A Perspective on Digital Knowledge Representation in Materials Science and Engineering

12 authors, including:

Bernd Bayerlein
Bundesanstalt für Materialforschung und -prüfung
6 PUBLICATIONS **112** CITATIONS

SEE PROFILE

Thomas Hanke
Fraunhofer Institute for Microstructure of Materials and Systems IMWS
7 PUBLICATIONS **4** CITATIONS

SEE PROFILE

Thilo Muth
Bundesanstalt für Materialforschung und -prüfung
58 PUBLICATIONS **1,282** CITATIONS

SEE PROFILE

Benjamí Moreno Torres
Bundesanstalt für Materialforschung und -prüfung
6 PUBLICATIONS **6** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Metaproteomics software development View project

Sequential Learning to accelerate Materials discovery View project

A Perspective on Digital Knowledge Representation in Materials Science and Engineering

Bernd Bayerlein, Thomas Hanke, Thilo Muth, Jens Riedel, Markus Schilling, Christoph Schweizer, Birgit Skrotzki, Alexandru Todor, Benjami Moreno Torres, Jörg F. Unger, Christoph Völker, and Jürgen Olbricht*

The amount of data generated worldwide is constantly increasing. These data come from a wide variety of sources and systems, are processed differently, have a multitude of formats, and are stored in an untraceable and unstructured manner, predominantly in natural language in data silos. This problem can be equally applied to the heterogeneous research data from materials science and engineering. In this domain, ways and solutions are increasingly being generated to smartly link material data together with their contextual information in a uniform and well-structured manner on platforms, thus making them discoverable, retrievable, and reusable for research and industry. Ontologies play a key role in this context. They enable the sustainable representation of expert knowledge and the semantically structured filling of databases with computer-processable data triples.

In this perspective article, we present the project initiative Materials-open-Laboratory (Mat-o-Lab) that aims to provide a collaborative environment for domain experts to digitize their research results and processes and make them fit for data-driven materials research and development. The overarching challenge is to generate connection points to further link data from other domains to harness the promised potential of big materials data and harvest new knowledge.

1. Introduction

Materials science and engineering (MSE) is an interdisciplinary field of its own that touches on both natural sciences and engineering. It aims to investigate the relationship between the manufacturing process, resulting microstructure, and properties to develop materials with optimized characteristics and maximize their reliability, service life, and recyclability.[1,2] A special feature and a challenge at the same time are the many structural scales that must be taken into account. They range from the atomic scale (nm) to the micro- (μm) to the macroscale (from mm to m). Another challenge is the variety of investigation methods, standards, and models used for characterization in this highly interdisciplinary field, up to and including the use of modeling and simulation techniques.[3–6]

The data generated in MSE is therefore per se characterized by a high degree of heterogeneity. In addition, the increasing integration of (partly robot-based) high-throughput methods and experiments, the continuously growing computing power and its resources, and the variety of software-based analysis options lead to a very high rate of data creation in a multitude of different formats.[7,8]

For some time now, the digital transformation has also been changing and shaping the MSE research landscape.[9,10] However, the promise of successively transforming the generated research data into knowledge by applying data-science driven methods (such as machine learning [ML] and deep learning approaches) and thus accelerating material discovery and material design according to the fourth paradigm (data-driven approach) has yet to be fulfilled.[11–14] In the context of optimal handling of MSE research data, the FAIR data principles provide orientation and enable the fundamental challenges of digitalization to be addressed. According to these principles, research data and corresponding metadata should be findable, accessible, interoperable, and reusable.[15] A powerful tool for implementing these principles are modularizable and extensible ontologies. They allow to semantically structure and annotate raw data, processed data, and contextual data using a commonly shared, consistent, and understandable vocabulary based on fundamental terms.[7,16–20]

B. Bayerlein, T. Muth, J. Riedel, M. Schilling, B. Skrotzki, B. Moreno Torres, J. F. Unger, C. Völker, J. Olbricht
Bundesanstalt für Materialforschung und -prüfung (BAM)
Unter den Eichen 87, 12205 Berlin, Germany
E-mail: juergen.olbricht@bam.de

T. Hanke
Fraunhofer-Institut für Mikrostruktur von Werkstoffen und Systemen IMWS
Walter-Hülse-Straße 1, 06120 Halle (Saale), Germany

C. Schweizer
Fraunhofer-Institut für Werkstoffmechanik IWM
Wöhlerstraße 11, 79108 Freiburg im Breisgau, Germany

A. Todor
Fraunhofer-Verbund Werkstoffe
Bauteile – MATERIALS
Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adem.202101176.

Different types and levels of ontologies exist and can be subdivided into three main categories. 1) Top-level (or upper) ontologies (TLOs) describe general terms that are common across many domains. One example of a standardized upper ontology is the basic formal ontology (BFO). 2) Midlevel (or core) ontologies (MLOs) represent higher-level, more abstract concepts that enable the complex and expressive domain ontologies to be interconnected. 3) Domain (or domain-specific) ontologies are developed based on explicit expert knowledge and represent concepts that belong to specific domains, for example, specific processes or research methods. Thus, this knowledge is also prepared in an organized and sustainable way comprehensible to others.[7,21,22]

While ontologies provide a formal description of the data, specific types of databases, so-called triple stores, typically support ontologies as database schema models. Triple stores allow storing data in a machine-processable form in three linked data pieces, so-called triples which describe subject–predicate–object relationships. Besides the direct integration of ontologies, the major benefits of using triple stores are linkage of diverse data and time efficiency of data retrieval: evolving data from heterogeneous sources can be dynamically queried using semantic search functions and can be further enriched to build knowledge graphs. In addition, the application of inference-based techniques such as reasoning aims to enable the retrieval of implicit knowledge.[17,19,20,23–26]

While the technological basis can be directly established with existing methods, there is still a great demand for action, especially in data exchange and data sharing culture. This does not only require customized data management platforms with user-definable search, visualization, and analysis options but also community-driven MSE digitalization initiatives and platforms that promote this practice.[23]

Pioneering among the MSE digitalization initiatives, the Materials Genome Initiative (MGI) was launched in 2011, whose primary goal was to accelerate the discovery and deployment of new advanced materials systems at a fraction of the cost. MGI provides the necessary materials innovation ecosystem, consisting of a digital data infrastructure including computational and experimental tools.[27,28]

Complementary current initiatives include the MaterialDigital (PMD) innovation platform, which was launched in 2019 by a German consortium. It aims to use newly generated database and software tools to better understand the properties and behavior of materials and optimize them in a more targeted manner to make production processes more efficient. In this context, partner projects from science and industry are working together to establish a virtual materials data space and thus jointly implement and sustainably design digitization tasks for materials and their production.[29]

Such large community efforts make it possible to develop and implement standardization at different levels close to MSE applications, concerning, for example, data formats, terminologies, ontologies, evaluation routines, with the aim to establish and accelerate the digital transformation of MSE.[30] Interdisciplinary research fields such as MSE require standardized digital workflows for covering a diverse set of data conversion and enrichment pipelines (e.g., with respect to data cleaning, processing, and annotation) and for integrating heterogeneous data sources using data management tools and developed MSE ontologies.

An important aspect in this context is the integration of already existing databases with specific (meta)data structures. Recently, the Event-Sourced Architecture for Materials Provenances (ESAMP) database, developed for the storage of materials research data, demonstrated the added value of comprehensive (meta)data structure modeling.[31] The instantiation of 6 million measurements of 1.5 million samples from the Materials Experiment and Analysis Database (MEAD) enabled specific structured query language (SQL) queries, analysis, and knowledge generation from the records of sample origin, performed experiments, and derived results thus deposited.[32] Another approach is used in the Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE) open-source software pipeline for which an ontological framework enables the acquisition of a comprehensive and extensible (meta)data structure from chemical experiments, leading to a final curated data frame.[33] Finally, the example of the high-throughput experimental materials (HTEM) database highlights how in-depth (meta)data structuring in materials science can be effectively used and transferred based on an advanced laboratory information management system (LIMS). The latter system is responsible for automatically harvesting, indexing, and archiving experimental (meta)data into a data warehouse.[34] These approaches also highlight the importance of properly documenting the origin of data and changes to the data over time as part of a comprehensive provenance management. The well-structured and enriched datasets may serve as high-quality input for modeling and ML approaches on data management platforms.

To be able to transfer data-driven platforms into both research and industrial applications, it is essential to also provide intuitive user interfaces (UIs) and extended documentation with user and developer tutorials. So far, the MSE field has been missing a perspective on a one-stop-shop solution for representing materials data and knowledge meeting the abovementioned community requirements.

In this publication, details on a corresponding current project, the joint initiative being Mat-o-Lab of Bundesanstalt für Materialforschung und -prüfung (BAM)[35] and Fraunhofer Group Materials and Components (MATERIALS),[36] are outlined. In this context, the challenges and tasks associated with digitization in MSE will be generally discussed and put in relation to the initial approaches and solutions of the particular project. To provide a comprehensive overview, first, the background and design concept of the joint initiative is introduced in Section 2 of this work. The following two Sections 3 and 4 highlight different types of ontologies, that is, top-level, midlevel, and domain ontologies, going into more detail about their usage and development. Section 5 presents the Data Space concept chosen in Mat-o-Lab. Here, the intended implementations of the data in triple store, data sovereignty, and data exchange are discussed. Aspects of scientific simulation in the field of MSE, workflows for data processing, and model parameter determination are considered in Sections 6 and 7. Finally, the much-enhanced availability of comprehensive machine-readable data will enable data-driven materials research. Data science approaches, based on ML or sequential learning, are addressed in Section 8. The paper concludes with our vision on the opportunities for future digital-supported materials research and development. Interested parties are kindly invited to

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**

www.aem-journal.com

participate in refining the initial concepts and approaches of our project.

2. The Mat-o-Lab Approach

The Mat-o-Lab project stands for the digitization of materials and components along their entire life cycle. Using specific material examples and use cases, applicable and practical solutions for the overarching digital challenges are developed and made available for public use following the Materials Data Space concept.[36] Essential results (datasets, ontologies, tools for data structuring, and data analysis) from previous and ongoing research projects will be consolidated and improved in a targeted manner. Mat-o-Lab stands for a new, open, and agile collaboration between the institutes of the Fraunhofer Group Materials and Components (MATERIALS), the BAM institute, and interested partners from industry and academia. In fact, Mat-o-Lab aims to actively network with similar groups and organizations that share their interests and progress.

The project is divided into three innovation stages. In its first stage, Mat-o-Lab is focused on a specific use case of aluminum alloys for elevated temperature applications. Existing data on this topic, established in previous joint research projects of the project partners, are semantically structured and transferred into a corresponding repository. In a second stage, the developed structures and toolchains are adopted for two additional use cases which cover a second material class (e.g., polymers), thereby extending the range of covered characterization and simulation methods and refining toolchains for the development of domain-specific ontologies. The ontologies are developed in a collaborative manner by multiple domain experts according to existing standards or scientific best practices. The proposed ontology architecture offers high degrees of granularity to capture every aspect of the characterization process, ensuring research reproducibility and dataset compatibility. The final third stage will open the project for selected external partners from academia and industry to extend the work on data exchange processes toward data curation and brokerage. Most importantly, crosslinks from basic materials research to design and production will be established in this phase, thereby allowing to represent the full development chain for materials applications.

The three-stage concept allows for a continuous up-scaling of the developed MSE framework. In each stage, a manageable degree of complexity is maintained as the scope of considered fields and methods and the related requirements for their digital representation are clearly defined. An agile management process and digital communication strategy were introduced to meet the needs for crossinstitutional collaboration.

Accompanying stages 1 and 2, five teams were set up. The teams focus on 1) setting up a collaborative IT infrastructure; 2) defining a common data structure philosophy for materials data (ontologies, workflow tools, exemplary datasets); 3) defining digital representations of microstructural data and analysis methods; 4) establishing similar ontology-based descriptions of mechanical data and characterization methods; and 5) linking these experimental data with materials models by establishing automated workflows for the model calibration processes.

To avoid any divergent developments, both within the three project stages and the different expert teams, clear guidelines are mandatory for the daily work in the project. The Mat-o-Lab teams are committed to 1) using similar toolchains for domain ontology development and referencing to similar TLOs and MLOs; 2) applying similar workflows for data structuring and fulfilling the requirements of the overall data space concept; and 3) providing data concepts that allow subsequent automated data analyses and model calibration.

Details on these key elements of the Mat-o-Lab framework will be outlined in the following sections, thereby opening a general perspective on the implementation of semantic web technologies in MSE in current and future projects.

3. Top-Level and Mid-Level Ontologies

For discovery, retrieval, exchange, as well as integration and analysis between different networks, information systems, and applications, heterogeneous data, as well as their contextual information, need to be converted into a computer-processable language.[15] Ontologies play a central role in this process. They represent a way to describe entities (classes and instances) from a certain area of reality, as well as their relationships to each other, in formal language. The explicit definition of clearly defined terms, their meanings, and their relations to each other make it possible to store data and information in a semantically organized way and represent knowledge sustainably.[37,38]

Ontologies are developed for a wide range of different use cases; however, they usually don't have reusability in mind, ending up with monolithic, single-purpose ontologies. To enable a more structured development process and better reuse, modular/layered ontology architectures have been developed. Depending on the degree of detail and the formal expressiveness, different types and levels of ontologies can be distinguished: upper-, foundational-, or TLOs, MLOs, and domain ontologies. TLOs represent the most abstract, independent level. Intending to interconnect as many ontologies as possible, the design incorporates universal and fundamental concepts to ensure expressiveness and generality across a wide range of domains. MLOs are based on TLOs and enhance their structure. Entities are represented more fine granularly, and they have meaning in different domains. MLOs also bridge the gap between TLOs and domain ontologies by providing a set of terms that can be shared among multiple domains, allowing for a higher degree of modularity. Domain ontologies are very expressive and extend the provided concepts of MLOs with specific terms of the domain to be represented.[39,40] Ontologies are widely developed and used since a long time in life sciences. The controlled vocabulary provided by them, together with standard identifiers and relationships, enables unified semantic annotation and description of a biological object within a certain domain. This allows integrated analyzes and interpretations of multimodal data.[41,42]

The Gene Ontology (GO) initiated by the Gene Ontology Consortium in 1998 is one of the most successful ontologies so far. Three independent aspects form the cornerstones of the GO to describe the knowledge of the biological domain: Molecular Function, Cellular Component, and Biological

Process. The overall goal is to develop an "up-to-date, comprehensive, computational model of biological systems from the molecular level to larger pathways, cellular and organism-level systems."[43] The success of GO is based on the fact that experimental knowledge can often be transferred from organism to organism, especially if they share relevant genes due to common ancestry. It all started with a common classification scheme for the gene functions of three model organisms, *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse), and *Saccharomyces cerevisiae* (brewer's or baker's yeast). Today, it is possible to compare homologous gene and protein sequences across the phylogenetic spectrum of over a thousand organisms.[43,44]

To enable uniform development and integration of different ontologies from the biological domain, the Open Biological and Biomedical Ontology (OBO) Foundry was created. The ontologies of the OBO, the GO is a part of, follow certain guidelines and principles. This is to ensure that the ontologies are interoperable and logically formed while representing biological reality as accurately as possible.[45–47]

In addition to the OBO, several other MLOs and domain ontologies refer to the Basic Formal Ontology (BFO) which was initiated in 2002. It is a well-supported and used TLO that has generalized the concepts of the GO. The BFO is specifically designed as a developmental framework and connecting point for science-based domain ontologies. The compact design (which does not include physical, chemical, biological, or other specific scientific terms) makes the BFO consistent with other TLOs. The crossinteroperability of the BFO results from the underlying definition and partitioning into the most general categories of entities: continuants and occurrents. Continuants are persisting entities (uniquely identifiable objects about which information is to be stored or processed) that exist at a time, such as 3D persistent objects, for example, the cellular component and molecular

function of GO. Occurrents are entities that occur, such as primarily time-dependent events or processes that are conceived in successive phases or occur at a particular time interval, for example, the biological process of GO. As granular extension through midlevel and domain ontologies is done in a top–down approach, it is important to understand the semantic framework of the BFO architecture.[48–50]

The European Materials Modelling Ontology (EMMO) is the result of efforts by the European Materials Modelling Council (EMMC) to present a standardized ontology framework for describing, processing, characterizing, and modeling materials and their properties. This successful upper- and midlevel ontology is based on a physical and materials science worldview developed in a bottom-up approach, in contrast to the BFO. Thus, low-level concepts, for example, from the perspective of experimental physics or scientific application, serve to develop the upper conceptual layers of the ontology. This ensures that the TLO concepts of EMMO can be understood by users without a philosophical background.[51,52]

Other important TLOs and initiatives include the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), the General Formal Ontology (GFO), the Suggested Upper Merged Ontology (SUMO), and the Industrial Ontologies Foundry (IOF) initiative.[53–60]

Basically, the main use of TLOs is to enable semantic interoperability of ontologies across multiple domains. As an architectural framework, TLOs provide general concepts that are common to all domains and thus are particularly useful as an ontological blueprint in development. Following the inheritance principle, constraints are propagated to the domain level (**Figure 1**). Thus, TLOs provide a proven means for verifying basic ontological relationships. In principle, domain ontologies oriented to the same TLO are compatible with each other.
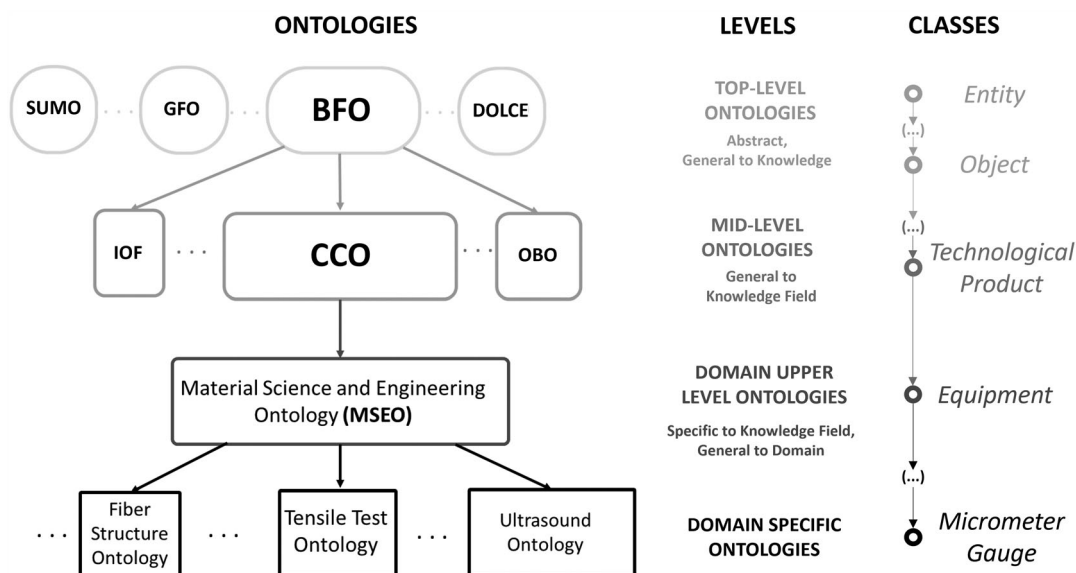

Figure 1. Hierarchical representation of relevant ontologies. A higher position in the representation means a higher level of abstraction. The examples of ontologies shown in the two lower levels are currently being developed by the Mat-o-Lab initiative. Using the example of the micrometer gauge, a device used to accurately measure specimen dimensions in the tensile test, specific-related entities are assigned according to the ontology levels (inheritance principle).

Noticing problems in the adoption and reuse of ontologies in the industrial domain, the Industrial Ontologies Foundry (IOF)[61] was created as a counterpart to the OBO Foundry. Developed by some of the same initiators, it adapts the OBO ontology architecture to the requirements of the industrial and engineering domain using the same open, community-based approach that has led to OBO's success. The IOF promotes the design and development of industrial ontologies by providing and formulating common principles, guidelines, and best practices. In addition, working groups take care of the maintenance, updating, and documentation of designated ontologies, which will also support the creation of domain-specific standards. The IOF ontologies are organized similarly to OBO ontologies, with top- and midlevel ontologies, with the BFO as the designated TLO, and the common core ontologies (CCO)[39] among others, forming the midlevel. The IOF goes one step further and splits the domain ontologies into the two subcategories, "domain upper level" and "domain-specific ontologies," to both drive the development of modularizable domain ontologies and ensure interoperability and reuse of these. This subdivision is a consequence of the ontology architecture that introduces a dedicated and abstract mid level. Other ontologies such as EMMO or MatOnto[62] do not require this subdivision as their midlevel is not abstract but domain specific.

While a certain number of developed ontologies from the field of MSE already exist,[25] it must be stated that that many of these material ontologies cannot be reused because they are not based on the BFO, and thus compatibility issues with the IOF would arise.

The ontology architecture used in the Mat-o-Lab project follows IOF approach due to its open nature, extensive documentation, high granularity, and expressivity of the ontologies. Accordingly, Mat-o-Lab uses the BFO as the top level and the CCO as the mid-level (**Figure 1**).

The midlevel of the CCO supports the representation of a wide range of entities. While the chosen level of abstraction does not permit direct annotation of MSE data, it does allow for the simplified development of domain ontologies as extensions of it. By following principles, the CCO stack provides a vocabulary that can be used to express complex relationships while remaining interoperable.[63] As an extension of the CCO, the Materials Science and Engineering Ontology (MSEO)[64] was developed, which is a domain upper-level ontology for materials science. Similar to the IOF approach, different working groups within Mat-o-Lab are involved in the development of MSE domain-specific ontologies, such as the Tensile Test, the Ultrasound, or the Fiber Structure Ontology. With respect to extensibility, reuse, and interoperability, the approach of designing the ontologies as independent modules is also pursued here. In the long term, these ontologies will be extended by others from the fields of mechanical testing and material structure. An important task will be to update, maintain, and add to them as necessary, with an emphasis on standardization and reuse by the community.

4. Domain-Specific Ontologies: Development and Application

Domain ontologies are used to describe organized and structured material science knowledge in such a way that it can be read, understood, and processed by computers. Specific material science methods and processes as well as characterization paths can be mapped by corresponding material concepts and their relationships to each other. This allows material research data and its related information to be semantically organized to enable knowledge representation. At this ontological level, material concepts acquire high expressiveness using material science domain-specific vocabulary, definitions, facts, statements, axioms, rules, and relations.[17] A basic requirement for the accurate representation of domain ontologies and the specific knowledge is therefore the involvement of MSE experts in the otherwise heavily computer science-driven development process.

Several guidelines are available that are recommended to be used as a basis for the creation of ontologies, the so-called "Ontology Engineering Methodologies" (OEngMs).[65] Therein, different stages, tasks, and actors as well as sequences of their interaction are defined with respect to ontology development. There are a variety of typical OEngMs available.[66]

The explicit definition of classes, properties, and instances as well as the transformation to the formal ontology language is possible with tools such as Protégé developed by Stanford University.[67,68] For instance, Protégé allows the construction of a hierarchical order of concepts and classes and the definition of rules in a neatly arranged structure to include knowledge about the relations of concepts. After having the ontology available, experimental data can be mapped accordingly. Thus, raw data, characteristic values, metadata, and data gained from simulations are stored following the ontology, that is, they will be able to be queried with explicit references. Tools such as OpenRefine[69] and OMERO[70] are suitable for such mapping procedures as they allow the annotation of experimental data in all common data formats. As ontology development is an innovative discipline, especially in the field of MSE, useful tools and converters continue to be generated.

Following the procedure for MSE domain ontology development in Mat-o-Lab (**Figure 2**), which is in accordance with typical guidelines, information on and parameters of the materials and processes to be described need to be collected, first. In this respect, standards and norms, scientific literature, manuals, and for example, also the header information of experimental data files are sources of information and knowledge. In addition, expert knowledge gained from interviews with scientists, engineers, and technicians is crucial to include their valuable experience. Due to the nature of standards to name and define terms, symbols, and processes in a specific context, they comprise information on syntax, semantics, and taxonomy that was furthermore already agreed on by a group of experts in a certain field (standardization committee). Therefore, the usage of standards as a starting point for the identification and definition of entities in a distinct domain such as, for example, the description of a test method, is considered to be a reasonable approach to facilitate the development of ontologies. Standards also provide an orientation for the structuring and categorization of entities and concepts that are represented by classes in the ontology.

Based on datasets resulting from specific processes and experiments, a process template is generated in Mat-o-Lab. This is enriched by valuable contextual information, so-called metadata (e.g., specifications of machines and software, project information, laboratory equipment, etc.). This results in a
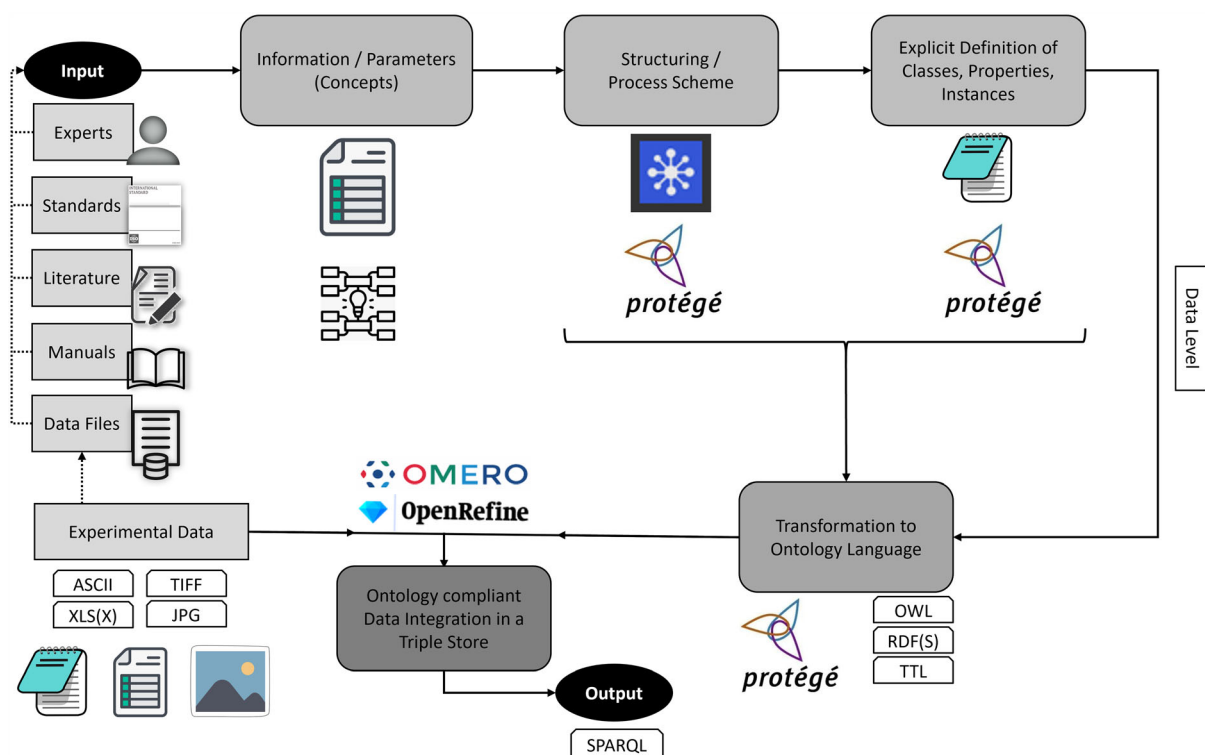
Figure 2. Procedure for domain ontology development within Mat-o-Lab. Reproduced with permission. KnowledgeBaseBuilder logo, 2022, Ingo Straub Softwareentwicklung; Protégé logo, 2022, The Board of Trustees of the Leland Stanford Junior University; Omero logo 2022, University of Dundee & Open Microscopy Environment; and OpenRefine logo, wikimedia.

human- and computer-readable representation of, for example, the process or the experiment. A variety of tools is available for ontology creation that support developers in collecting and visualizing information, drafting process schemes and templates to be represented, and transforming knowledge into the final ontology using proper formal language. The collection of concepts and parameters as well as their categorization and structuring in natural language can easily be performed using the well-known spreadsheet or mind managing software. For the creation of process schemes, the KnowledgeBase Builder by InfoRapid[71] is one well-suited option as it allows for a taxonomic representation of processes in which entities, classes, and properties are marked, respectively. Converter apps developed in Mat-o-Lab then facilitate the enrichment of the created process schemes with uniquely assigned ontology terms, the creation of unified process templates, the instantiation of classes and properties, and the assignment of values and information to corresponding ontology entities.

The development of domain ontologies enables the integration of experimental data and the associated contextual information, that is, the metadata, in data repositories. The structured and stringent mapping of various methodological procedures in process templates creates a digital, comprehensible, human-, and machine-readable representation of materials science reality. This standardizes data formats, supports further use and thus comparability, and paves the way for the creation of automated data pipelines. Furthermore, the high degree of granularity offered by the chosen ontology architecture together with

the semantic expressivity and background knowledge emended in the standards-based ontologies developed in the project enable the execution of logical rules, that are implemented in the form of SHACL shapes and constraints in the pipeline, and perform various validation tasks.[72] These tasks include the verification of adherence to the standard, validation of the correspondence of the values to the measurement units and ranges, detection of missing values or metadata, as well as an overall dataset quality estimation.

For the advancement of ontology development in the field of MSE, there is a need for comprehensible, application-related use cases. Within the Mat-o-Lab projects, dedicated examples are considered in the first stage, including, for example, the detailed digital representation of the classical mechanical tensile test. This widely used and well-standardized test method, with which comparable material parameters are generated, has the potential to appropriately convey the profitable benefits and advantages that arise using ontologies.

5. Concept of Data Framework and Data Space

The aforementioned strategies ensure that datasets of a standardized content and structure are created. However, their integration and application require a dedicated infrastructure in which datasets can be created, curated, aggregated, and analyzed. The concept of such a data framework and how it might be integrated into a decentral data space is outlined in **Figure 3**.

Starting with referenceable raw data with a known uniform resource identifier (URI) or, more specifically, in this case, a
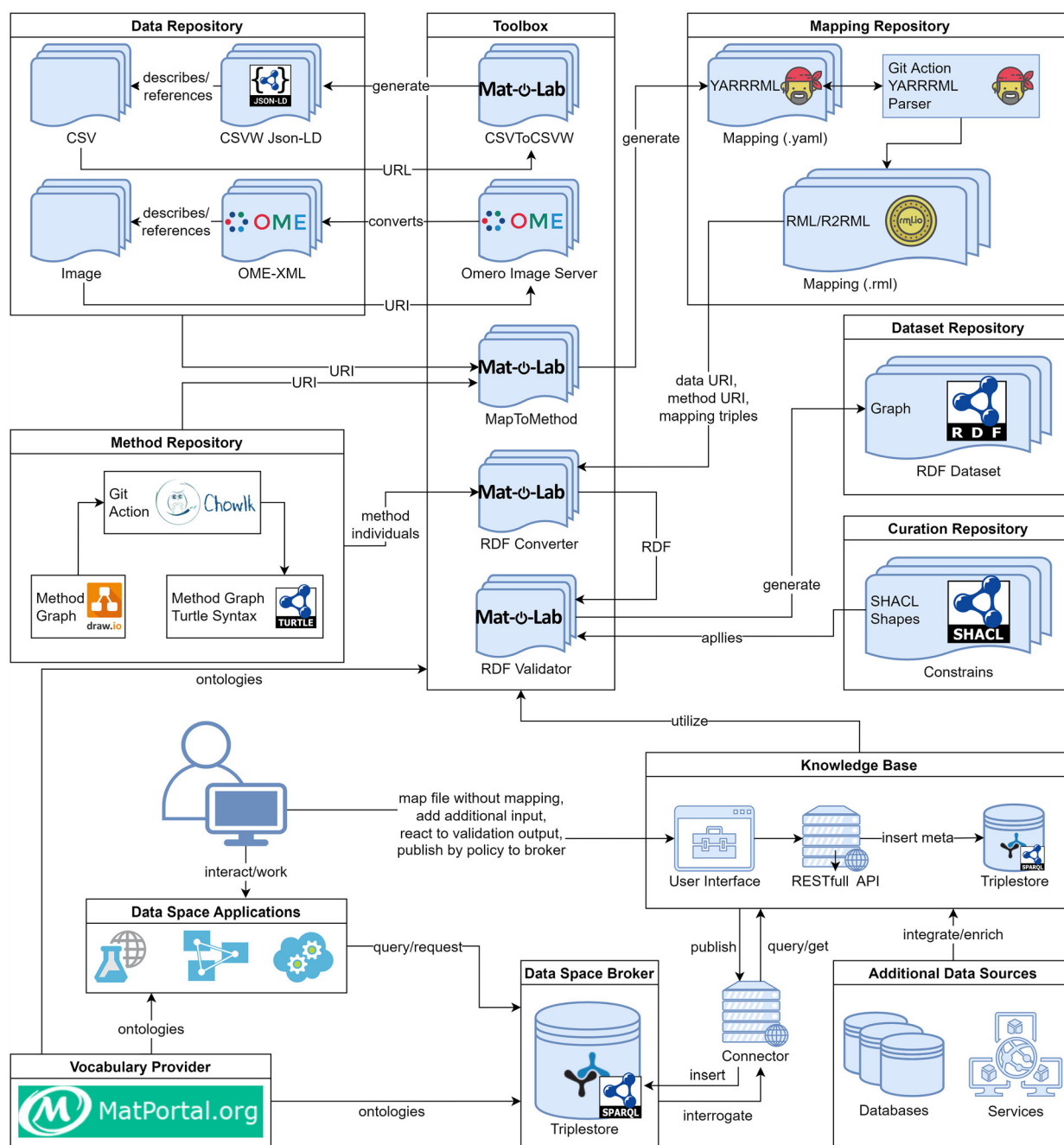
Figure 3. Data framework and data space integration of the Mat-o-Lab initiative. Reproduced with permission, RML.io, 2022, IMEC/Ghent University; OME logo under CC BY 4.0, RDF; JSON-LD logo under CC0; YARRRML logo under CC BY 3.0; draw.io logo under CC BY 4.0; and Chowlk logo under Apache License 2.0.

uniform resource locator (URL) file in a repository, it allows metadata to be created that further describes the origin and the context of those files. Mat-o-Lab supports comma-separated value (CSV) files and image data in various formats, which represent a large part of the file types from the MSE domain.

The CSV files provided are usually very inhomogeneous in their structure and the data presented. They may differ, for example, in the encoding and delimiters for the columns, the decimal format, the amount of header information, and often even

contain valuable information in an additional header row before the tabular data (i.e., the actual content of the CSV files). Therefore, metadata is generated to describe this additional information and provide context for reading the CSV files using a tool from our toolbox called CSVToCSVW.[73] It utilizes the vocabulary for CSVs on the web (CSVW) provided by W3C to create individual metadata in JavaScript Object Notation for Linked Data (JSON-LD) format, without any further input from a user.

Image data, in turn, are converted to OME-TIFF as intermediate format by extraction of metadata of various proprietary image formats used in the microscopic domain with the help of the bioformat library.[74] As a community developed tool, it allows reading more than 140 proprietary formats. The services can be used as part of an image server application called Omero,[70] which has an integrated web application and viewer and provides static URLs for all imported images, including several API endpoints where metadata can be extracted.

To enrich the data, the metadata extracted from the files is combined with method graphs that represent the specific chain of processes, equipment, specifications, and objects relevant to producing the results. Because of the diversity of methods, a toolchain is provided for domain experts to draw these method graphs using freely available software. The graphs are constructed in draw.io,[75] which can be used at the public web platform.[76] The shapes necessary are provided by a library of the Chowlk Converter.[77] The resulting xml files can then be converted to turtle syntax using a web service,[78] by running the python code, or with a runner of a GitLab repository.[79]

The information contained in the metadata file and the method graph are then mapped to each other using the MapToMethod[80] tool. The tool's simple UI allows a user to point to a metadata file with a URL and select the method graph to map against. The use of CCO concepts allows querying for all information artifact bearing (IAB) instances in the metadata file and information content entities (ICE) in the method graph. UI elements are provided to the user, and an inherent mapping IAB can be selected for each ICE. The result is a YARRRML[81] file that captures the mapping in the form of mapping rules in a human-readable form. The file can be easily converted to RML,[82] the less readable but state-of-the-art mapping language of the sematic web, or to JSON-LD. The RDF data, from the metadata file, the method graph, and the mapping will then be merged into an RDF dataset using the outlined, but not yet fully developed, RDF converter.[83] In the process, the resulting RDF is to be curated. The design foresees using SHACL shapes[72] for this purpose. If the validation is successful, the RDF file is placed into a repository where it can be easily processed by a knowledge base triple store.

In this way, very heterogeneous data files from the domain can be combined into a central "Knowledge Base." The figure shows how this fits into a data economy from the perspective of a peer participating in this data space. The knowledge base would be exposed to the users by a UI where they will interact with the Toolbox to create their data pipelines. At this stage, it is useful to integrate data from external databases (e.g., a legacy database) to provide additional information, such as measurement devices, locations, business units, or costs. This data may not be made public but must be tracked to ensure a high level of data interoperability.

The developed domain ontologies described in the previous chapter are made available to all data space participants by a "Vocabulary Provider" via the ontology portal MatPortal.org.[84]

Utilizing the International Data Spaces (IDS) architecture[85] prevents any data leakage/loss and ensures full data sovereignty of the data providers. In this approach, each peer hosts its own triple store, containing the datasets along with metainformation, in the data space. Its participants are identified and authorized by a central identity provider, while all datasets are fully traceable by their unique resource identifiers (URIs). Datasets can be published by exporting the selected data to a central data broker of the dataspace through the Connector by selecting the applicable publication policy (e.g., public, specific peer, etc.).

The "data space broker" instance, to which all peers have access, transmits only the approved data according to policy over a secure transport layer. This makes it possible to query, aggregate, and request another peer's datasets via certified "data space applications" if the corresponding permission is available. Using this functionality, specific applications and digital workflows can address individual problem solving and data analysis issues by submitting appropriate automated queries to the data space broker. The user, as a data space peer, will have access to all data available to him in the broker, as well as all data in his knowledge base, as if interacting with a central data store and not a federated data space.

6. Scientific Simulation and Data Processing Workflows

In recent years, the development in scientific computing has been very dynamic and there are many different specialized tools and procedures available. Analyzing and processing scientific data generally is a complex task that often involves many different steps to be performed in sequential order. This includes pre- and postprocessing of experimental data (e.g., the generation of RDF data), subsequent analysis of multiscale simulation models, or ML types of procedures that are usually based on different models that are coupled either sequentially or even are called in a cyclic order. The challenge is to combine these tools into an automated, reproducible workflow that can be shared with other researchers with complete documentation of all intermediate results, because only then the complete reproducibility of all data stored as an output of such a workflow can be assured. This could be a pure experimental data processing workflow that performs the annotation of the data or a complex simulation workflow coupling different simulation models.

In this context, a workflow is a combination of different software modules that are independently developed and can be combined sequentially, in parallel or in a cyclic order, with dependencies of their inputs/outputs and together provide new scientific knowledge.[86,87] The connection and execution of these individual modules is performed in a workflow management system (WMS). The requirements for a workflow system can be summarized as follows.[86] 1) Data exchange between and within the different modules; 2) embedding heterogeneous modules into a common framework with a definition of relations/dependencies including a portable data structure; 3) hierarchical allocation and provisioning of resources (ranging from individual jobs of a single user up to the coordination of all users with all their jobs on a cluster); 4) execution and scheduling including task launching and data transfer must be coordinated. In particular, this includes observational workflows with scheduling instruments and experimental workflows with manual tasks (e.g., lab tests that are not automated or require some manual input); and 5) tracking the provenance and monitoring and

validation of the data workflow as well as the simulation processes.

The definition of the dependencies between different modules is often defined through a direct acyclic graph (DAC). However, such a static definition might lead to limitations that could be avoided by dynamic generation of the dependencies, taking into account intermediate results.[88] This is, from our perspective, most relevant for situations, where callback functions are used, for example, in a (Bayesian) optimization of model parameters. The steps within the optimizer depend on the forward model to optimize and vice versa; thus, there is cyclic dependency that is difficult to capture with a DAC. A difference between existing WMS is the UI. Most WMS define the workflow through a scripting language that, based on a preprocessing step, builds the computation graph.[89–92] Other tools such as AiiDA[93] directly provide an interface to the programming language. Other options are Jupyter notebooks as used, for example, in Pyiron,[94] where individual modules are directly implemented in the workflow environment and then connected via a notebook that can be shared and executed online. In a similar approach, AiiDAlab[95] provides a cloud platform to set up, exchange, and execute workflows. In addition, Snakemake[96] is also a popular workflow system that also provides support for execution of steps on external HPC resources.

One example of a scientific workflow is the generation of a reproducible journal paper that publishes the results of multiple simulation workflows. An example of a paper workflow can be downloaded via Zenodo.[97] This includes a hierarchical Pydoit-based workflow implementation[98] that covers the complete generation of the paper, from input parameters over all simulation steps up to the postprocessing of individual graphs as well as the LaTeX compilation of the paper. The compute environment is provided via a docker container; alternatively, Conda and PyPy can be used, but they do not provide information on the operating system itself. An important feature from our perspective is caching and only selective recomputation of steps whose dependencies have changed. This is because the main effort in this use case is setting up the workflow with iterative modifications within some processing steps of the complete workflow. Having set up a workflow system right from the beginning ensures data transparency and reproducibility in accordance with typical guidelines on scientific publishing.[99]

A challenge for the future is the integration of standardized ontology-based data nodes that can be used by multiple packages which require the same inputs and outputs (with the structure defined by an ontology). An interesting approach is the common workflow language[100] to define software- and hardware-independent portable workflows. An overview of workflow tools with examples and a discussion of their advantages is published in the study by Ashby et al.[101]

7. Parameter Identification and Model Calibration of Physics-Based Simulation Models

Generalizing experimental information to other setups can be either done by a pure data-driven approach, that is, interpolating or extrapolating the data using ML approaches to other situations not considered in the experiment or by generating a physics-based simulation model that correctly represents the problem at hand. As there are often only a limited number of datasets available, the latter approach usually has the advantage that the number of free parameters in these models is rather limited and thus the number of data points required to calibrate/train these models is significantly reduced compared with a purely data-driven approach. On the other hand, the physics-based modeling assumptions are constraining the solution space and might not adequately represent the real physics; thus, there is always a model bias. This was expressed by George Box with the famous phrase "All models are wrong, but some are useful,"[102] which is particularly true for situations that are related to cracking, failure, damage, or other "extreme" conditions (high temperature, pressure, etc.) for the material.

Consequently, any model prognosis is naturally not a deterministic value, but rather a probability distribution. Model calibration in form of identification of model parameters by numerically solving inverse problems with Bayesian methods or classical regularization has a long tradition.[103] The idea of Bayesian inference is to combine prior knowledge, for example, using a probability distribution both on the parameter level or potentially also on the model level together with a likelihood function that characterizes the conditional probability that observed measurements were created from the model given the parameters. Using Bayes theorem allows computing posterior estimates of the model parameters that could subsequently be used for prognosis purposes. However, there are several challenges related to this approach. 1) The computation of the posterior is usually intractable.[104] Therefore, approximation methods such as Markov Chain Monte Carlo methods (MCMC) are often used. There are a variety of different sampling schemes available[105] and implemented in different software packages. We are using Emcee,[106] Pyro,[107,108] and PyMC3[109,110] that differ in the way samples are generated, if derivatives of the likelihood function can be computed, what is to be computed (e.g., posterior parameter distribution, model evidence, point estimates vs. distributions), and potentially including additional prior knowledge; 2) The computational effort of a single forward model evaluation is often computationally expensive; higher-order derivatives might not be available. Consequently, a specific response surface based on the proper general decomposition is developed.[111] This allows to perform a (computationally more demanding) computation before the model calibration task and build an Abaqus of the full model. In the subsequent model calibration task, the full model is replaced by the Abaqus that can be evaluated in real time. 3) Neglecting model discrepancies and calibrating insufficient models lead to biased parameter estimates and model predictions.[112] Therefore, we propose to identify model bias by adding additional terms to the model with an automatic relevance determination (ARD) prior in combination with variational approaches to select only the relevant terms that better explain the data.[113] This is an iterative process allowing the user to subsequently improve the models. 4) In many situations, it can be advantageous to include additional knowledge about the simulation model into the calibration procedure. This is particularly true for situations, where a significant model bias is expected (as is often the case for failure models due to an insufficient model in these extreme scenarios or due to randomness in crack

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**

www.aem-journal.com

localization). Applying methods such as FEMU-F[114] with stochastic models[115] allows improving the quality of the calibration. The essential idea is to interpret the finite element model (FEM) solution as a stochastic variable (or random field when taking into account correlations) that is identified. Besides the terms related to the discrepancy between the measured and identified fields, the likelihood contains terms related to the residual of the underlying partial differential equation. 5) Correlations both in the data and in the parameters are present that must be considered but are often difficult to be described objectively. The correlation structure can be characterized by additional hyperparameters that are simultaneously identified together with the model parameters. An alternative is to use the model evidence (e.g., computed using nested sampling[116] or approaches based on the evidence lower bound which are often directly computed in variational approaches) and compare different choices for these hyperparameters.

Finally, the result of a model calibration using Bayesian inference is often reused in subsequent computations (e.g., when using the calibrated parameters of a model calibrated with lab data to predict properties for an industrial use case). As a result, it is of utmost importance to store in a database not only the final result (i.e., a posterior distribution of the model parameters) but to also include the characteristics of how those parameters have been obtained. This includes, for example, the (experimental) data, the models including the compute environments within a scientific workflow, the definitions of priors and likelihood in a probabilistic graph, as well as the inference engine used. Consequently, an ontology to describe this calibration process is developed within Mat-o-Lab to characterize the calibration process together with the model parameters and ultimately allow publication of all this data to describe the complete data provenance.

8. Toward a Platform-Supported, Data-Driven Research Approach in MSE

ML-based analysis of materials data has shown great success in predicting material properties for a variety of materials, from ordinary Portland cement[117] to aluminum alloys[118] and superconductors,[119] to name a few. The key concept is to match empirically observed sample characteristics with experimentally confirmed (or simulated) materials properties. ML exploits patterns and correlations influenced by several, intermixed theoretical concepts and multistage, complicated processes in the data, that cannot be described with a closed formulation.[120]

Specifically, the task of ML is often to find desired materials properties in a high-dimensional search or discovery space (DS) spanned by millions of possible material mixtures, of which only a small fraction of material properties has been experimentally explored.[121] The challenge for the ML model is to make the best use of the limited knowledge from the few available data points to effectively explore the DS. Ultimately, closing the loop between exploration with very little available data (as is often the case when faced with a novel research problem), through ML and leveraging knowledge from ontologies, could enable the next generation of self-improving materials research with ML.

Cyber infrastructures, such as Mat-o-Lab, can play a central role to form a shared intelligent materials data ecosystem. They provide digital tools to enrich data with information along the process of materials science knowledge creation (including the most detailed parts of the domain expertise) to create increasingly detailed data that seamlessly integrates with ML tools.

However, many ML models require large amounts of data for making predictions. For new materials, hundreds to thousands of laboratory tests would be required. The generalizability of existing models to new materials outside of the training data is additionally challenging.[122] These circumstances are a great obstacle for practical applications of ML in MSE. Using the broad information from only a few samples remains one of the central challenges.

To get by with less data, sequential learning (SL) is frequently recognized as having great potential to accelerate materials research.[121,123] Instead of making accurate predictions about a particular set of experiments, SL sorts them according to their expected utility. This is done by coupling the predictions of a ML model with a decision rule that guides the experimental procedure. Each new experiment is selected to maximize the amount of useful information, for example, according to Lindley,[124] using previous experiments as a guide for the next experiment. The underlying idea is that not all experiments are equally useful. Some experiments provide more information than others. In contrast to the classical design of experiments, where (only) the experimental parameters are optimized, the potential outcomes of the experiments themselves are the decisive factor.

SL was used for experimental material data from the laboratory, for example, for the search for suitable alloys or cement.[125–127] However, almost exclusively examples exist in the literature that show the potential of SL methods based on simulated experiments, where the ground truth labels for all data points are already known. Yet even in simple statistically motivated scenarios, specific performance is usually highly dependent on the data and the problem, that is, what information is available and what is the sought-after target. The exact relationships are still largely unknown.[128]

In practice, data from multiple experiments often have gaps, for example, when experimental design or resource availability do not allow data collection. This does not necessarily mean that the data are incomplete but is due to the rationale that the cost and effort of data collection must be proportional to the expected gain in knowledge, that is, data from which no gain in knowledge is expected are not collected. However, most ML approaches require invariant tuples and are not designed for constantly changing scenarios. The relatively new field of graph-based ML overcomes this by capturing the data in its context as a graph. This has led to elegant solutions for predicting the properties of crystals[129] or reducing the error of concrete properties predictions with incomplete data.[130] Work from other fields (toxicity prediction,[131] zero-shot image classification,[132] transfer learning[133]) shows that ontologies can be used as the underlying graphs. The integration of structural information from ontologies in ML has the crucial advantage that empirical observations are not required for each case under investigation, but existing knowledge that has been modeled into the ontology can be transferred.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**

www.aem-journal.com

In conclusion, ML and SL benefit from material data platforms such as Mat-o-Lab as a source of high quality and detailed data. In addition, ontologies as comprehensive knowledge frameworks could play an important role in developing the next generation of intelligent algorithms that are able to better process incomplete data and novel scenarios.

9. Summary and Outlook

In the present work, the state of the art and current perspectives of digital knowledge representation in MSE were discussed, yielding the following conclusions: 1) Innovation in MSE is fostered when materials data with their contextual information is made available, discoverable, interoperable, and reusable for research and industry. 2) The initiative Mat-o-Lab provides the potential of establishing a newly developed collaborative framework for data-driven materials research and engineering. 3) The conceptualization, collaborative work structure, and technical implementation of the Mat-o-Lab platform and its upcoming connections to a data space environment were outlined. 4) Ontologies are a core component as domain expert-guided knowledge representations for semantically structuring data in machine-processable formats, providing a valuable basis for modern data-driven methods of ML, modelling, and simulation. By linking these elements, the Mat-o-Lab concepts support advanced materials research and development approaches.

In a nutshell, Mat-o-Lab empowers the field of MSE by upgrading its (meta)data infrastructure to semantic web standards according to the well-approved linked open data (LOD) principles. In the near future, this bears the potential to fundamentally enhance and accelerate the dissemination of MSE research results. One possible scenario is an effective new route for knowledge distribution as a direct alternative to the current state-of-the-art in which results are published in peer-reviewed articles. In this well-established but somewhat outdated dissemination process, the metadata and the contextualization, interpretation, and annotation of the data are written out in a verbose form.

As the ontologies for experimental tests and experiments developed within Mat-o-Lab include a comprehensive and unambiguous digital representation of the analog process description, they immediately act as a sort of digital twin (digital representation) and can at the same time be translated into scripts for highly automated systems. In the same way, they directly allow a computer to compare experimental results with simulated predictions. Thus, they bear the potential to act as the ideal framework for smart autonomous research, in which robot-supported synthesis with parameters based on in silico simulations is iteratively optimized toward a desired property, which is also determined in the same loop. As discussed earlier in more detail, an unambiguous digital representation of scientific experiments including contextualized metadata and results potentially opens the door to a long-anticipated evolution of scientific publication of results. In accordance with the proposed FAIR data principles, the coming paradigm of knowledge dissemination should encourage the direct publication of data, ontologies, and workflows also outside of scientific articles. The Mat-o-Lab framework allows for such a publication of citable and annotated data including contextualization without having to go the detour of verbose explanatory text.

The short road of LOD history is paved with unique success stories, most of which use the data along a path different from the originally designated purpose. In most cases, the resulting creative and sometimes even unconventional solutions could not be foreseen prior to the provision of the respective data. Just a few examples are social networks, unified traffic services, or legal tech. In the same way, Mat-o-Lab will bring forth collateral, indirect innovations. Potential outcomes are a faster form of patenting novel processes simply by staking the claim on certain combinations of literals within a given ontology framework or fully machine-readable norms in the form of semantic operating protocols. While the first would vastly accelerate research, the latter would help facilitate new technologies into safety-relevant fields of application.

This semantic framework introduced in Mat-o-Lab is not without competition and is only one possible framework. In other data provision technologies, the selection process toward one widely used standard is typically decided by the amount of accessible content while in other technologies the user-friendliness usually plays a pivotal role. A unique asset of the Mat-o-Lab framework is the fact that the orthogonal skillsets of complex semantic database design and complicated domain science are separated by predefining TLOs and MLOs and providing a step-by-step DIY instruction to translate the domain-specific expertise into a domain ontology without any previous training or knowledge in data science. This feature directly addresses the abovementioned quality gates by being user friendly and by inherently carrying the potential to become widely filled by nontrained domain scientists. We are, thus, optimistic that the Mat-o-Lab framework is an attractive pathway to laboratory digitalization and will soon mature to an accepted data handling standard. This progression is facilitated not only by its open structure and context but also by its strong emphasis on coworking. Mat-o-Lab lives a collaborative spirit and is always strongly encouraging peers to join the team. Its agile management allows cooperation partners to contribute not merely content but codetermine the structure of the ontologies and actively shape the framework.

Conflict of Interest

The authors declare no conflict of interest.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**

www.aem-journal.com

[1] M. Ashby, H. Shercliff, D. Cebon, *Materials 4th Edition - Engineering, Science, Processing and Design*, Butterworth-Heinemann, Oxford, UK **2018**.

[2] D. G. R. William, D. Callister Jr., *Callister's Materials Science and Engineering*, 10th ed., Global Edition, John Wiley & Sons Inc, New York, NY **2020**.

[3] J. M. Rickman, T. Lookman, S. V. Kalinin, *Acta Mater.* **2019**, *168*, 473.

[4] K. Rajan, *Annu. Rev. Mater. Res.* **2015**, *45*, 153.

[5] The Minerals, Metals & Materials Society; on behalf of the National Science Foundation (NSF), *Building a Materials Data Infrastructure - Opening New Pathways to Discovery and Innovation in Science and Engineering*, The Minerals, Metals & Materials Society, Pittsburgh, PA **2017**, https://doi.org/10.7449/mdistudy_1, ISBN: 9780692860441.

[6] The Minerals, Metals & Materials Society (TMS), *Advanced Computation and Data in Materials and Manufacturing: Core Knowledge Gaps and Opportunities*, The Minerals, Metals & Materials Society (TMS), Pittsburgh, PA **2018**, https://doi.org/10.7449/coreknowledge_1

[7] T. Ashino, *Data Sci. J.* **2010**, *9*, 54.

[8] A. Prakash, S. Sandfeld, *Pract. Metall.* **2018**, *55*, 493.

[9] M. Heilmaier, M. Zimmermann, J. Seifert, A. Weidenkaff, B. Jahnen, *Digitaler Wandel in der Wissenschaft: Herausforderungen und Chancen für das Fachgebiet Materialwissenschaft and Werkstofftechnik - Anmerkungen der Fachkollegien Materialwissenschaft and Werkstofftechnik der Deutschen Forschungsgemeinschaft*, DFG, Bonn, Germany **2018**.

[10] S. Sandfeld, T. Dahmen, F. Fischer, C. Eberl, S. Klein, M. Selzer, *Digitale Transformation in der Materialwissenschaft and Werkstofftechnik*, Deutsche Gesellschaft für Materialkunde e.V., Frankfurt, Germany **2018**.

[11] A. Agrawal, A. Choudhary, *APL Mater.* **2016**, *4*, 053208.

[12] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Adv. Sci.* **2019**, *6*, 1.

[13] J. Schmidt, M. R. G. Marques, S. Bott, M. A. L. Marques, *npj Comput. Mater.* **2019**, *5*, 1.

[14] J. Kimmig, S. Zechel, U. S. Schubert, *Adv. Mater.* **2021**, *33*, 2004940.

[15] M. Wilkinson, M. Dumontier, I. Aalbersberg, J. Jsbrand, G. Appleton, M. Axton, A. Baak, N. Blomberg, J-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, et al., *Sci. Data* **2016**, *3*, 1.

[16] N. Guarino, D. Oberle, S. Staab, *Handbook on Ontologies. International Handbooks on Information Systems*, Springer, Berlin, Heidelberg. **2009**, pp. 1–17, https://doi.org/10.1007/978-3-540-92673-3_0.

[17] C. Feilmayr, W. Wöß, *Data Knowl. Eng.* **2016**, *101*, 1.

[18] M. Hepp, *IEEE Internet Comput.* **2007**, *11*, 90.

[19] L. Takahashi, K. Takahashi, *Phys. Chem. Lett.* **2019**, *10*, 7482.

[20] J. Glick, *Informatics for Material Science and Engineering* (Eds: K. Rajan), Butterworth-Heinemann, Oxford, UK **2013**, Chapter 8, 10.1016/B978-0-12-394399-6.00008-4.

[21] A. Spear, W. Ceusters, B. Smith, *Appl. Ontol.* **2016**, *11*, 103.

[22] J. Domingue, D. Fensel, J. A. Hendler, *Handbook of Semantic Web Technologies*, Springer Berlin (Verlag), Berlin, Germany **2011**.

[23] S. Ramakrishna, T-Y. Zhang, W-C. Lu, Q. Qian, J. S. C. Low, J. H. R. Yune, D. Z. L. Tan, S. Bressan, S. Sanvito, S. R. Kalidindi, *J. Intell. Manuf.* **2019**, *30*, 2307.

[24] J. Davies, R. Studer, P. Warren, *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*, John Wiley and Sons, New York, USA **2006**.

[25] X. Zhang, C. Zhao, W. Xiang, *Comput. Ind.* **2015**, *73*, 8.

[26] S. Zhao, Q. Qian, *AIP Adv.* **2017**, *7*, 105325.

[27] D. L. McDowell, S. R. Kalidindi, *MRS Bull.* **2016**, *41*, 326.

[28] National Science and Technology Council, *Materials genome initiative for global competitiveness*, **2011**, https://www.mgi.gov/ (accessed: June 2021).

[29] Karlsruher Institut für Technologie, *Platform MaterialDigital*, **2019**. https://www.materialdigital.de/ (accessed: July 2021).

[30] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, *MRS Bull.* **2016**, *41*, 399.

[31] M. Statt, B. A. Rohr, K. S. Brown, J. S. Hummelshoej, L. Hung, A. Anapolsky, J. Gregoire, S. Suram, *Mater. Chem.* **2021**, *1–10*, https://doi.org/10.26434/chemrxiv.14583258.v1.

[32] E. Soedarmadji, H. S. Stein, S. K. Suram, D. Guevarra, J. M. Gregoire, *Comput. Mater.* **2019**, *5* 1.

[33] I. M. Pendleton, G. Cattabriga, Z. Li, M. A. Najeeb, S. A. Friedler, A. J. Norquist, E. M. Chan, J. Schrier, *MRS Commun.* **2019**, *3* 846.

[34] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, C. Phillips, *Sci. Data* **2018**, *5*, 180053.

[35] *Bundesanstalt für Materialforschung and -prüfung*, **2021**. www.bam.de (accessed: July 2021).

[36] *Fraunhofer Materials" **2021**. www.materials.fraunhofer.de (accessed: July 2021).

[37] G. Stephan, H. Pascal, A. Andreas, *Semantic Web Services* (Eds: R. Studer, S. Grimm, A. Abecker,), Springer, Berlin, Heidelberg **2007**, pp. 51–105, https://doi.org/10.1007/3-540-70894-4_3.

[38] R. Gayathri, V. Uma, *ICT Express* **2018**, *4*, 69.

[39] R. Rudnicki, B. Smith, T. Malyuta, W. Mandrick, *White Paper: Best Practices Of Ontology Development*, CUBRC, Buffalo, NY, USA **2016**.

[40] L. Obrst, *The Ontology Spectrum Semantic Models*, **2006**. https://slidetodoc.com/the-ontology-spectrum-semantic-models-dr-leo-obrst/. (accessed: July 2020).

[41] M. Hartung, T. Kirsten, E. Rahm, *Data Integration in the Life Sciences, DILS 2008. Lecture Notes in Computer Science* (Eds: A. Bairoch, S. Cohen-Boulakia, C. Froidevaux, Vol. *5109*, Springer, Berlin, Heidelberg **2008**, pp. 11–27, https://doi.org/10.1007/978-3-540-69828-9_4.

[42] R. Hoehndorf, P. N. Schofield, G. V. Gkoutos, *Briefings Bioinf.* **2015**, *16*, 1069.

[43] GENEONTOLOGY Unifying, About the GO, **2019**. http://geneontology.org/docs/introduction-to-go-resource/ (accessed: July 2021).

[44] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. Michael Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, *Nat. Genet.* **2000**, *25*, 25.

[45] M. Ashburner, C. J. Mungall, S. E. Lewis, *Cold Spring Harbor Symp. Quant. Biol.* **2003**, *68*, 227.

[46] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, The OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, S. Lewis, *Nat. Biotechnol.* **2007**, *25*, 1251.

[47] The OBO Foundry, *The Open Biological and Biomedical Ontology (OBO) Foundry*, **2007**. http://www.obofoundry.org (accessed: July 2021).

[48] R. Arp, B. Smith, A. D. Spear, *Building Ontologies with Basic Formal Ontology*, The MIT Press, Boston, USA **2015**, p. 248.

[49] C. Partridge, A. Mitchell, A. Cook, J. Sullivan, M. West, *A Survey of Top-Level Ontologies - To Inform the Ontological Choices for a Foundation Data Model*, University of Cambridge, New York, USA **2020**, p. 143, https://doi.org/10.17863/CAM.58311.

[50] A. Ruttenberg, *BFO Basic Formal Ontology*, **2002**, http://basic-formal-ontology.org (accessed: July 2021).

[51] J. Friis, *The European Materials & Modelling Ontology (EMMO)*, **2019**, https://github.com/emmo-repo/EMMO, (accessed: July 2021)

[52] P. Mason, C. R. Fisher, R. Glamm, M. V. Manuel, G. J. Schmitz, A. K. Singh, A. Strachan, in *Proc. of the 4th World Congress on Integrated Computational Materials Engineering (ICME 2017)*, The Minerals, Metals & Materials Society (TMS), Warrendale, Pennsylvania, USA **2017**.

[53] V. Mascardi, V. Cordì, P. Rosso, in *WOA 2007: Dagli Oggetti agli Agenti. 8th AI*IA/TABOO Joint Workshop "From Objects to Agents": Agents and Industry: Technological Applications of Software Agents*, Seneca Edizioni Torino, Genova, Italy **2007**, pp. 55–64, ISBN 978-88-6122-061-4, https://dblp.org/db/conf/woa/woa2007.html.

[54] Laboratory for applied Ontology, *DOLCE: Descriptive Ontology for Linguistic and Cognitive Engineering*, **2002**, http://www.loa.istc.cnr.it/dolce/overview.html (accessed: July 2021).

[55] Ontologies in Medicine and Life Sciences Foundations, *Development and Applications, „General Formal Ontology (GFO)*, **1999**, https://www.onto-med.de/ontologies/gfo#pubs (accessed: July 2021).

[56] H. Herre, B. Heller, P. Burek, R. Hoehndorf, F. Loebe, H. Michalek, *General Formal Ontology (GFO) - A Foundational Ontology Integrating Objects and Processes [Version 1.0]*, University of Leipzig, Leipzig, Germany **2006**.

[57] Ontology Portal, *Suggested Upper Merged Ontology (SUMO)*, **2021**, [accessed July 2021].

[58] I. Niles, A. Pease, in *FOIS '01: Proc. of the international conference on Formal Ontology in Information Systems*, Vol. *2001*, Ogunquit, ME **2001**.

[59] Industrial Ontologies Foundry (IOF), IOF, **2018**. https://www.industrialontologies.org (accessed: July 2021).

[60] B. Kulvatunyou, M. Lee, M. Katsumi, in *Industrial Ontology Foundry (IOF) - Achieving Data Interoperability Workshop*, Online, **2020**.

[61] M. Karray, N. Otte, R. Rai, F. Ameri, B. Kulvatunyou, B. Smith, D. Kiritsis, C. Will, R. Arista, in *Industrial Ontology Foundry (IOF) - achieving data interoperability Workshop, Inter. Conf. on Interoperability for Enterprise Systems and Applications*, online, **2021**.

[62] K. Cheung, J. Drennan, J. Hunter, in *AAAI Spring Symp. - Technical Report*, California, USA **2008**.

[63] R. Rudnicki, *An Overview of the Common Core Ontologies*, CCO White Paper, pp. 1-27, **2019**.

[64] Mat-o-Lab, MSEO, **2021**, https://github.com/Mat-O-Lab/MSEO (accessed: November 2021).

[65] N. F. Noy, D. L. McGuinness, *What Is an Ontology and Why We Need It*, https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html, (accessed: April 2021).

[66] B. Moreno Torres, C. Völker, S. Nagel, T. Hanke, S. Kruschwitz, *Remote Sens.* **2021**, *13*, 13122426.

[67] Stanford University, Protégé, **2021**. https://protege.stanford.edu/ (accessed: July 2021).

[68] M. A. Musen, *AI Matters* **2015**, *1* 4.

[69] D. Huynh, OpenRefine, October 2012. https://openrefine.org/. (accessed: July 2021).

[70] Environment University of Dundee & Open Microscopy, Omero, **2005**. https://www.openmicroscopy.org/omero/ (accessed: July 2021).

[71] I. Straub, *InfoRapid KnowledgeBase Builder Web Edition*, **2014**. https://inforapid.org/webapp/login.php (accessed: July 2021).

[72] W3C, *Shapes Constraint Language (SHACL)*, https://www.w3.org/TR/shacl (accessed: November 2021).

[73] Mat-o-Lab, *Mat-o-Lab Repository – CSVToCSVW*, https://github.com/Mat-O-Lab/CSVToCSVW (accessed: November 2021).

[74] Openmicroscopy, *Openmicroscopy - Bio-formats*, https://www.openmicroscopy.org/bio-formats (accessed: November 2021).

[75] Draw.IO, Draw.IO Repository, https://github.com/jgraph/drawio (accessed: November 2021).

[76] Draw.io Public Web Platform, https://app.diagrams.net/ (accessed: November 2021).

[77] Chowlk Converter, Chowlk Repository, https://github.com/oeg-upm/Chowlk (accessed: November 2021).

[78] M. Poveda-Villalón, S. Chávez-Feria, Chowlk Converter, Chowlk, **2021**. https://chowlk.linkeddata.es/ (accessed: November 2021).

[79] Mat-o-Lab, Mat-o-Lab Repository – MSEO, https://github.com/Mat-O-Lab/MSEO/blob/main/.github/workflows/drawiotordf.yml (accessed: November 2021).

[80] Mat-o-Lab, Mat-o-Lab Repository, https://github.com/Mat-O-Lab/MapToMethod (accessed: November 2021).

[81] imec — Ghent University — IDLab, YARRRML, https://rml.io/yarrrml/ (accessed: November 2021).

[82] IDLab - imec - Ghent University, RDF Mapping Language (RML), https://rml.io/specs/rml/ (accessed: November 2021).

[83] Mat-o-Lab, Mat-o-Lab Repository – rdfconverter, https://github.com/Mat-O-Lab/RDFConverter (accessed: November 2021).

[84] Matportal, Matportal, https://matportal.org/ (accessed: November 2021).

[85] B. Otto, S. Steinbuss, A. Teuscher S. Lohmann, Zenodo, **2019**, https://zenodo.org/record/5105529#.YRNzSYgzbcs (accessed: August 2021).

[86] E. Deelman, T. Peterka, I. Altintas, C. Carothers, K. Kleese van Dam, K. Moreland, M. Parashar, L. Ramakrishnan, M. Taufer, J. Vetter, *Int. J. High Perform. Comput. Appl.* **2018**, *32*, 159.

[87] E. Deelmana, D. Gannon, M. Shields, I. Taylor, *Future Gener. Comput. Syst.* **2008**, *25*, 528.

[88] M. Uhrin, S. P. Huber, J. Yu, Nicola Marzaria, G. Pizzi, *Comput. Mater. Sci.* **2021**, *187*, 110086.

[89] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, S. Mock, in *Proc. 16th Inter. Conf. on Scientific and Statistical Database Management*, IEEE, New York, NY **2004**, pp. 423–424.

[90] D. Barseghian, I. Altintas, M. Jones, in *Proc. of the Environmental Information Management Conf.*, Albuquerque, NM, USA **2008**, https://www.yumpu.com/en/document/read/39431796 (accessed: March 2022).

[91] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, K. Wenger, *Future Gener. Comput. Syst.* **2015**, *46*, 17.

[92] M. Wilde, M. Hategan, J. M. Wozniak, B. Clifford, D. S. Katz, I. Foster, *Parallel Comput.* **2011**, *37*, 633.

[93] M. Uhrin, S. P. Huber, J. Yu, N. Marzari, G. Pizzi, *Comput. Mater. Sci.* **2021**, *1*, 110086.

[94] J. Janssen, S. Surendralal, Y. Lysogorskiy, M. Todorova, T. Hickel, R. Drautz, J. Neugebauer, *Comput. Mater. Sci.* **2019**, *1*, 24.

[95] A. V. Yakutovich, K. Eimre, O. Schütt, L. Talirz, C. S. Adorf, C. W. Andersen, E. Ditler, D. Du, D. Passerone, B. Smit, N. Marzari, G. Pizzi, C. A. Pignedoli, *Comput. Mater. Sci.* **2021**, *188*, 110165.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**

www.aem-journal.com

[96] J. Köster, snakemake, **2020**, https://snakemake.github.io/ (accessed: July 2021).

[97] C. Pohl, J. F. Unger, V. Smilauer, *A three-phase transport model for high-temperature concrete simulations validated with X-Ray CT data*, https://doi.org/10.5281/zenodo.4890635, **2021**.

[98] E. Schettino, Pydoit, **2018**, https://pydoit.org/ (accessed: July 2021).

[99] DFG, *Guidelines For Safeguarding Good Research Practice. Code Of Conduct*, Deutsche Forschungsgemeinschaft, DFG, Bonn, Germany **2019**.

[100] M. R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilovic, C. Goble, https://doi.org/10.48550/arXiv.2105.07028.

[101] BAM GitHub Repository, BAMresearch - NFDI4Ing Scientific Workflow Requirements, BAM, https://github.com/BAMresearch/NFDI4IngScientificWorkflowRequirements (accessed: November 2021).

[102] G. E. P. Box, *J. Am. Stat. Assoc.* **1976**, *71*, 791.

[103] J. P. Kaipio, E. Somersalo, *Statistical and Computational Inverse Problems*, Vol. 160, Springer, New York, NY **2005**, p. 340.

[104] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, *J. Am. Stat. Assoc.* **2017**, *112*, 859.

[105] S. Sharma, *Annu. Rev. Astron. Astrophys.* **2017**, *55*, 219.

[106] D. Foreman-Mackey et al., emcee, **2021**, https://emcee.readthedocs.io/en/stable/ (accessed: July 2021).

[107] PYRO, **2017**, https://pyro.ai/examples/. (accessed: July 2021).

[108] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, N. D. Goodman, *J. Mach. Learn. Res.* **2019**, *20*, 1.

[109] J. Salvatier, T. V. Wiecki, C. Fonnesbeck, *PeerJ Comput. Sci.* **2016**, *2*, e55.

[110] PyMC3, **2018**, https://docs.pymc.io/ (accessed: July 2021).

[111] A. Robens-Radermacher, J. F. Unger, *Adv. Model. Simul. Eng. Sci.* **2020**, *7*, 1.

[112] J. Brynjarsdóttir, A. O'Hagan, *Inverse Probl.* **2014**, *30*, 114007.

[113] M. A. Chappell, A. R. Groves, B. Whitcher, M. W. Woolrich, *IEEE Trans. Signal Process.* **2009**, *57*, 223.

[114] S. Avril, M. Bonnet, A-S. Bretelle, M. Grédiac, F. Hild, P. Ienny, F. Latourte, D. Lemosse, S. Pagano, E. Pagnacco, F. Pierron, *Experimental Exp. Mech.* **2008**, *48*, 381.

[115] L. Bruder, P-S. Koutsourelakis, *Int. J. Uncertainty Quantif.* **2018**, *8*, 447.

[116] J. Skilling, *AIP Conf. Proc.* **2004**, *735*, 395.

[117] W. B. Chaabene, M. Flah, M. L. Nehdi, *Constr. Build. Mater.* **2020**, *260*, 119889.

[118] J. Li, Y. Zhang, X. Cao, Q. Zeng, Y. Zhuang, X. Qian, H. Chen, *Commun. Mater.* **2020**, *1*, 1.

[119] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, *npj Comput. Mater.* **2018**, *4*, 1.

[120] C. Suh, C. Fare, J. Warren, E. Pyzer-Knapp, *Annu. Rev. Mater. Res.* **2020**, *50*, 1.

[121] T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, *npj Comput. Mater.* **2019**, *5*, 2.

[122] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehtaf, L. Ward, *Mol. Syst. Des. Eng.* **2018**, *3*, 819.

[123] K. Reyes, W. B. Powell, SIAM Review (preprint), pp. 1-47, arXiv:2004.05417, **2020**.

[124] D. V. Lindley, *Ann. Math. Stat.* **1956**, *27*, 986.

[125] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, B. Meredig, *Integr. Mater. Manuf. Innovation* **2017**, *6*, 207.

[126] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, *Nat. Commun.* **2016**, *7*, 11241, https://doi.org/10.1038/ncomms11241.

[127] C. Völker, R. Firdous, D. Stephan, S. Kruschwitz, *J. Mater. Sci.* **2021**, *56*, 15859.

[128] Y. Kim, E. Kim, E. Antono, B. Meredig, J. Ling, *npj Comput. Mater.* **2020**, *6*, 131.

[129] S. Sadeed Omee, S-Y. Louis, N. Fu, L. Wei, S. Dey, R. Dong, Q. Li, J. Hu, https://doi.org/10.48550/arXiv.2109.12283.

[130] J. You, X. Ma, D. Y. Ding, M. Kochenderfer, J. Leskovec, *Comput. Sci. Mach. Learn.* **2020**, *15*, https://doi.org/10.48550/arXiv.2010.16418.

[131] E. B. Myklebust, E. Jimenez-Ruiz, J. Chen, R. Wolf, K. E. Tollefsen, *Lect. Notes Comput. Sci.* **2019**, *11779*, https://doi.org/10.48550/arXiv.1907.01328.

[132] Y. Geng, J. Chen, Z. Chen, J. Z. Pan, Z.. Ye, Z. Yuan, Y. Jia, H. Chen, in *Int. World Wide Web Conference WWW '21: Proceedings of the Web Conf. 2021*, Association for Computing Machinery, New York, NY, United States **2021**, pp. 3325–3336, https://doi.org/10.1145/3442381.3450042.

[133] S. Mei, W. Fei, S. Zhou, *BMC Bioinf.* **2011**, *12*, 12.

Bernd Bayerlein has been a postdoc at the Federal Institute for Materials Testing and Research (BAM) in Berlin, Germany, since 2017. He received his Ph.D. in material science from the Technical University of Berlin by studying the structure–property correlation of a mussel shell layer at the Max Planck Institute of Colloids and Interfaces in Potsdam. Currently, the material science expert is promoting digitization in various initiatives (e.g., Platform MaterialDigital, Mat-o-Lab) in interdisciplinary teams, thus working toward realizing the vision of a shared materials data space in accordance with the FAIR principles.