Contextualization for Generating FAIR Data: a Dynamic Model for Documenting Research Activities

Osman Altun 1[0000-0002-6107-2601], Marc Hinterthaner 2[0009-0004-4232-5732],
Khemais Barienti 1[0000-0002-0764-8980], Florian Nürnberger 1[0000-0002-7824-0675],
Roland Lachmayer 1[0000-0002-3181-6323], Iryna Mozgova 3[0000-0002-6761-0220],
Oliver Koepler 2[0000-1111-2222-3333] and Sören Auer1 [0000-0002-0698-2864]

[1] Leibniz University Hannover, Welfengarten 1A, 30167 Hannover, Germany
[2] Leibniz Information Centre for Science and Technology (TIB),
Welfengarten 1B, 30167 Hannover, Germany
[3] Paderborn University, Warburger Str. 100 33098 Paderborn, Germany
iryna.mozgova@uni-paderborn.de

Abstract. The digitization of technologies in product manufacturing results in the availability of large amounts of process and product data. To gain knowledge from this data and fully leverage its potential, its structuring and semantically annotation is essential. This allows preserving the context of data generation and makes the data machine-readable and interpretable. Contextualization is the key to generating FAIR (Findable, Accessible, Interoperable, Reusable) data. The documentation of research activities and provenance of generated data is usually achieved by protocols. However, there is often a tension between the desire to document data generation in a structured, semantically rich form and the need to design research and process parameters flexibly as experimental conditions change.

To resolve these contradictions, a dynamic model is described that allows to document research activities and implemented into a knowledge and research data management system to resolve these contradictions. The model allows a formal, semantic representation of research steps, parameters and gathered data, while also providing flexibility in the generation of protocol templates and individual experiments through the reuse of semantic building blocks. The approach is carried out within the context of a large collaborative research center, showcasing its use in managing and providing data for heterogeneous research tasks, documentation, and data types across interdisciplinary projects.

Keywords: FAIR Data, Semantic Annotation, Data Management, Knowledge Management.

1 Introduction

The development of novel production technologies is often associated with complex, interdisciplinary research questions that are investigated collaboratively across several project teams and from different perspectives in order to be able to penetrate and describe processes to be researched in their entirety. During the research activities, experiments, simulations or observations generate large heterogeneous data sets that increasingly merge with data from other activities into joint analysis processes. A comprehensible and reproducible documentation of the methods, materials and tools used in the execution of research activities as well as documentation of measurement results is therefore of great importance.

Documentation in the form of protocols enables contextualization of the research and the data generated, the why, how, when, where, with what and by whom. Logging is manifold and increasingly digital. Realized as locally stored text documents, Excel spreadsheets, or supplementary ReadMe files to data files, such documentation is loosely linked to the corresponding data, weakly structured and not sufficiently annotated in a machine-readable way. Software tools like digital log books, Electronic Lab Notebooks (ELN) or Knowledge Management Systems improve the structured documentation of research activities. Nevertheless, we observe few accepted standards for documenting research activities in engineering. With the increasing availability and acceptance of public data repositories, at least the standardized provision of data is becoming easier. The structured deposition and description of research activities together with generated data is yet another key to enable reproducibility and reusability of data. Looking at increasing amounts of data available, semantic annotation of data is a prerequisite for machine-readable and -interpretable data enabling machine learning and data-driven research.

However, there is a tension between making the documentation process easy to use and the need to have fixed interfaces to capture structured data. On the one hand, researchers require a user-friendly interface that enables them to quickly and easily document their constantly changing procedures, experimental conditions and parameters without disrupting their research workflow. On the other hand, a fixed interface is necessary to ensure that data is captured in a structured and consistent manner that can be easily analyzed and shared. In the following an approach is described to develop a flexible and extensible framework that allows researchers to document their procedures in a structured and semantically rich way, while also providing the necessary flexibility to accommodate changes in experimental conditions. This can be achieved by a “generic protocol structure” applying semantic building blocks that can be customized to specific experimental procedures and easily assembled to form a structured representation of experiments. The implementation is carried out within the knowledge management system of the Collaborative Research Centre (CRC) 1368 „Oxygen-free production“ [1].

2 Research Data Management in engineering sciences

Data is the foundation of research and science, providing the raw material from which knowledge and understanding are derived. It is used to test hypotheses, develop theories, apply scientific methods, and gain insights. The availability of research data is an important factor for the transparency and reproducibility of scientific results. The FAIR principles describe the framework of how data should be prepared and made available [2, 3]. In recent years, both the availability of data repositories and the portfolio of standards and tools for comprehensive descriptions of data generation in engineering have improved [4]. With the increasing data available the comprehensive documentation of data generation comes into focus as it provides essential context about the research methods and procedures, enabling other researchers to evaluate the validity and reliability of the data and derived results. In addition, documentation of how data was generated can be used to identify potential sources of error or bias, and to improve the design of future experiments. Therefore, the availability of data alone is not sufficient for ensuring reproducibility, as it must be accompanied by detailed documentation of how the data was generated and processed. Software tools such as ELN or knowledge management systems are also increasingly used in the engineering sciences to document the context of data generation in the form of protocols [5, 6]. E. g., the ELN Software elabFTW provides a generic documentation of research activities due to its flexible way to capture experimental procedures, process steps and parameters in less-structured text documents. Additional tools can be applied on such protocols to increase the structure and semantics of information and data [7]. When using semantic tools such as Semantic MediaWiki (SMW) [8] for the documentation of research activities, information and data are basically captured in a structured and semantically annotated form from the very beginning due to the native functionality of SMW [9-11]. Research data can be described applying the generic DataCite metadata schema [12] or using a domain-specific ontology [13]. For the description of the provenance of digital objects the PROV ontology can be used [14]. Terms from PROV have also been imported into the Metadata4Ing ontology for engineering. Metadata4Ing provides a model for documenting research activities and research data in engineering [15]. It has been developed in the context of the NFDI4Ing consortium within the National Research Data Infrastructure (NFDI) in Germany [16]. It enables researchers to document the origin and path of data created or modified during the research process. For this purpose, Metadata4Ing applies a generalized process model centered on the Class *Processing Step*.

3 Problem analysis and research overview

Research problems in production technologies are often complex and require interdisciplinary collaborations among researchers from various subdisciplines within the engineering sciences. The merging and analysis of data and intermediate results from multiple research projects create the need to harmonize not only the data itself, but also the documentation of its generation and processing, in order to fully capture the context of the data. Although engineering subdisciplines have distinct standards and practices for Standard Operating Procedures (SOPs), documenting research activities can be challenging due to the constantly evolving experimental parameters and the absence of predefined standard procedures. To address this challenge, electronic lab notebooks and knowledge management systems are increasingly used to support a fully digitized documentation of experiments and data generation. These tools offer templates that can be used to structure and even semantically describe experiments and gathered data, facilitating the documentation of research activities in a harmonized and structured manner in accordance with the FAIR data principles. In the aforementioned interdisciplinary research projects, standardized templates for protocols are only partially suitable for documentation purposes. Instead, flexible documentation tools are needed that can be easily adapted and reused in parts. However, such documentation should still be structured and semantically annotated, enabling machine-readability and facilitating the combined analysis with large data collections. This allows the documentation to be effectively integrated with other data, enabling joint analysis and interpretation of interdisciplinary research questions. Adapting these ideas to fully digitized processes and linked data it is intended to create a framework of semantically annotated data and documentation of data generation. To address these competing demands, a dynamic model is being developed to describe and document research activities, data generation, and data provenance. This model enables the development and combination of semantic building blocks to create generic protocols. The implementation of this model and the generic protocols is facilitated by a knowledge management system that utilizes the SMW software.

4 Contextualization of data generation

4.1 Requirements

Large Collaborative Research Centres (CRCs) with multiple sub-projects are a good example of how overarching research problems are broken down into smaller research activities. For example, in the development of new or adaptation and modernization of existing production technologies, various sub-projects may examine the same prototype (referred to as "specimen" below). Different parameters (referred to as "variables" below) of both the specimen and the production process are considered using different methods. Sub-projects partially capture, measure or modify the same variables. Data and insights generated in sub-projects should be structured, networked, and made available for a comprehensive analysis.

The documentation of research activities, such as experiments, should be available across all sub-projects through digital protocols. Experiments are often repeated while varying conditions such as process variables or specimens. The basic description of the steps, specimen, and variables of an experiment should be represented by generic protocol types, which can be flexibly composed of reusable semantic building blocks for the representation of specimens and variables. Protocol types serve as an easy-to-use template for creating a protocol to document a specific experiment by capturing specific values for the variables assembled in the protocol type. Creating and adapting protocol types should be easy and intuitive so that both documentation and its structure can be quickly and independently adjusted to changes in experiment design. Semantic building blocks can be reused, so that a new declaration is not necessary when creating a new protocol type. The declaration and definition of variables is done once, separated from the protocol types, with a unique naming convention and is subsequently referenced in protocol types.

4.2 Modeling of a generic protocol structure

With the aim to achieve reproducibility, reusability, and interoperability of data generation research activities are modeled reusing parts of the Metadata4Ing ontology. The Metadata4Ing ontology enables a description of data generation processes, associated artifacts, and procedures for data manipulation [15]. It implements concepts of inheritance and modularity, making it ideal for the modular approach of semantic building blocks described here. The names of entities may differ from the terms in

Metadata4Ing to maintain the terms used in the joint project or to avoid reserved terms in the software.

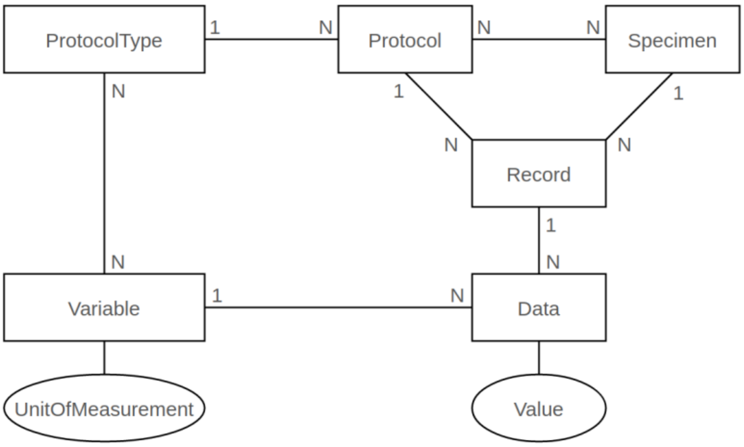


Fig. 1. Entity–relationship model of the generic protocol structure

At the center of our model (Fig. 1) is the entity *Protocol*, whose instance documents the performance of an individual experiment. A protocol type describes the basic steps and conditions. The same process variables are always used or variables are measured during the execution of these steps, but with varying values and from varying specimens. In addition, multiple specimens can be considered in a protocol.

Table 1. Mapping of the model to concepts in the Metadata4Ing ontology.

Entity	Category in SMW	Class Metadata4Ing	Description
Specimen	Specimen		specimen to be investigated
Variable	Variable	*PIMS-II:Variable*	variable of the specimen
Protocol	Protocol	*ProcessingStep*	documentation of an experiment
Protocol type	ProtocolType	*PROV:Activity*	type of experiment
Record (Dataset)	Record	*PIMS-II:Assignment*	container for metadata of a specimen within a protocol
Key-value-pair	Data („Value“ is reserved by the system)	*PIMS-II:Value*	assignment of a value to a variable in a record
Quantity	UnitOfMeasurement (“Quantity is reserved by the system)	*EMMO:MeasurementUnit*	measured quantity including the definition of admissible units

The *Variables* are separate entities to ensure consistent naming and reusability, including a unique definition with associated units. The entity *Record* serves as a container for measured values of variables in the relation between *Protocol* and *Specimen*, represented as key-value pairs. An example of a variable in the model is the oxygen partial pressure or the diameter of a specimen. During the execution of an experiment, the measured value, along with other variables and their values, is summarized and documented in a dataset within a protocol. Table 1 shows the mapping of the model to concepts in the Metadata4Ing ontology and the corresponding categories for implementation in SMW as described in the following section 4.3.

4.3 Implementation into the Knowledge Management System Semantic MediaWiki

The generic protocol structure is implemented in the knowledge management system SMW [8]. SMW is a semantic extension of the software MediaWiki, known from Wikipedia. With SMW, data and information can be semantically annotated. Semantic statements about data in the form of subject, predicate, object can be modeled and represented on wiki pages using a specific syntax through categories and so-called properties.

For each entity in the model, an equivalent SMW category with required properties was created in SMW. An instance of a category and its associated data, as well as their assignment to semantic properties, are captured as key-value pairs with a form and displayed in a structured manner through templates. In general, static forms are created in SMW, so that the input form of a category always captures the same fields. Therefore, for a static protocol structure, it is necessary to define a separate form for each protocol type. In the presented approach of the generic protocol structure, the process variables and the variables to be recorded for each specimen of a protocol should be specified by the corresponding protocol type. For this purpose, instances of the variable category are created for all required variables, linked to a unit of measurement that describes the measured quantity and determines its units.

The form for the protocol type is extended with a field for a list of variables to be examined in the experiment. In order to capture exactly these variables in records of a corresponding protocol instance for each specimen, the record-form is dynamically generated in an outsourced template. This template receives the instance of the parent protocol, through which the list of variables to be examined is available through a query. Using this list, a Wiki-syntax with escape sequences is generated, which, when included in the record-form, dynamically generates the desired fields for the user. Thus, the fields in the form are dynamically generated and do not need to be implemented statically for each protocol type. The representation of the data thus captured is done similarly in the template of the record page.

5 Use case - knowledge management in the Collaborative Research Centre “Oxygen-free Production”

5.1 Status quo of research activity documentation

In practice, research involves a wide range of experiments and detailed documentation. In the context of the oxygen-free production application example, experiments as well as the outcomes are very heterogeneous. Typical experiments for material investigation like atomic emission spectroscopy, x-ray diffraction and scanning electron microscopy lead to high resolution images including the corresponding value tables. Experiments like hardness tests and tensile tests lead to large data series. Virtual experiments such as casting simulations are often also documented additionally in form of video content. Consequently, in addition to subjective or organization-internal documentation, this leads to a further lack of structured documentation that is traceable across projects. Figure 2 shows the status quo of data collection and documentation starting with the initial documentation of experiments by handwritten notes.

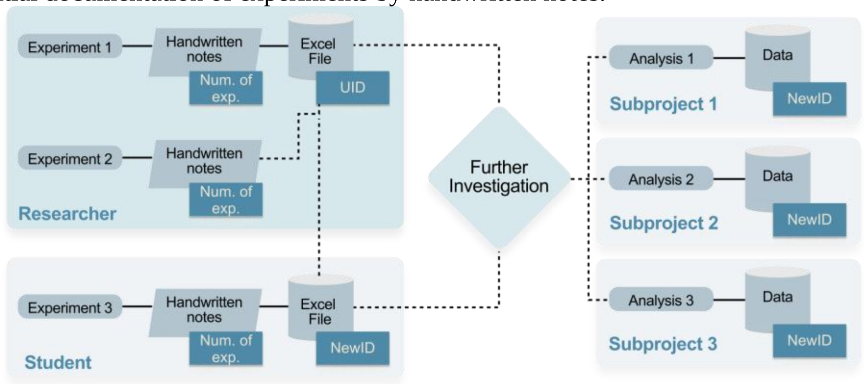


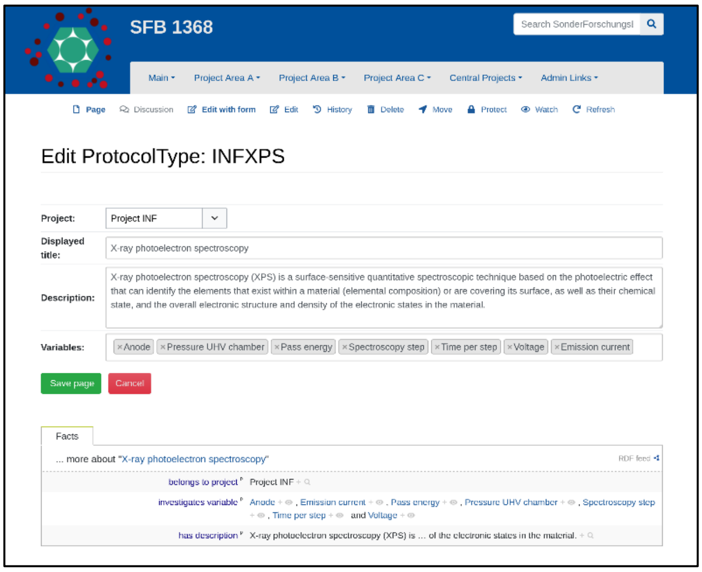
Fig. 2. Flowchart describing a common way of data documentation done by a researcher involving interaction with students and other subprojects.

For practical reasons, the specimens generated are temporarily given an ascending numerical designation, but not yet their unique identifier (UID). The handwritten protocols are transferred to a summary database, such as an Excel file. This process assigns each specimen its UID, but retains its temporary numbering from the hands-on experiment, as the specimens are often already labeled with this temporary ID. As an additional challenge, experiments may also be carried out by students using their own slightly different way of documentation, which may vary from those of the researcher. This may result in minor discrepancies in the data structure and additional efforts in merging data. In addition, specimens may be passed on to other subprojects for analysis. After successful analysis, the additional data are returned to the researcher without any information about the parameters used during the analysis, making it difficult to transfer these relevant parameters into one's own database. In the end, it is often only documented that a particular analysis was carried out.

5.2 Documentation with the generic protocol model implemented in Semantic MediaWiki

The implementation of the generic protocol structure in SMW as described in section 4.3 leads to the following workflow from a researcher's point of view. For each specific experiment, a protocol type needs to be created once (Fig. 3.).

Afterwards the researcher has to choose from the list of existing variables in the system, exactly those variables for the protocol type that are used for the experiment. In case of a missing variable, the user can create and describe it beforehand, so that a unit of measurement is linked accordingly. Each time an experiment is conducted, a new protocol must be created, selecting the protocol type and the specimen to be analyzed. The process is illustrated in Figure 4.



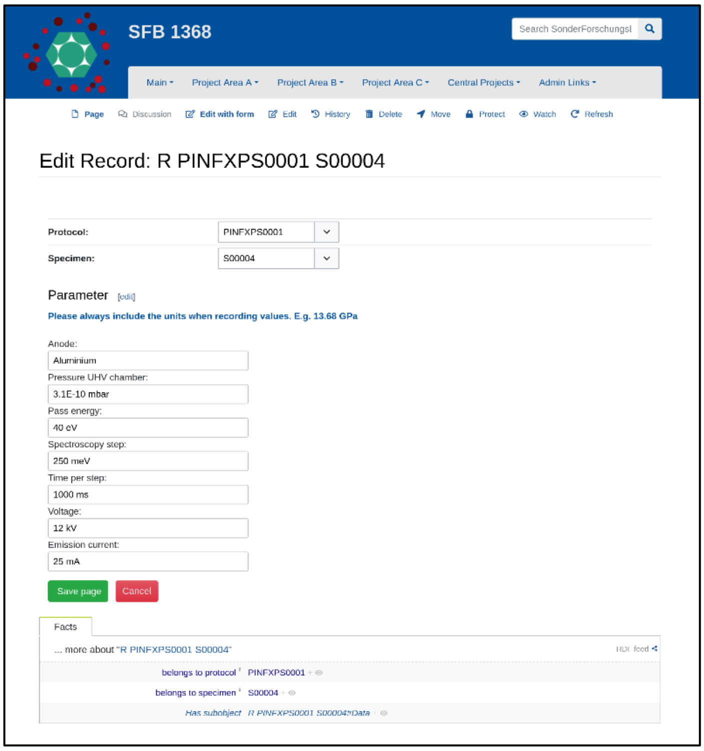
Edit ProtocolType: INFXPS

Project:	Project INF
Displayed title:	X-ray photoelectron spectroscopy
Description:	X-ray photoelectron spectroscopy (XPS) is a surface-sensitive quantitative spectroscopic technique based on the photoelectric effect that can identify the elements that exist within a material (elemental composition) or are covering its surface, as well as their chemical state, and the overall electronic structure and density of the electronic states in the material.
Variables:	×Anode ×Pressure UHV chamber ×Pass energy ×Spectroscopy step ×Time per step ×Voltage ×Emission current

... more about "X-ray photoelectron spectroscopy"

belongs to project	Project INF
investigates variable	Anode , Emission current , Pass energy , Pressure UHV chamber , Spectroscopy step , Time per step and Voltage
has description	X-ray photoelectron spectroscopy (XPS) is ... of the electronic states in the material.

Fig. 3. Creation of a research protocol template in Semantic MediaWiki using the generic protocol structure



Edit Record: R PINFXPS0001 S00004

Protocol:	PINFXPS0001
Specimen:	S00004

Parameter [edit]

Please always include the units when recording values. E.g. 13.68 GPa

Anode:
Aluminium

Pressure UHV chamber:
3.1E-10 mbar

Pass energy:
40 eV

Spectroscopy step:
250 meV

Time per step:
1000 ms

Voltage:
12 kV

Emission current:
25 mA

... more about "R PINFXPS0001 S00004"

belongs to protocol	PINFXPS0001
belongs to specimen	S00004
Has subobject	R PINFXPS0001 S00004#Data

Fig. 4. Entering values of a protocol record

A link to the corresponding data record is located on the protocol's wiki page for each specimen. This link provides access to a form that can be used to enter the specific variables for the specimen, which were selected when the protocol type was created. Thus the process of documenting research activities has not changed significantly. The user can flexibly create protocols himself and fill them out. With these generated data, the history of each specimen with the changes of variables through all subprojects can be viewed, traced and semantically queried. In addition, specimens and protocols are assigned UIDs that are generated uniformly by the system and are to be used automatically by all subprojects. This further standardization in the documentation increases the findability of the data sets and their documentation. In order to evaluate the effectiveness of the generic protocol structure and its implementation into the knowledge management system in documenting experimental procedures and parameters, an initial test with researchers from the Collaborative Research Centre was conducted. Three Participants were assigned tasks to model and document their experimental procedures, parameters, materials and instruments used with the SMW system, either by designing semantic building blocks for reuse in their protocols or by embedding already existing building blocks. The results showed that the researchers were able to independently transfer their existing documentation, which was in the form of spreadsheets, logbooks, and text documents, into the SMW system without any significant loss of information. They were also able to create their own variables and new specimens for protocols. However, the navigation between forms and templates of different categories was not clear enough, requiring too many interactions, and users indicated a desire for simpler usage, such as creating multiple specimens at once or having a better flow through multi-step forms. Additionally, users requested new features, such as personal specimen numbers and a relation of specimens to a materials database. These requirements can be implemented with reasonable effort from the current development stage and are compatible with our model and the generic protocol structure. The results highlight the importance of prioritizing user experience.

6 Conclusion and future work

In this paper, we introduced an approach for modeling research activities using semantic building blocks, enabling the flexible generation of structured, semantically annotated protocol templates to document research activities in a discipline-specific manner. The approach was implemented and evaluated using the SMW knowledge management system within the CRC 1368, which facilitated the documentation of experiments conducted in subprojects. Our approach shows that protocols can be created by researchers with reasonable efforts providing the necessary flexibility while generating structured documentation about data generation. In the next steps we will apply the generic protocol structure to more subprojects of the CRC creating a collection of semantically interlinked research data and data provenance. Future work will include analysis of relations between research activities, data and data provenance, enabling logical inferences from research data to support decision-making and ensuring the accessibility and interpretability of large amounts of complex structured information.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 394563137 – SFB 1368

References

1. Maier, H. J. et al.: Towards Dry Machining of Titanium-Based Alloys: A New Approach Using an Oxygen-Free Environment, Metals 10 (9), p. 1161. (2020). DOI:10.3390/met10091161.
2. Wilkinson, M. D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3:160018. (2016). DOI:10.1038/sdata.2016.18.
3. Devaraju, A. et al.: From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. Data Sci J., 20: 4, pp. 1–14. (2021). DOI:10.5334/dsj2021-004.
4. Amorim, R. C. et al.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. UAIS 16, pp. 851–862. (2017). DOI:10.1007/s10209-016-0475-y.
5. Grönewald, M.: Forschungsdatenmanagement mit elabFTW im SFB/TRR 270. Zenodo (2022). DOI:10.5281/zenodo.6772197.
6. Brandt, N. et al.: Kadi4Mat: A research data infrastructure for materials science. Data Sci J. 20 (2021). DOI:10.5334/dsj-2021-008.
7. Gohlke, H.: LISTER: Semi-automatic metadata extraction from annotated experiment documentation in eLabFTW. Chemotion/NFDI4Chem Stammtisch; 2022 Nov 25; Homepage https://www.nfdi4chem.de/index.php/abstract-gohlke/. last accessed 2023/02/25.
8. Krötzsch, M. et al.: Semantic MediaWiki. In: The Semantic Web - ISWC (2006). Lecture Notes in Computer Science, vol 4273 (2006). DOI:10.1007/11926078_68.
9. Mozgova, I. et al.: Research Data Management System for a large Collaborative Project. DS 101: Proceedings of NordDesign 2020, 12th - 14th August 2020, Lyngby, Denmark (2020). DOI:10.35199/NORDDESIGN2020.48.
10. Mozgova, I. et al.: Knowledge Annotation within Research Data Management System for Oxygen-Free Production Technologies. pp. 525-532. Proceedings of the Design Society (2022). DOI:10.1017/pds.2022.54.
11. Altun, O. et al.: Integration eines digitalen Maschinenparks in ein Forschungsdatenmanagementsystem. pp. 1-10. Proceedings of the 32nd Symposium Design for X (DFX2021) (2021). DOI:10.35199/dfx2021.23.
12. DataCite Metadata Working Group: DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V. (2021). DOI:10.14454/3w3z-sa82.
13. Bruno, G. et al.: Efficient management of product lifecycle information through a semantic platform. International Journal of Product Lifecycle Management, 9(1), 45 (2016). DOI:10.1504/ijplm.2016.078864.
14. Moreau, L. et al.: The rationale of PROV. Web Semantics: Science, Services and Agents on the World Wide Web. 35, pp. 235–257. (2015) DOI:10.1016/j.websem.2015.04.001.
15. Fuhrmans, M., Iglezakis, D.: Metadata4Ing - Ansatz zur Modellierung interoperabler Metadaten für die Ingenieurwissenschaften. Zenodo (2020). DOI:10.5281/zenodo.3982367.
16. Schmitt, R.H. et al.: NFDI4Ing-the National Research Data Infrastructure for Engineering Sciences. Zenodo (2020). DOI:10.5281/ZENODO.4015201.