutes a subset of the broader category of artificial intelligence (AI), which encompasses the development of algorithms and models that enable systems to learn and improve from data without explicit programming. AI comprises a broader range of technologies integrated into a system that aims to facilitate reasoning, learning, and problem-solving in order to address complex problems.

ML algorithms analyse vast amounts of data, extracting insights that inform decision-making processes [58]. This is exemplified by the work of Cloud and Google, who have developed an AI system that can detect and extrapolate patterns from data sets. The popularity ML is growing worldwide, driven by an increasing demand for data analysis solutions. This is evidenced by [14]. However, the necessity for large amounts of data may present a challenge in many areas, particularly given the requirement for sophisticated large-scale laboratory tests. The application of ML methodologies in materials science research has increased in recent times, as evidenced by [15]. It is evident from research findings that the limited practicality of AI in certain domain-specific contexts, partly due to the necessity for comprehensive laboratory tests on a large scale, represents a significant challenge to its implementation.

A novel approach to the applicability of AI techniques, termed Usable Artificial Intelligence (UAI), has been developed with a focus on industrial requirements. This approach is outlined in reference to the work of Wiemer et al. [59]. Despite the considerable progress that has been made in the development of data-driven, machine learning, and artificial intelligence methods, these techniques are not yet fully utilised to address the associated technical challenges, particularly in industrial applications. This is primarily attributable to the restricted practical applicability of AI solutions. It is often the case that technical practitioners depend on collaboration with data science specialists in order to fully exploit the capabilities of AI methods [60]. In this work, a flexible, tractable, scalable, and adaptable technique for constructing anticipatory models has been introduced and demonstrated on two use cases.

E. Own contribution

The novelty of this approach lies in its ability to shape a metadata model that is harmonised, platform-independent and facilitates the reuse of high-quality datasets across laboratory, team and process boundaries. Furthermore, the metadata schema supports the applicability of AI/ML methods, making this subject highly relevant to the ever-increasing need for digitalisation.

III. Solution concept

The following section proposes a solution concept based on a collaborative metadata management approach. The approach allows researchers to focus on data provision by reducing

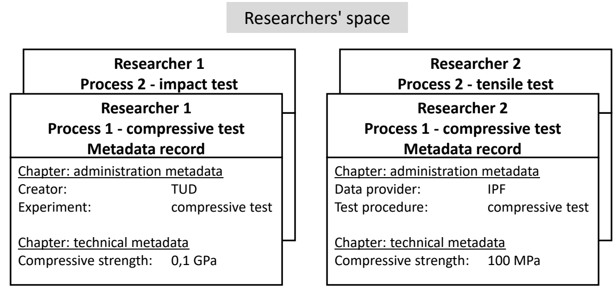

Figure 2. Typical metadata management in case research data for different processes are documented by different researchers / research groups illustrated for a simple example.

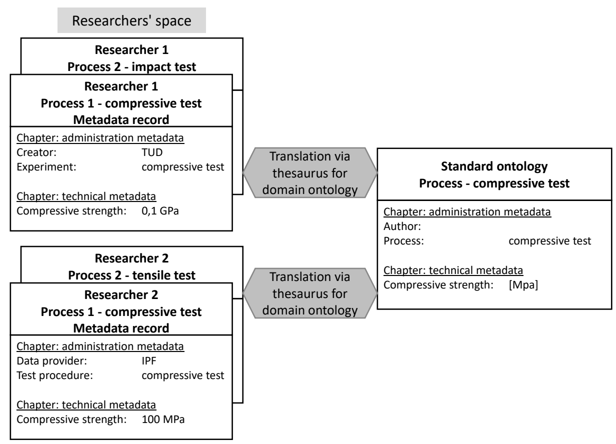

Figure 3. Metadata management of research data using a thesaurus as translation layer illustrated for a simple example.

routine activities rather than aligning with similar research groups. This means, it enables researchers to concentrate on their experiments and research questions. An initial solution concept based on a translation of the ontology has already been proposed in reference [60].

As previously stated, a significant challenge inherent to metadata management is the existence of disparate representations of the same metadata. This challenge will be illustrated by a simplified example. The objective is to annotate a tensile test by recording the name of the experiment, the name of the data provider, and the value of the compressive strength. As illustrated in Figure 2, researchers from disparate organisations employ disparate terminologies for the representation of the experiment name (e.g., "Process," "Experiment," "Test procedure") and the data provider name (e.g., "Author," "Creator," "Data provider").

The core concept of the proposed solution to this challenge is the implementation of a translation layer based on a thesaurus. This enables researchers to utilise their own terminology. Figure 3 illustrates the fundamental concept illustrated in the aforementioned example. As illustrated, the terms "Creator" and "Data provider" are translated via a thesaurus to yield the term "Author," which represents the author's name. Similarly, the terms "Experiment" and "Test procedure" are translated via a thesaurus to yield the term "Process," which represents the

TABLE I. TERMS WITH DESCRIPTIONS AND SYNONYMS FOR THE GIVEN EXAMPLE.

Term	Description	Synonyms
Author	Person who has created the dataset	Creator, Data Provider
Process	Name of experiment / process	Experiment, Test procedure

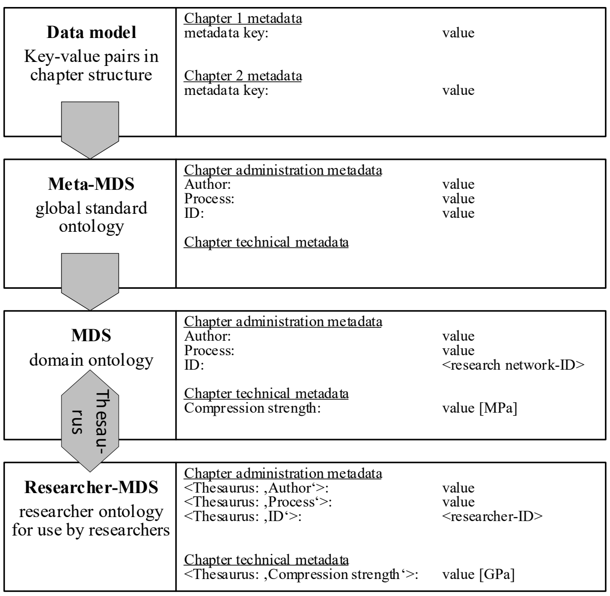

Figure 4. Workflow for the metadata management within research networks.

name of the experiment. Table I illustrates the mapping of the terms and synonyms employed in the aforementioned example.

In the following, the various components of the proposed model for the documentation of research data in the terminology used by researchers themselves will be outlined at a high level of abstraction.

Model 11 (Documenting data in researchers' lang.) *The model comprises the following components:*

a) *Metadata model: A given metadata model with key-value pairs of the process for which data should be documented. Chapters within the model containing only static metadata have the same keys for all processes and are mandatory. Chapters within the model containing only dynamic metadata must be redefined for each process. It is essential that the redefinition is conducted in collaboration with domain experts in order to guarantee the reproducibility of the investigation.*

b) *Meta-metadata schema: A meta-metadata schema (MMDS) which is based on the given metadata model. Consequently, the MMDS incorporates the key-value pairs of the metadata model into the identical chapter structure as that specified in the metadata model, including chapters pertaining to administrative, organisational, and technical metadata. The keys within the MMDS are constituted by class names derived from a given global ontology. The values serve to exemplify the potential forms that the keys may assume, such as co-domains for numerical properties. In this manner, the keys and values in the MMDS are populated with specific terms from the global ontology, thereby obtaining concrete values.*

c) *Metadata schema: A metadata schema (MDS) which is based on the MMDS. Similarly, the MDS incorporates the key-value pairs of the MMDS into the same chapter structure as that defined in the MMDS. The key names within the MDS are drawn from the class names of the given domain ontology. The values indicate the potential forms of the keys.*

d) *Researcher MDS: A researcher metadata schema (researcher MDS) which is based on the MDS. The researcher MDS is represented in accordance with the local ontology employed by the researcher group. The metadata within the researcher MDS are translated into the domain ontology via the use of local thesauri. The translation process entails the terms associated with the keys and the physical units represented in the values of the key-value pairs within the researcher MDS. Furthermore, identifiers must be translated, as each laboratory or researcher employs a distinct system of identification.*

The model is described in terms of natural language. However, as natural languages are inherently ambiguous and imprecise, the model will be formulated in a mathematical setting, and its feasibility will be demonstrated. In order to facilitate comprehension, it is first necessary to define some basic terms in greater detail.

Definition 12 (Research group, research network) *A group of researchers, not necessarily belonging to the same organisation (e.g., laboratory, research institution) is a* research group. *A research network (RN) is a set of research groups working on the same research project.*

Definition 13 (Alphabet) *Let Σ be a finite, non-empty set of distinguishable symbols. Then the set Σ is an* alphabet. *The symbols of the alphabet are also named characters or letters.*

Definition 14 (Technical term) *Given a domain, let Σ be an alphabet, let $a \in \Sigma^m$ be a string containing $m \in \mathbb{N}$ characters forming a string which has a special meaning in the given domain. Then a is a* technical term.

Definition 15 (Synonym of technical term) *let Σ be an alphabet, Let $a \in \Sigma^n$, $b \in \Sigma^m$ with $n, m \in \mathbb{N}$ be two technical terms. If b has the same meaning as a, then b is a* synonym of *a, denoted by synonym(a, b).*

Definition 16 (Thesaurus) *Let Σ be an alphabet, $A, B \subseteq \mathcal{P}(\Sigma)$ be given sets, A contains $p \in \mathbb{N}$ technical terms, B contains $q \in \mathbb{N}$ technical terms. Let T be a set of ordered pairs, defined as follows:*

$$\mathcal{T} := \{(x; y), x \in A, y \in B \,|\, \textit{synonym}(x, y)\}. \tag{1}$$

Then the set T is a thesaurus related to the sets *A* and *B.*

Assumption 17 (Uniqueness of terms) *It is assumed that all terms employed in the technical language are pairwise unique.*

Remark 18 (Uniqueness of terms and main term) *It follows from assumption 17, that is always possible to uniquely assign synonyms to a main technical term.*

Definition 19 (Set of synonyms and main term) *Let Σ be an alphabet, $\mathcal{P}(\Sigma)$ the Kleene closure of Σ, A, $B \subseteq \mathcal{P}(\Sigma)$ be given sets, A contains $p \in \mathbb{N}$ technical terms, B contains $q \in \mathbb{N}$ technical terms, let T be a thesaurus related to the sets A and B. Then the* set of synonyms related to the thesaurus *T for a given term $x \in A$ is denoted by synonyms$(x; T) \subseteq B$, and the* main technical term related to the thesaurus *T for a given term $y \in B$ is denoted by main$(y; T) \in A$.*

Theorem 20 (Compatibility researcher MDS with MDM) *Given the following components as input:*

- *One research network with $m \in \mathbb{N}$ research groups G_j, $j = 1, 2, \ldots, m$,*
- *$n \in \mathbb{N}$ processes P_i, $i = 1, 2, \ldots, n$,*
- *One global metadata model per process P_i with $p \in \mathbb{N}$ key-value-type triples $(k_l; v_l; d_l)$, $l = 1, 2, \ldots, p$ with keys k_l, corresponding values v_l and data types d_l representing the metadata of the model. The triples of the model are ordered in such a way that there are a chapter with only static metadata and chapters with dynamic metadata.*
- *One global ontology with a set of technical terms given by $\mathcal{N}_G$,*
- *One domain ontology with a set of technical terms given by $\mathcal{N}_D$,*
- *One local ontology per research group G_j, $j = 1, 2, \ldots, m$, with set of technical terms given by $\mathcal{N}_{L_j}$,*
- *One global thesaurus $\mathcal{T}$ related to the sets $\mathcal{N}_L$ and $\mathcal{N}_D$ with $\mathcal{N}_L := \cup_{j=1}^{m} \mathcal{N}_{L_j}$, and*
- *One local thesaurus $\mathcal{S}_j$ per research group G_j, $j = 1, 2, \ldots, m$, related to the sets $\mathcal{N}_{L_j;syn}$ and $\mathcal{N}_{L_j}$ with $\mathcal{N}_{L_j;syn}$ a set of technical terms containing synonyms of the technical terms in $\mathcal{N}_{L_j}$.*
- *One researcher MDS per process P_i with $p \in \mathbb{N}$ triples $(\tilde{\tilde{k}}_{l,j}; \tilde{\tilde{v}}_{l,j}; d_l)$, $l = 1, 2, \ldots, p$ representing the metadata of the model. The triples of the researcher MDS are ordered in the same way as the triples of the metadata model.*

Then each research group G_j can use the technical terms of their own local ontology L_j given by the sets $\mathcal{N}_{L_j}$ and $\mathcal{N}_{L_j;syn}$ for documenting research data via the given researcher MDS while being compatible with the global metadata model for the process P_i.

Proof: It has to be shown that the keys k_l of the metadata model are identical to the keys that will be be obtained when applying the given domain ontology and the given local ontologies to the keys $\tilde{\tilde{k}}_{l,j}$.
On the one hand, the keys k_l are the technical terms of the given domain ontology by definition. Thus, it be can concluded that $k_l \in \mathcal{N}_D$.
On the other hand, applying the local thesaurus in Eq. (2) to the keys $\tilde{\tilde{k}}_{l,j}$ yields the keys of the local ontologies (set $\mathcal{N}_{L_j}$).

$$\hat{k}_{l,j} := \text{main}(\tilde{\tilde{k}}_{l,j}; L_j) \qquad \text{for } l = 1, 2, \ldots, p. \tag{2}$$

Thus, it be can concluded that $\hat{k}_{l,j} \in \mathcal{N}_{L_j}$. Applying the global thesaurus in Eq. (3) to the keys $\hat{k}_{l,j}$ yields the keys of the domain ontology (set $\mathcal{N}_D$).

$$\hat{k}_l := \text{main}(\hat{k}_{l,j}; T) \qquad \text{for } l = 1, 2, \ldots, p. \tag{3}$$

Thus, $\hat{k}_l \in \mathcal{N}_D$.
In conclusion, the keys $\hat{k}_l$ are identical to the keys k_l. $\square$

Example 21 *The result of the Theorem 20 will be illustrated by a simple example. Given a process P_i, let $\mathcal{N}_D =$ „author", $\mathcal{N}_{L_1} = \{$„data provider"$\}$, $\mathcal{N}_{L_2} = \{$„creator"$\}$, $\mathcal{N}_{L_1,syn} = \{$„author", „data recorder"$\}$, $\mathcal{N}_{L_2,syn} = \{$„author", „data creator"$\}$, $\tilde{\tilde{k}}_{1,1} =$ „data recorder", $\tilde{\tilde{k}}_{1,2} =$ „data creator". Then, applying the local and global thesauri yields:*

$$\hat{k}_{1,1} = S_1(\tilde{\tilde{k}}_{1,1}) = S_1(\text{„data recorder"}) = \text{„data provider"},$$
$$\hat{k}_{1,2} = S_2(\tilde{\tilde{k}}_{1,2}) = S_2(\text{„data creator"}) = \text{„data provider"}.$$

Figure 4 illustrates the aforementioned workflow in great detail, based on the proposed metadata model.

It is essential that the translation layer is in place for each process for which research data should be documented.

The proposed model permits researchers to view the metadata schema in their native languages. This implies that each researcher is able to continue utilising their customary ontology. Nevertheless, there are still higher-ranking, standardised ontologies that it is imperative that all parties are able to comprehend. In the event that research data is to be shared with other research groups, the metadata must be translated back into the global ontology, thus ensuring that it can be understood by other researchers.

In conclusion, metadata schemas grounded on a domain-specific ontology and a thesaurus guarantee that the corresponding research data is understandable, usable for further analyses, interoperable across laboratory boundaries, replicable at a qualitative level, complete, and of superior quality.

IV. USE CASES

The following section will illustrate the feasibility of the proposed metadata model, as outlined in Section III. For this, it is applied within the research data management of the following research projects:

A. An interdisciplinary research training group comprising over 30 researchers from 13 different research institutions, investigating mineral-bonded composites with the objective of improving structural impact safety,
B. a joint research project in the engineering domain, involving over 20 researchers from six different research institutions, investigating process-structure-property relationships for additively manufactured components.

This section is organised as follows: Subsection IV-A will provide a comprehensive account of the use case pertaining to the interdisciplinary research training group. The subsection presents the solution concept employed for the RDM and demonstrates how the proposed metadata model facilitates the collection of research data. Subsequently in Subsection IV-B, the same approach is applied to the second use case, which pertains to the joint research project in the engineering domain.

A. Use case GRK 2250

The GRK 2250 [61] is an interdisciplinary research training group, or in German, a "Graduiertenkolleg." It is comprised of over 30 researchers from 13 different research institutions. The primary objective of the GRK is to investigate mineral-bonded composites with the aim of enhancing structural impact safety.

This investigation is conducted through the utilisation of experimental-numerical and data-driven methodologies, as evidenced by the reference [59]. The research encompasses a range of scales, from the microscale to the structural scale. At the microscale, investigations include fibre pull-out tests and the corresponding simulations. At the structural scale, investigations include drop tower tests (utilising a 10-metre drop tower with plates measuring 1.5 metres by 1.5 metres by 30 centimetres) and corresponding simulations. The research team comprises researchers from nine different departments from four faculties at TU Dresden and the Leibniz Institute of Polymer Research (IPF Dresden). The GRK 2250 is structured in three successive cohorts, each comprising three years of study. The research domains represented in each cohort include textile technology, polymer and material sciences, construction materials, structural engineering, continuum mechanics, numerical modelling, 3D optical monitoring techniques, sustainability, resilience, and machine learning. These domains are represented by more than 10 researchers from a variety of academic backgrounds. Each cohort is typically addressed by a distinct team of researchers. It is therefore of great importance that a systematic approach to research data management is employed in order to ensure the success of the project.

The quantity of data varies significantly, with a range extending from a few megabytes to several hundred megabytes per experiment. A total of 3 terabytes of cumulative data has been stored on 15 test systems at six different laboratories. The number of experiments conducted on each test system ranges from 20 to 300.

In the following, the RDM concept is presented as it is used in the GRK 2250. The objective is to establish a methodology for linking data along process chains and for documenting process data in a sustainable manner. The implementation of the RDM concept within the GRK 2250 will be outlined, with an illustrative example provided to demonstrate the documentation of data from tensile tests.

1) Research data infrastructure: As illustrated in Figure 5, the research data infrastructure (RDI) comprises a shared drive and an RDM frontend, accessible to all project partners. The frontend provides the functionality of data search and access, visualisation and analysis. Furthermore, researchers are able to upload data directly via the frontend and create metadata for each dataset.

2) Data documentation: As illustrated in Figure 6, the RDM workflow offers two data storage options. The first option is to manually save the research data in the dedicated folders within the group drive. The second option is to upload the data via the RDM frontend, whereby the files are saved in the specified folder on the group drive. In the initial phase, the data can be stored in one of two ways. For the documentation of data, templates for the annotation in terms of CSV and Excel spreadsheets are provided for all investigations. The template for compression tests is illustrated schematically in Figure 7, which also contains an abridged version of the chapter pertaining to technical metadata. The templates have been designed in such a way that metadata can be recorded in a standardised manner with minimal effort. The Excel spreadsheet templates include a number of helpful features, such as automatic naming upon saving, the ability to automatically add a new template with the same keys, and predefined lists for metadata values. Metadata files are linked to the corresponding research data files via an identifier in the filename. The format of the research data files is unrestricted.

Two options exist for the documentation of research data via metadata.

1) In the initial option, the researchers input the metadata manually into an Excel spreadsheet, which is subsequently stored in the corresponding research data folder on the shared drive. An alternative approach is to upload the data together with the metadata via the frontend, where the files are saved in the selected folder.
2) In the second option, which is represented by step 2 in Figure 6, the metadata file, which is based on the existing template, can be created via the frontend and saved in the same folder as the corresponding research data as a CSV file. Furthermore, research data can be subjected to analysis and visualisation in the frontend, provided that they are stored in the predefined file formats CSV and XLSX. Researchers have also the option of integrating self-written Python scripts into the frontend, thus enabling the support of additional file formats.

Although the Excel-based approach is relatively straightforward and readily implementable, it also presents certain disadvantages in terms of usability. The manual input of data documentation is a time-consuming and error-prone process.

The data obtained from each process must be exported in the appropriate file format and manually merged into a global data set for further data analysis.

One potential solution is the utilisation of a modelling tool based on a specified process data model. This enables the straightforward integration of processes and the selection of pertinent metadata for processes through a graphical user interface. The process data model comprises a variety of block types for the representation of materials, processes, devices, and experiments. A set of metadata can be associated with each block, thereby forming a metadata schema. The primary metadata structure is consistent across all blocks. The addition

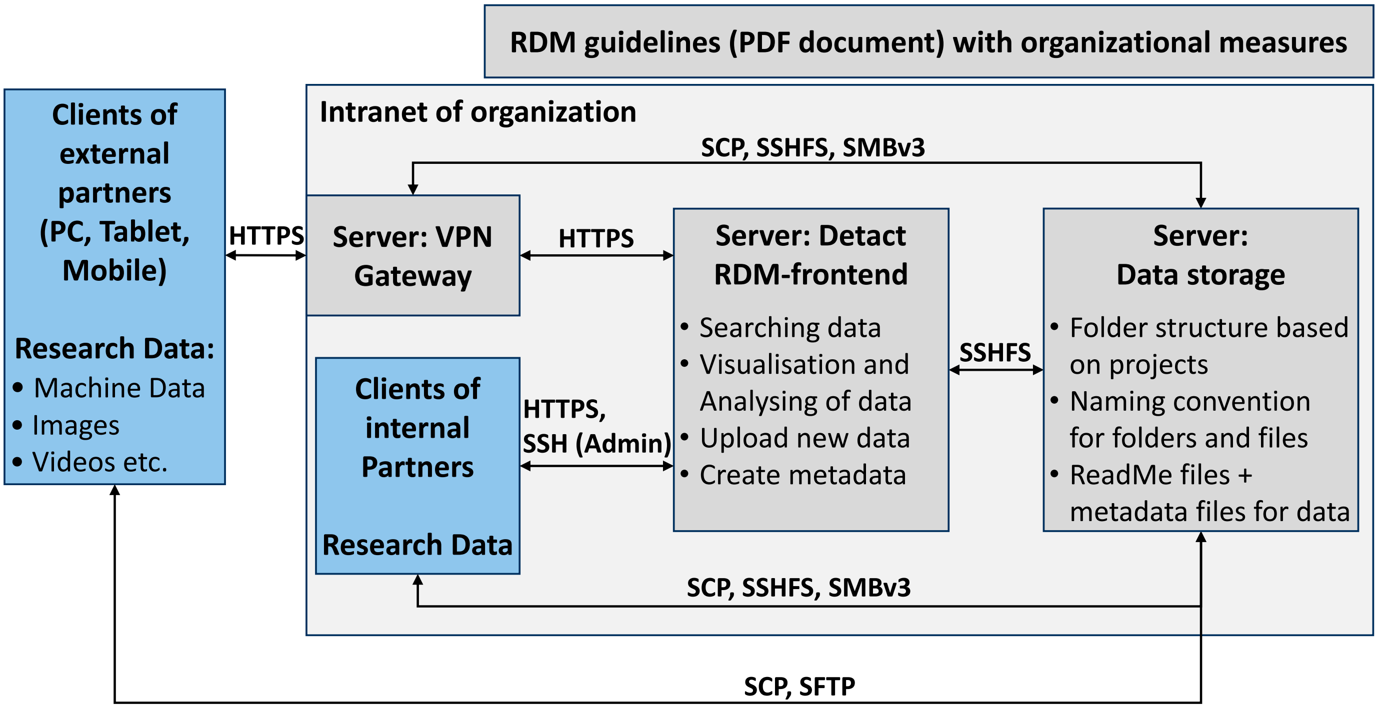

Figure 5. RDI within the solution concept for the research data management in GRK 2250.

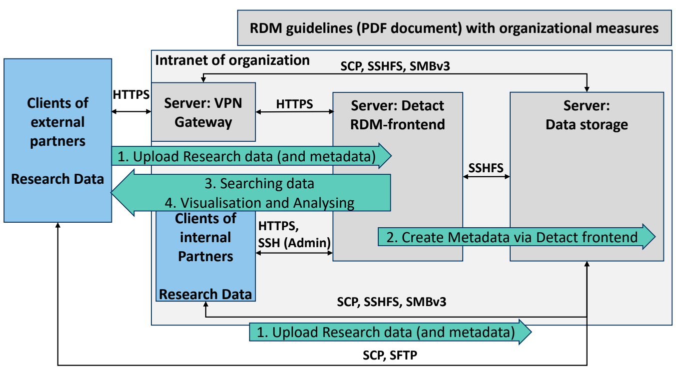

Figure 6. RDM workflow for storing and documenting research data within the solution concept used at the GRK 2250.

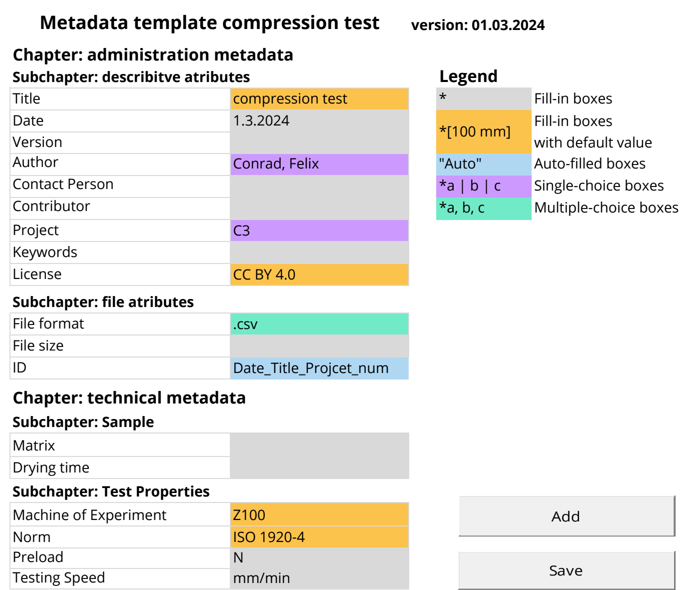

Figure 7. Metadata template for compression tests in GRK 2250.

of metadata keys to a block can be accomplished by utilising a predefined library, designated as the "metadata library." Figure 8 shows the basic modules of the process data model.

The process data model facilitates comprehension and adoption of the workflows of other researchers and it allows communication between researchers. The model itself provides the instructions for carrying out the experiment or the entire workflow.

The standardisation of workflows represents a fundamental step towards the establishment of sustainable data management and research practices. A standardised workflow permits the comparison of experiments, thereby facilitating their reuse. At present, each researcher develops their own workflow throughout the course of their research. A significant proportion of experiments remain non-standardised because of the need to develop a novel experimental setup or to produce a new material, the latter of which is not yet standardised. Even minor discrepancies in the manufacturing process can have a significant effect on the target properties. This renders comparison and reuse of data a challenging undertaking. Furthermore, the standardisation of workflows serves to reduce the unnecessary replication of common process steps.

The GRK 2250 is currently engaged in a collaborative endeavour with Symate [62] with the objective of integrating a process modeller within the software Detact [63]. The Detact software is a cloud-based, closed-source program developed and

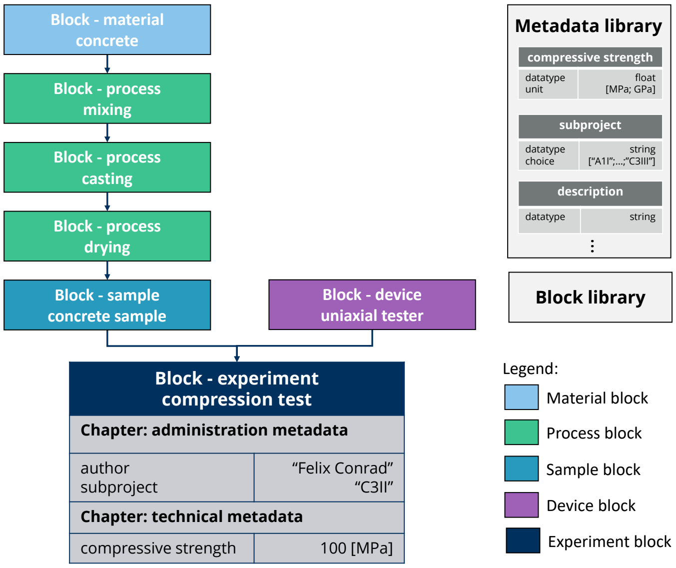

Figure 8. Schematic representation of the process modeller diagram for the use case "uniaxial tensile test of a textile-reinforced concrete test specimen".

maintained by Symate for the purpose of collecting data from a variety of sources along process chains, thereby facilitating the application of automated data analysis.

Figure 8 illustrates the manner in which the aforementioned process modeller can be deployed in the context of a compression test for a concrete test specimen. For the sake of simplicity, only a portion of the total process chain within the GRK 2250 has been selected. The specific material utilised is delineated within the block designated "material concrete." Furthermore, the precise composition will be documented. The mixing of the concrete is defined in the block entitled "process mixing," the casting of the material is defined in the block entitled "process casting," and the subsequent drying of the material is defined in the block entitled "process drying." The aforementioned steps culminate in the formation of the final test specimen, which is represented by the "concrete sample" block. The testing apparatus is represented by the block "uniaxial tester." Subsequently, the test procedure can be recorded within the aforementioned block, designated as "compression test." Again for the sake of simplicity, only the metadata associated with the "compression test" block are illustrated in Figure 8.

It should be noted that the process chain for concrete compression tests is relatively straightforward and concise. In this instance, the metadata may also be incorporated into a unified metadata schema for the specific compression test experiment, obviating the necessity for the process modeller. Nevertheless, process chains in practical applications are frequently more extensive and intricate.

It is imperative that all recorded metadata, as outlined in the metadata schema, align with the FAIR data principles for the research data associated with the process. Attempting to integrate this set of metadata into a single metadata schema without the involvement of a process modeller is likely to result in the omission of crucial parameters. As shown in Figure 8, the drying time of the concrete exerts a profound impact on the strength of the concrete. In the absence of a record of the drying time, the resulting dataset would be essentially unusable for further analysis, as a major influence factor would not have been recorded.

In conclusion, the utilisation of the process modeller as an implementation of the presented process data model offers researchers the following advantages in their day-to-day operations:

- Researchers are not required to peruse lengthy "metadata lists" but may instead utilise the graphical representation of the process model,
- The software can be accessed via a web browser, thereby enabling researchers to record metadata directly in their laboratories.
- A comment function enables researchers to flag instances where deviations from the norm may have occurred, such as noting that the concrete was particularly adhesive today.
- Metadata can be captured and recorded rapidly and in a straightforward manner.
- All individual blocks and overall process data models can be collated and disseminated via an internal Detact library, thus facilitating the standardisation of workflows amongst collaborating researchers. Subsequently, these can be shared in the Detact frontend. The metadata schemas thus created can be exported in the form of CSV files.
- New tests may be initiated with the default values and subsequently modified with minimal effort. This is of great advantage, as it has been shown that the current metadata is insufficient for the purposes of complete capture.
- Perceived costs associated with a comprehensive recording can be significantly reduced.

3) Data processing: One of the research objectives within GRK 2250 is to examine the relationships between the material used, its structure, and its properties. To this end, the same material has been subjected to analysis in a series of tests, including compression and shear tests. The following experiments are detailed in the following sections.

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

Figure 9 depicts the recorded input parameters, designated as "features," and output parameters, designated as "labels," for the specified experiments in a matrix schema. The terminology has been selected as it is typical of the field of machine learning. In the context of a test process, the features represent the set-up parameters, whereas the output parameters correspond to the determined characteristics of the test. Each row represents a discrete test process. The designation of the experiment is indicated in the boxes pertaining to the features, while the

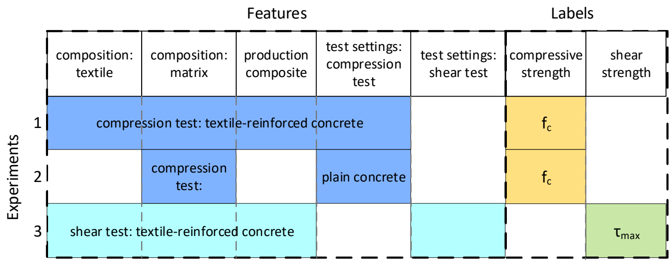

Figure 9. Schematic illustration of an extract of the global data set within GRK 2250. The columns comprise a list of features and labels pertaining to a variety of processes. The rows represent the identifiers of the experiments that have been conducted.

symbols denoting the determined characteristics are displayed in the boxes corresponding to the labels.

The parameters that are available for utilisation in the experiment are indicated by coloured boxes, whereas those that are not available for a particular test are indicated by white boxes. To illustrate, the compression test of unreinforced concrete must furnish data regarding the concrete's composition, including the water content and the quantity of binder. However, data pertaining to the textile reinforcement are not available, as this was not included in the investigation and was therefore outside the scope of the inquiry. The label for this test is the compressive strength, denoted by the symbol f_c. For each test, the sets of features and labels are recorded and stored in a corresponding process-local data set.

- Experiments 1 and 3 employ similar descriptions of the textile-reinforced concrete, but differ in their characterisation of the experimental procedure, reflecting the use of distinct tests to ascertain the compressive and shear strengths. Furthermore, the determination of different properties is indicated herewith.
- Experiments 1 and 2 are identical in terms of the features used to describe the experiment and the label assigned (the material property determined in both cases is the same). However, with regard to Experiment 2, the features pertaining to textile reinforcement are absent, as they are not applicable to the experiment.
- Experiments 2 and 3 employ the same features to describe the concrete matrix. The features are distinct for each experiment, as are the labels. In Experiment 2, the features pertaining to textile reinforcement are absent, as they are not pertinent to the experiment.
- Experiments 2 and 3 employ the same set of characteristics to delineate the concrete matrix. The features are distinct for each test, as indicated by the different labels. In Experiment 2, the features pertaining to textile reinforcement are absent.

The application of the presented process data model enables the consolidation of data sets from disparate experiments into a unified global data set.

The global data set contains a greater quantity of information that can be considered in data-driven models which are capable of simultaneously mapping several material properties. In this manner, the global data set serves as the foundation for the implementation of data-driven models. The creation of a global data set that encompasses the full range of test processes facilitates a comprehensive understanding of the underlying test specimens from a data-driven perspective.

4) Summary and conclusion: This section presented the solution concept which has been used for the management of research data within the GRK 2250. The used approach to metadata management, which involves the use of separate CSV or Excel files, is, in principle, an effective method. However, it is a time-consuming process, which often results in incomplete or inadequate data entry. Consequently, essential details pertaining to the experiment are irretrievably lost.

It is evident that researchers require guidance in the process of recording metadata. It is therefore essential that the tools in question have a low learning curve, given the limited time available for introducing new tools into day-to-day laboratory operation and that these tools are user-friendly, allowing researchers to readily adapt to them.

To address this issue, a novel methodology was proposed, utilising a process data model. Then, the recording of metadata is a relatively straightforward process, which helps researchers to standardise their workflows. This approach is currently being implemented as a prototype.

B. Use case AMTwin

AMTwin is a joint research project in the engineering domain, comprising over 20 researchers from six different research institutions. The project is cited [30], [60], [64]–[66]. The primary objective of the research project is to gain insights into the relationships between processes, structures, and properties of additively manufactured components, with the aim of establishing a systematic knowledge base. The fundamental principle of additive manufacturing (AM) is the construction of three-dimensional geometries through the addition of materials, typically in a layer-by-layer manner [67]. AM offers the potential to produce components in small batches with complex geometry and a high degree of lightweight construction in a flexible manner. The research is focused on components manufactured from Ti6Al4V using Selective Laser Melting (SLM), a specific AM method. The present study employs an experimental-numerical approach to investigate the interactions between the manufacturing process, the material, and the final component properties. A variety of testing processes, including computed tomography, light microscopy, tensile tests, and fatigue tests, were employed to gain insights into the process-structure-property relationships of the manufactured components.

The consistent acquisition of data from materials, processes, and components creates a digital twin, that is to say, a digital image of the AM process that can be used for monitoring and optimisation. The analysis of data that will be available in the future using machine learning methods offers significant potential for innovation in this context. This is achieved by quantitatively mapping the process-structure-property relationships for AM components under static and cyclic load.

In the following, the used RDM concept will be presented. It provides a solution approach to the documentation of data based on ontologies. Subsequently, the implementation of the RDM concept within AMTwin is outlined, with illustrative examples of the documentation of data from tensile tests provided.

1) Research data infrastructure: A solution concept for a practicable RDM in AMTwin has been developed which takes the RDM requirements in Section I-C into account. It is based on the solution concept presented by the authors in [64]. Figure 10 illustrates the architectural solution concept employed in the research data infrastructure (RDI) in great detail. A particular emphasis has been placed on the practicality of the concept, with a specific focus on the rapid provisioning and deployment of the solution.

The RDI within the solution concept for the research data management in AMTwin is comprised of the following components:

ELN: The ELN is a web-based frontend accessible via HTTPS (Hypertext Transfer Protocol Secure), which allows all project partners (both internal and external) to access the main component in a secure manner only through the use of an ordinary web browser. In particular, no additional software installation is necessary to access the ELN. The management of research data is conducted via search masks and filters, while the data documentation is accomplished through the use of forms, both of which are accessible via the web frontend. The ELN provides functionalities for the documentation of research data and the searching of such data, thus ensuring that researchers can easily access the information they require (see Req:Findability). It also provides functionalities for the documentation of research data and the searching of such data, thus ensuring that researchers can easily access the information they require (see Req:ComprehensiveDatadoc).

Central data storage system: The research data are stored on a central data storage system. The system contains file servers that adhere to the traditional file system model, which can be accessed both via the RDM web frontend and directly (see Req:Accessibility). The data store is organised according to a directory structure that reflects the data flows. Access permissions to research data may be set in such a way that project partners are permitted to read and write data from "own" processes (i.e., processes initiated by them) and read data from "external" processes (i.e., processes not initiated by them and not yet subjected to their own processing). The documentation pertaining to the research data is stored in the form of text files within the same folder.

Agent: An agent is connected to the ELN in terms of a web server. The agent facilitates to flexibly customise the RDM, to enhance the functionality of the ELN and thus, to meet project-specific requirements (see Req:Usability). It also guarantees the synchronisation of the data documentation in the ELN and the data storage system. Furthermore, the agent permits the interconnection of IT services that are unable to communicate directly with the RDM web frontend. This way, high-performance computing services can also be utilised on the web frontend, for instance, to undertake computationally demanding data processing operations.

Server with data repositories: It is possible to archive released process data sets within a data repository. Furthermore, data sets can be published using a persistent identifier via the attached publication service (see Req:Citation).

Compute server: The ELN enables the initiation of workflows for data pre- and post-processing, which can then be executed on attached compute servers. In the event of a workflow requiring significant computational resources, it can be executed on a high-performance computing (HPC) system (see Req:Workflows).

User management system (IDM): All components of the RDI are linked to an IT service for user management (see Req:Accessibility and Req:Usability) in order to facilitate the centralised management of researcher roles and rights, as well as the provision of RDI services via standardised credentials. Consequently, external partners can be readily incorporated via guest access.

Labelling management system: A unique global identifier will be allocated to each created sample (see Req:Labelling). This approach enables the tracking of samples throughout the process chain. Consequently, the management of samples can be conveniently conducted via web forms accessible from the frontend.

Term set management system: A term set management system, structured according to a taxonomy, has been integrated into the ELN. The system contains a set of default terms and synonyms that can be employed by researchers during the data documentation process (see Req:TechnicalLanguage).

Rule set: A set of rules governs the rights and obligations associated with the utilisation of the research data infrastructure. The regulations encompass a range of provisions pertaining to access to the RDI, the storage and documentation of research data, and other related matters.

Secure network connections: The fundamental elements of the RDI, namely the electronic laboratory notebook for the documentation and retrieval of data and the central data storage for the storage of data, are accessible via the secure network protocols HTTPS and Secure Shell (see Req:DataSecurity). Additional RDI components, such as external computing services for data processing, are accessible solely from within the intranet. External partners may gain access to the intranet via the Virtual Private Network (VPN) gateway. This guarantees a high level of IT security while maintaining unrestricted access to the fundamental functionality of the RDI (see Req:DataSecurity and Req:Usability). Connections to the data storage are always possible via the secure file transfer protocol SFTP (SSH File Transfer Protocol/Secure File Transfer Protocol). In the event that users are situated within the intranet of the organisation, connections to the data storage may also be established via the protocol SMBv3 (Server Message

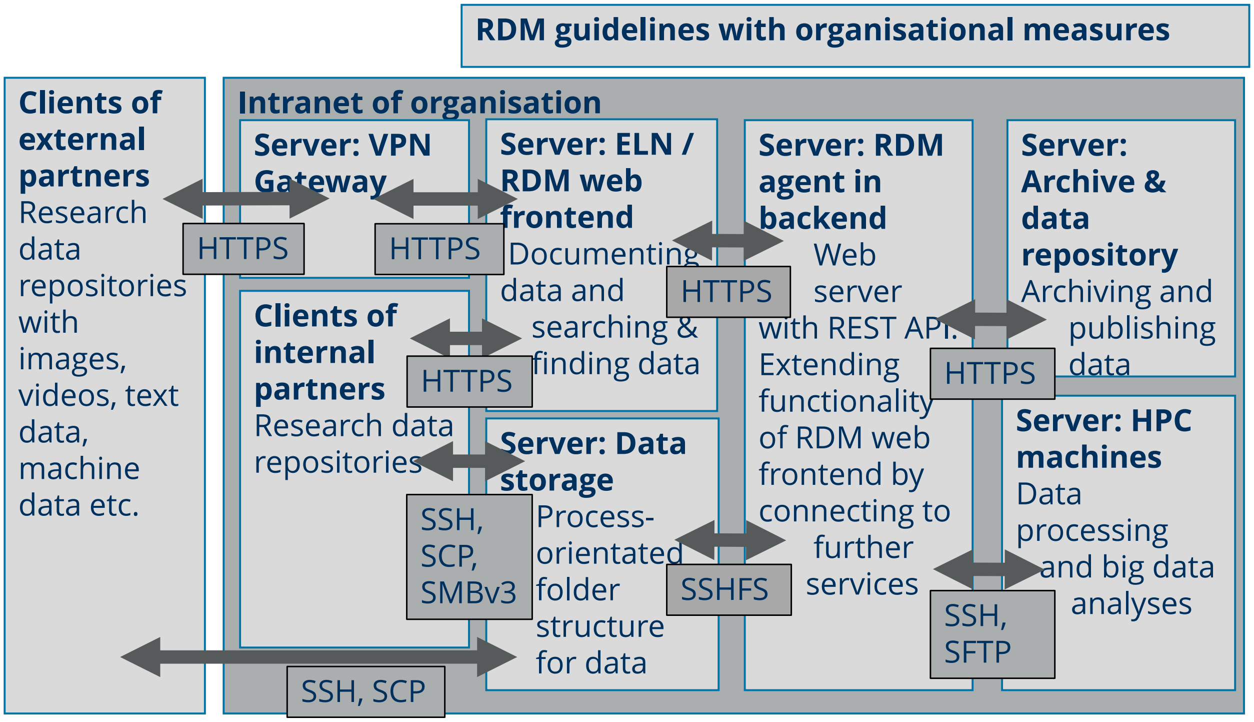

Figure 10. RDI within the solution concept for the research data management in AMTwin.

Block version 3). SMBv3 enables the data storage to be accessed as a network drive on operating systems such as Windows, thereby facilitating convenient browsing of the data storage contents through graphically based file explorers like the Windows Explorer. In order to permit end-to-end encrypted transmissions, it is necessary to enforce version 3 or higher of the protocol. Once more, this guarantees a high level of IT security while imposing only minimal restrictions on usability (see Req:DataSecurity and Req:Usability).

The key strategy is to base RDI components on existing IT services wherever possible. This approach has the advantage of reducing the effort required to set up and maintain the RDI (see Req:Applicability). Examples of existing IT services that could be leveraged include those provided by the internal data centre of the organisation.

In conclusion, the concept encompasses all the requisite components to facilitate researchers in the fundamental data-related workflows throughout the conventional research data life cycle. In particular, researchers are able to document research data in accordance with standardised designations. Although this concept was developed for AMTwin, it has broader applicability.

Figure 11 shows the used hardware and software components, given in yellow coloured boxes. The following components are used in the RDI:

1) The TUD network for establishing secure network connections between the services within the RDI (official service of TUD data centre).

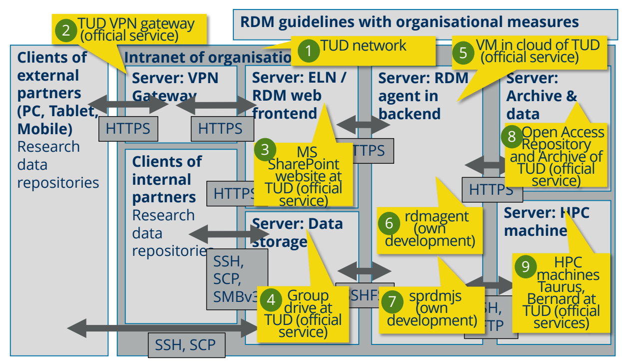

Figure 11. Used hardware and software components in RDI of AMTwin (given in yellow coloured boxes).

2) the VPN gateway servers of the TUD (official service of TUD data centre).
3) A project website at the Microsoft SharePoint® (abbreviated: SP) at the TUD serving as ELN (official service of TUD data centre).
4) A group drive at the TUD storage system as central data storage system (official service of TUD data centre).
5) Virtual machines in the TUD research cloud (official service of TUD data centre) for hosting the web-based agent.
6) The self-developed software tool "rdmagent" as an implementation of the web-based agent for connecting the ELN to external storage and compute systems.
7) The self-developed software tool "sprdmjs" for enhancing the functionality of SP websites related to the RDM.

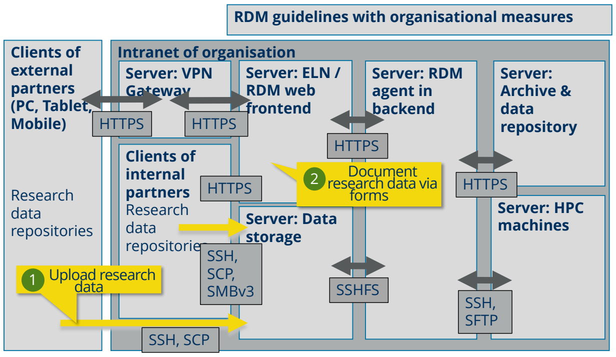

Figure 12. Workflow for storing and documenting data within used RDM concept.

8) The institutional data repository and archive for archiving and publishing research data (official services of TUD data centre).
9) The HPC machines Taurus and Barnard of the TUD for big data processing and analyses (official services of TUD data centre).

As it is not obvious to use SP as an ELN for research data management, the main reasons are described below:

Support: The software is an official service of the TUD and as such is free of charge and supported by the TUD. This has the consequence that no effort related to the installation of SP is required on the part of the user of data backups are created in an automated way.

Data security: As the service is operational as a premise solution at the TUD, the data stored on SP is located on TUD servers in Saxony, Germany. This is in accordance with the General Data Protection Regulation (GDPR) of the European Union.

Access: SP web sites are accessible from any location worldwide via ordinary web browsers, the access is independent of the device, and connections are encrypted. Mobile applications for SP are available for both the Android and iPhone operating systems. The authentication and authorisation processes are facilitated by the identity management system of the TUD. Furthermore, a fine-grained role-rights management system is available.

Document management: The document management system offers versioning and full-text search and indexing.

2) Data documentation: The following section provides a comprehensive account of the procedures employed in AMTwin for the storage and documentation of research data. The general workflow is as follows, compare Figure 12:

Step 1. Store: Research data are uploaded from local research data repositories into the central file storage system at the RDI. The storage system contains a process oriented folder structure for the data.

Step 2. Document: The process-specific metadata schemes for the research data are filled via web forms in the web frontend. This facilitates a convenient data documentation management via a graphical user interface.

This way, the documentation of the research data is stored in two locations: at the web frontend and at the central storage system in terms of text files (named README files). The text files are automatically stored together with the associated research data in a shared folder on the central file server. This facilitates access to the data documentation via the RDM web frontend and directly via the central file server.

The research data is documented in a structured form using metadata schemas with key-value pairs, in accordance with the solution concept for documenting data proposed in Section III of this contribution. The key-value pairs are based on a given domain ontology. The class names of the ontology serve as the keys, while the potential characteristics of the classes constitute the values. This guarantees that all terminology employed is derived from a unified technical lexicon that is comprehensible to all project partners. In addition to subject-specific entries, the metadata schemas also contain general key-value pairs (e.g., the licence model used) in order to facilitate the publication and archiving of research data on publication platforms.

In order to facilitate communication within the relatively nascent field of additive manufacturing, an application ontology has been developed, designated as OFAM (Ontology for Additive Manufacturing). This enables the comprehensive description of all processes, machines, and materials related to AM. The ontology is based on the upper ontology EMMO. As a consequence of the extension of EMMO, the application ontology OFAM is compatible with ontologies from other domains, in particular in the domain of applied sciences. Consequently, it can be merged with these ontologies with relative ease.

An exemplar of a researcher MDS for engineering processes in terms of a JSON-LD file is provided in Listing 1. Lines 2-5 delineate the context through the utilisation of specified vocabularies, namely schema.org and the branch designated as "holistic" within the EMMO framework. The abbreviations of the aforementioned vocabularies are defined (sorg, emmoholistic), which may subsequently be utilised in the key names of the MDS. Lines 6-20 contain administrative metadata, including the name of the author, the date of recording, and the type of recording. Lines 21-22 contain metadata pertaining to the governance of research data, including the data licence and the degree of public access. Lines 23-24 contain further metadata, including the process name and a list of items involved in the process. Lines 25-31 contain process-specific technical metadata, including the machine used within the process, setup parameters and characteristic values as process output quantities.

```
1 {
2   "@context": {
3     "sorg": "https://schema.org/version/14.0/",
4     "emmoholistic": "http://emmo.info/emmo/1.0.0-beta/
          middle/holistic#"
5   },
6   "sorg:author": {
7     "sorg:affiliation": "",
8     "sorg:email": "",
9   },
10  "sorg:description": "",
11  "sorg:keywords": [],
12  "sorg:inLanguage": "en",
```

```
"dataReadingSoftware": "",
"sorg:recordedAt": {
  "sorg:startDate": "2024-08-22T00:00:00+01:00",
  "sorg:endDate": "2024-08-22T00:00:00+01:00"
},
"dataRecordingType": "Machine",
"degreeAggregation": "010",
"anomalyMarking": "NA",
"sorg:license": "group",
"degreePublicAccess": "group",
"emmoholistic:Process": "",
"items": [],
"machine": {
  "sorg:name": "",
  "id": ""
},
"setupParameters": [],
"characteristics": []
}
```

Listing 1. Example of a researcher MDS.

The use of JSON metadata schemas ensures structured input. To ensure the integrity and accuracy of the data, it is essential to validate the completed researcher MDS. This is done by automatically validating the researcher MDS in JSON format using a JSON validator. This requires all researcher MDS to be based on JSON Schema [68]. JSON Schema is a vocabulary for annotating and validating JSON data. Data structure constraints can be defined to catch errors, inconsistencies and invalid data. Schema validation then automatically checks the data for types and value ranges. AJV [69] was selected as the JSON schema validator due to a high runtime performance during validation and support for various standards (e.g., JSON Schema Drafts, JSON Type Definition). An extract of the JSON schema for the Researcher MDS, as detailed in Listing 1, is provided in Listing 2. `$schema` in line 3 describes the JSON schema used to validate JSON files, `$id` in line 4 the global identifier of the schema, including a version number in semantic style. The vocabularies used from json-schema.org are specified in lines 6-8. Line 10-29 state that the JSON files to be validated consist of a JSON object with properties `context`, `sorg:description`, `machine`. For each property, a description and the type are given. The elements required in the researcher MDS are given in the array named `required`, see lines 31-34.

```
1  {
2      "title": "JSON schema for researcher MDS of process A"
       ,
3      "$schema": "https://json-schema.org/draft/2019-09/
           schema#",
4      "$id": "https://<XXX>/schemas/A/0.1.0/",
5      "$vocabulary": {
6          "https://json-schema.org/draft/2019-09/vocab/core"
               : true,
7          "https://json-schema.org/draft/2019-09/vocab/
               format": true,
8          "https://json-schema.org/draft/2019-09/vocab/
               content": true
9      },
10     "type": "object",
11     "properties": {
12         "sorg:description": {
13             "$comment": "sorg:Text",
14             "description": "Short description of data set"
                   ,
15             "type": "string"
16         },
17         "machine": {
18             "description": "Used machine during (
                   manufacturing) process",
19                 "type": "object",
20                 "properties": {
21                     "id": {
22                         "description": "Unique id of
                               machine",
23                         "type": "string"
24                     },
25                     "sorg:name": {
26                         "$comment": "sorg:Text",
27                         "description": "Descriptive name
                               of machine (if available)",
28                         "type": "string"
29                     },
30 [...]
31     "required": [
32         "machine",
33         "emmoholistic:Process",
34         "sorg:description"
35     ]
36 }
```

Listing 2. JSON schema for researcher MDS as detailed in Listing 1.

An example of how to model tensile tests within the RDM concept and document the associated process data is given in the following. Figure 13 shows the steps involved in this process. In addition to general information such as the test date and name of the tester, the modelling of tensile tests also requires process-specific information on the test setup and test execution (e.g., in the form of specimen ID, test program, and test speed) as well as on the test results (e.g., yield strength, breaking limit, and elongation at break). Figure 13 a) shows an excerpt from the application ontology OFAM to structure and model the process. The excerpt shows the selected classes for the process together with their cross-relationships and properties. The process itself is modelled by the class "Tensile test", and the properties of the process by "HAS-A" relationships to other classes. Each class in yellow nodes is a subclass from the upper ontology EMMO. This ensures that all class names and terms used are clearly identifiable and that all project partners have the same understanding of them.

To illustrate, the class designated as "Tensile test" is a subclass of the class identified as "Destructive test methods." This class is, in turn, a subclass of the EMMO class designated "Processes". The measurement results are modelled using the EMMO classes "Yield strength", "Breaking point", and "Elongation at break", which are subclasses of the EMMO class "Physical quantity". The classes "floating point number" and "SI unit" are used to represent the measured values and units of physical quantities. The ontology serves as the foundation for the metadata schema utilised for data documentation purposes. Figure 13 b) illustrates a portion of the associated metadata schema. The schema comprises key-value pairs, wherein the keys are the class names for the process "Tensile test". The format of the values is dependent on the permitted data types. It is necessary to implement the metadata schema in practice so that the test data for the process can be documented, for example, via an input mask.

Figure 13 c) illustrates the practical implementation of the RDM web frontend SP. The metadata schema is implemented via a web form within SP. The form contains all the key-value pairs from the metadata schema pertaining to the tensile test. For each key in the schema, there is a corresponding input field

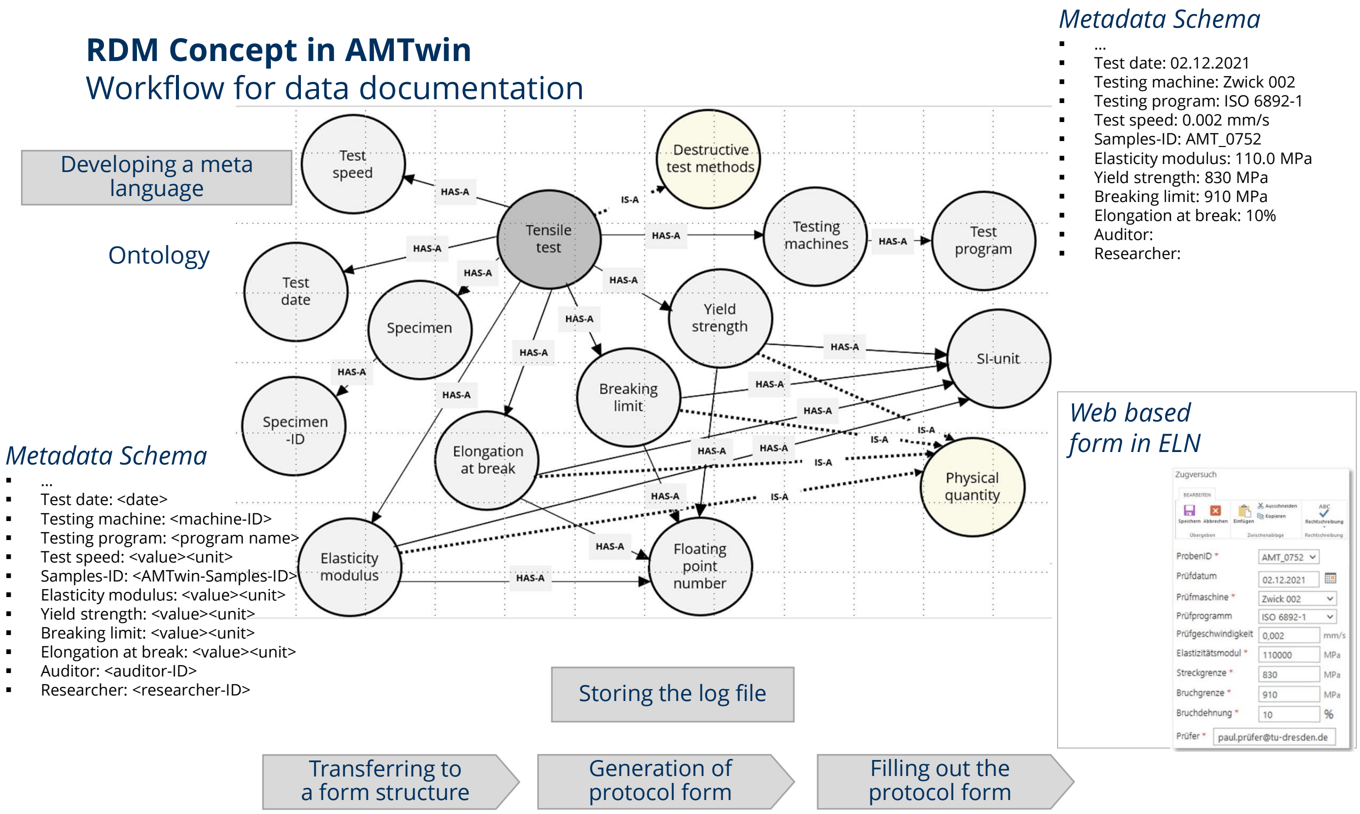

Figure 13. Interaction of ontology, metadata schema and test protocol: a) section of the OFAM ontology for the definition of classes (circles) and their semantic linkage (arrows) for exemplary modelling of tensile tests, b) metadata schema as form structure, c) graphical user interface for the use of the form in the exemplary implementation in SP, d) section of the protocol file.

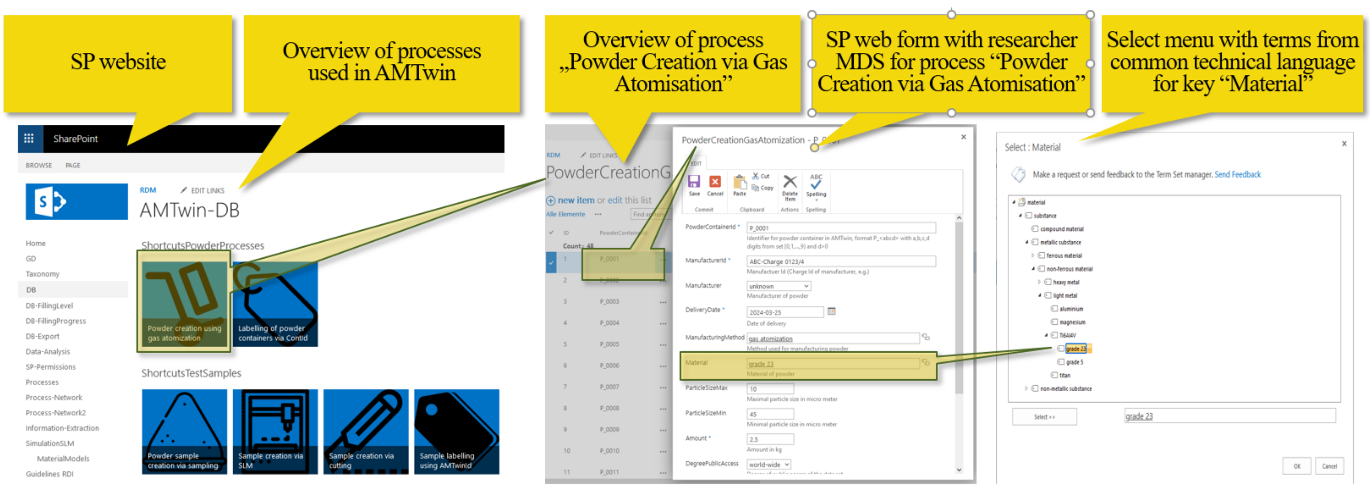

Figure 14. Workflow for documenting research data via web form in SP.

in the form. The type of field is determined by the permitted data type associated with the key. For example, a number field is used for the yield strength, while a text selection menu is employed for the testing machine. Additionally, each field includes a description to facilitate data entry. The entered values are subjected to a validity check, whereby the adherence to the prescribed value ranges and the plausibility of the entered values are verified. The web form in Figure 13 c) has already been populated with specific values for a tensile test in AMTwin, using the sample ID AMT_0752. Figure 14 illustrates in detail the workflow for documenting research data via SP web forms.

In order to ensure the sustainable utilisation of the data documentation, it is essential that it is stored in an easily accessible location. Consequently, upon saving the SP form, a new entry is automatically generated in the corresponding SP list within the RDM web frontend, comprising all the details entered for the test. Concurrently, a log file containing the process data is saved on the central file server. Figure 13 d) illustrates the section of the automatically generated documentation file pertaining to the tensile test. The file contains all key-value pairs in accordance with the metadata schema, together with the specific values pertaining to the tensile test on the given specimen.

3) Summary and conclusion: The metadata model proposed in this contribution has been considered and could be successfully integrated in the RDM solution concept. Official services of the TUD data centre have been used extensively, resulting in low set-up and maintenance costs. The amount of training required for using SP as ELN is generally low, as many researchers are already familiar with SP due to its widespread use at research institutions. In conclusion, the solution concept used in AMTwin allows for data integration which in turn allows for enhanced collaboration and improves overall efficiency. This concept enables the creation of a more comprehensive dataset, facilitating the analysis and interpretation of the integrated data.

V. Outline of the results

A solution concept has been proposed whereby research data can be documented in accordance with subject-specific ontologies. The viability of the proposed metadata model has been evaluated through its integration into the research data management processes of two joint engineering research projects. The concept is largely independent of the specific research projects, as previously outlined, and can therefore be readily transferred to other joint projects. Thus, it can be concluded that the presented solution concept can be applied to a large class of research and engineering projects with only minor adaptations, reducing the setup and maintenance effort in joint research projects and enhancing the reusability and reproducibility of the results.

The strategy facilitates a paradigm shift from subjectively designed individualistic conceptions to the handling of research data in a manner that is objectively aligned with established, harmonised solutions. The motivation for this work is the recognition of the importance of harmonised data preparation and subsequent documentation in the engineering domain. The impetus for this work stems from the recognition of the pivotal importance of standardising data preparation and subsequent documentation in the engineering domain.

The proposed metadata model facilitates the integration of disparate domain-specific languages and work cultures by providing a common language that all researchers and engineers from different domains can comprehend. This is achieved through the utilisation of metadata, which facilitates the unification of physical units and the interconnection of disparate domain-specific languages. In particular, the metadata model facilitates the unification of terms. The data are stored and documented in such a way that data from different processes along a process chain can be merged, thereby creating a single overall dataset. Consequently, cross-process data analysis methods may be employed. It is possible that the proposed metadata model may prove too general or abstract for application across the entire engineering domain. The solution approach permits the consolidation 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.

The metadata model also allows for inter-domain communication by defining a common set of concepts and relationships that can be used across different domains. The model offers a methodology for the management of metadata, delineating a set of rules governing its structure and storage.

While the metadata model has many advantages, it also has some disadvantages.

- The creation of a metadata template necessitates the input of metadata experts, whereas the objective should be for researchers or domain experts to utilise or construct it independently.
- The final result is contingent upon the input of domain experts. In conclusion, the proposed solution enables the creation of global datasets in a manner that facilitates analysis. This approach will facilitate enhanced interoperability and collaboration among disparate engineering research groups.
- A common language was established through a survey of data providers, with the objective of defining a shared technical vocabulary.

In conclusion, this metadata model offers a promising approach to addressing the challenges of research data management and improving collaboration among researchers and engineers from different domains. The solution guarantees that the data are documented in a comprehensible manner, thus ensuring that other researchers can understand them. The proper identity management of components, processes, and machines across laboratory boundaries ensures the interoperability of data. This facilitates the availability of the data for subsequent data-driven analyses across laboratory and process boundaries. The analysis results based on the documented research data can be reproduced at a high level of quality, due to the comprehensive data documentation.

VI. Conclusion and future research perspective

A. Conclusion

The proposed strategy facilitates the navigation of disparate working cultures by offering a unified approach that is comprehensible to researchers and engineers from a range of domains. This is accomplished through the utilisation of metadata augmented by the formulation of suitable ontologies. In particular, the metadata model facilitates the storage and documentation of data, thereby enabling the merging of data from disparate processes within a process chain, thus allowing for the utilisation of cross-process data analysis methods.

This article employs a use case approach to provide a summary of the existing requirements for practical research data management in the AMTwin and GRK 2250 joint projects. It also presents a solution concept that allows for the documentation of research data based on a subject-specific ontology. The feasibility of the concept was validated as an exemplar for the documentation of tensile tests as part of the AMTwin joint project, as well as for the investigation of mineral-bounded composites within the GRK 2250 project. The concept is largely independent of both use cases, and thus can be readily transferred to other collaborative endeavours. The ontology OFAM can be readily linked and reused with other ontologies, particularly those from the materials sciences, due to the extension of the EMMO basic ontology.

It has been demonstrated that researchers require assistance in establishing structured process-data models. It is challenging for researchers to identify all the metadata that must be recorded in order to document the experiment in a repeatable manner. Consequently, it is possible that crucial influencing factors in the experiments were not documented, resulting in the generated data being of limited reusability. The structured process data model (as exemplified in Figure 8) is designed to assist in the identification of all requisite steps and influences.

B. Future research perspectives

In order to consider and analyse cross-process relationships, it is first necessary to obtain a global view of the dataset in an analysable form. This necessitates the availability of meticulously documented data that can be integrated into global datasets [14], [15]. This is because subsequent data-driven modelling is not within the purview of this study.

The issue of usability is not addressed in sufficient depth in this paper. It remains unclear whether and how this issue can be adequately addressed, given that establishing an MDM system that adheres to the FAIR data principles is a significant undertaking in itself. Ultimately, however, the efficacy of such a system hinges on the active participation of all relevant stakeholders. As evidenced by the experience gained through the use cases presented, the MDM system often proves ineffective due to an inherent overhead burden on the individual researcher. In addition to the implementation of the FAIR data principles, it is essential that the MDM system generates overhead for the researcher, while also providing short-term benefits and facilitating their work.

It is essential that the data management system be designed to facilitate rapid implementation and straightforward adaptation to the evolving needs of the research network. In other words, it must be highly customisable. A considerable amount of resources are currently being allocated to the development of various data management systems, including AMTwin and GRK2250. Nevertheless, it will not be feasible to construct a novel MDM system from the ground up for each research project in the future, as this would necessitate a considerable investment of resources. It is therefore essential that the developed systems be customisable and reusable. This aspect aligns with the concept of usable AI, as discussed in [59]. Nevertheless, this study does not present a solution to the aforementioned challenge, which pertains to MDM.

The proposed model may be employed as a framework for the management of digital objects in other research domains, including the social sciences and natural sciences. Furthermore, additional research could be conducted to investigate the potential for integrating this metadata into existing RDM systems, or to identify areas for improvement to better meet the needs of different users.

The two principal chapters (static and dynamic metadata) may be subdivided into further subchapters (e.g., administrative, organisational, and technical metadata) for the purpose of containing attribute-value pairs. It should be noted, however, that this general schema has not yet been finalised. It is also necessary to determine which dynamic metadata must be captured in order to ensure the reproducibility and repeatability of the experiment.

The objective of the Industrial Ontologies Foundry (IOF) initiative [70] is similar to that proposed for the OBO Foundry (for biomedicine) [71]. In both cases, adherence to a standard upper-level ontology is of paramount importance in facilitating harmonisation. This upper-level ontology is designated Basic Formal Ontology (BFO) [72]. It would be beneficial to consider the relationship between the current effort and the wider initiative to curate and facilitate access to industrial ontologies.

Acknowledgment

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). The German Federal Ministry for Economic Affairs and Climate Protection (BMWK) has provided funding for this research project based on decisions made by the German Bundestag within the joint research projects "SWaT" (grant number 20M2112F) and "LaSt" (grant number 20M2118F). Additionally, the BMWK has provided funding for the project through the funding guideline "Digitization of the vehicle manufacturers and supplier industry" within the funding framework "Future investments in vehicle manufacturers and the supplier industry", which is 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).

REFERENCES

[1] M. Zinner, F. Conrad, K. Feldhoff, H. Wiemer, J. Weller, and S. Ihlenfeldt, “A Metadata Model for Harmonising Engineering Research Data Across Process and Laboratory Boundaries,” in *COGNTIVE 2024: The Sixteenth International Conference on Advanced Cognitive Technologies and Applications*, 2024, pp. 30–39. [Online]. Available: https://www.thinkmind.org/articles/cognitive_2024_1_50_40019.pdf

[2] 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.

[3] 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.

[4] 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: September 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2295861

[5] 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: September 2024. [Online]. Available: https://doi.org/10.1002/wat2.1424

[6] 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: September 2024. [Online]. Available: https://www.nature.com/articles/sdata2017191

[7] 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: September 2024. [Online]. Available: https://doi.org/10.1016/j.mspro.2013.02.005

[8] 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: September 2024. [Online]. Available: https://doi.org/10.1007/978-3-662-62217-9

[9] 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: September 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389921000581

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

[11] 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: September 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0263822321003779

[12] 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: September 2024.

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

[14] 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: September 2024. [Online]. Available: https://www.nature.com/articles/s41598-022-23327-1

[15] 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: September 2024. [Online]. Available: https://www.nature.com/articles/s41524-018-0081-z

[16] 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

[17] N. Brandt *et al.*, “Kadi4Mat: A research data infrastructure for materials science,” *Data Science Journal*, vol. 20, pp. 8–8, 2021, retrieved: September 2024. [Online]. Available: https://doi.org/10.5334/dsj-2021-008

[18] D. Fensel, U. Simsek, K. Angele, E. Huaman, K. Elias, O. Panasiuk, I. Toma, J. Umbrich, and A. Wahler, *Knowledge Graphs - Methodology, Tools and Selected Use Cases*, D. Fensel, Ed. Springer, 2020.

[19] 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: September 2024. [Online]. Available: https://opus4.kobv.de/opus4-fhpotsdam/files/192/1.1_Research_Data_Management.pdf

[20] A. M. de Carvalho Moura, M. L. Machado Campos, and C. M. Barreto, “A survey on metadata for describing and retrieving internet resources,” *World Wide Web*, vol. 1, no. 4, pp. 221–240, 1998.

[21] princeton.edu. (2024) What is a data model? [Online]. Available: https://cedar.princeton.edu/understanding-data/what-data-model

[22] J. Riley, “Seeing Standards: A Visualization of the Metadata Universe,” 2018, retrieved: September 2024. [Online]. Available: https://doi.org/10.5683/SP2/UOHPVH

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

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

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

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

[27] 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: September 2024. [Online]. Available: https://arxiv.org/pdf/2005.01637

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

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

[30] K. Feldhoff, H. Wiemer, and S. Ihlenfeldt, “FDM als Service für ein typisches Verbundprojekt in den Ingenieurwissenschaften auf Basis einer ontologie-basierten Verschlagwortung; Eng : RDM as a service for a typical joint project in the engineering sciences applied on an ontology based keywording,” 2021, retrieved: September 2024. [Online]. Available: https://doi.org/10.5281/zenodo.5718660

[31] B. Glimm and Y. Kazakov, “Classical algorithms for reasoning and explanation in description logics,” *Reasoning Web. Explainable Artificial Intelligence: 15th International Summer School 2019, Bolzano, Italy, September 20–24, 2019, Tutorial Lectures*, pp. 1–64, 2019, retrieved: September 2024. [Online]. Available: https://doi.org/10.1007/978-3-030-31423-1_1

[32] J. Pomerantz, *Metadata*. MIT Press, 2015, retrieved: September 2024. [Online]. Available: https://vtechworks.lib.vt.edu/bitstream/handle/10919/52547/DL_4-b_2009-10-07.pdf?sequence=1&isAllowed=y

[33] GitHub, “Elementary Multiperspective Material Ontology (EMMO),” *Generated by GitHub Pages*, 2022, retrieved: September 2024. [Online]. Available: https://github.com/emmo-repo/EMMO

[34] H. Li, R. Armiento, and P. Lambrix, “An ontology for the materials design domain,” in *International Semantic Web Conference*. Springer, 2020, pp. 212–227, retrieved: September 2024. [Online]. Available: https://arxiv.org/pdf/2006.07712

[35] I. E. Castelli *et al.*, “Data Management Plans: the Importance of Data Management in the BIG-MAP Project,” *Batteries & Supercaps*, vol. 4, no. 12, pp. 1803–1812, 2021, retrieved: September 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/batt.202100117

[36] 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.

[37] S. Bechhofer *et al.*, “Owl web ontology language reference,” *W3C recommendation*, vol. 10, no. 2, pp. 1–53, 2004.

[38] D. L. McGuinness, F. Van Harmelen *et al.*, “Owl web ontology language overview,” *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004, retrieved: September 2024. [Online]. Available: https://static.twoday.net/71desa1bif/files/W3C-OWL-Overview.pdf

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

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

[41] E. Ghedini, “Europeean Materials Modelling Ontology Version 1.0.0-beta,” *Generated by GitHub Pages*, 2020, retrieved: September 2024. [Online]. Available: https://emmo-repo.github.io/versions/1.0.0-beta/emmo.html

[42] 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: September 2024. [Online]. Available: https://doi.org/10.1145/2815833.2816955

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

[44] 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: September 2024. [Online]. Available: http://ceur-ws.org/Vol-897/session4-paper20.pdf

[45] 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: September 2024. [Online]. Available: https://doi.org/10.1186/2041-1480-5-25

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

[47] M. M. Ali, R. Rai, J. N. Otte, and B. Smith, “A product life cycle ontology for additive manufacturing,” *Computers in Industry*, vol. 105, pp. 191–203, 2019, retrieved: September 2024. [Online]. Available: https://doi.org/10.1016/j.compind.2018.12.007

[48] M. Mayerhofer, W. Lepuschitz, T. Hoebert, M. Merdan, M. Schwentenwein, and T. I. Strasser, “Knowledge-driven manufacturability analysis for additive manufacturing,” *IEEE Open Journal of the Industrial Electronics Society*, vol. 2, pp. 207–223, 2021, retrieved: September 2024. [Online]. Available: https://doi.org/10.1109/OJIES.2021.3061610

[49] J. Cuenca, F. Larrinaga, and E. Curry, “Dabgeo: A reusable and usable global energy ontology for the energy domain,” *Journal of Web Semantics*, vol. 61, p. 100550, 2020, retrieved: September 2024. [Online]. Available: https://doi.org/10.1016/j.websem.2020.100550

[50] ——, “A Unified Semantic Ontology for Energy Management Applications.” in *WSP/WOMoCoE@ ISWC*, 2017, pp. 86–97, retrieved: September 2024. [Online]. Available: http://www.edwardcurry.org/publications/WOMoCoE_17.pdf

[51] M. R. Neuhaus and M. Stappel. The open energy ontology. retrieved: September 2024. [Online]. Available: https://svn.aksw.org/papers/2020/EKAW_OEO/public.pdf

[52] H. Li and T. Hong, “A semantic ontology for representing and quantifying energy flexibility of buildings,” *Advances in Applied Energy*, vol. 8, p. 100113, 2022, retrieved: September 2024. [Online]. Available: https://doi.org/10.1016/j.adapen.2022.100113

[53] D. Garijo, M. Osorio, D. Khider, V. Ratnakar, and Y. Gil, “Okg-soft: An open knowledge graph with machine readable scientific software metadata,” in *2019 15th International Conference on eScience (eScience)*. IEEE, 2019, pp. 349–358, retrieved: September 2024. [Online]. Available: https://doi.org/10.1109/eScience.2019.00046

[54] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, and R. Rosati, “Ontology-based database access.” in *SEBD*, 2007, pp. 324–331.

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

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

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

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

[59] 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

[60] ——, “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: September 2024.

[61] 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: September 2024.

[62] (2024) Home Page Symate GmbH. retrieved: September 2024. [Online]. Available: https://www.symate.de/

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

[64] K. Feldhoff and H. Wiemer, “Praktikables, Ontologie-basiertes Forschungsdatenmanagement in der Additiven Fertigung; Eng: Practical, Ontology-Based Research Data Management in Additive Manufacturing,” *Brockmann, S. ;Krupp, U. (Hrsg.): 39. Vortrags- und Diskussionstagung Werkstoffprüfung„Werkstoffe und Bauteile auf dem Prüfstand: Prüftechnik Kennwertermittlung – Schadensvermeidung“. Düsseldorf: Stahlinstitut VDEh, 2021, ISBN 978-3-941269-98-9*, 2021.

[65] M. Kästner *et al.* (2020) Data-driven process, material and structure analysis for additive manufacturing. retrieved: September 2024. [Online]. Available: https://tu-dresden.de/ing/maschinenwesen/ifkm/nefm/forschung/projekte/amtwin?set_language=en

[66] M. Zimmermann. (2022) Data-Driven Process, Material and Structural Analysis for Additive Manufacturing (AMTwin). retrieved: September 2024. [Online]. Available: https://www.iws.fraunhofer.de/en/technologyfields/materials_characterization/amtwin.html

[67] M. Pérez, D. Carou, E. M. Rubio, and R. Teti, “Current advances in additive manufacturing,” *Procedia Cirp*, vol. 88, pp. 439–444, 2020, retrieved: September 2024. [Online]. Available: https://doi.org/10.1016/j.procir.2020.05.076

[68] A. of jsonschema.org. (2024) Json schema: Build more. break less. empower others. [Online]. Available: https://json-schema.org/

[69] A. of ajv.js.org. (2024) Ajv json schema validator. [Online]. Available: https://ajv.js.org/

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

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

[72] 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: September 2024. [Online]. Available: https://ceur-ws.org/Vol-3240/paper3.pdf