Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-10 (2022) 2030–2035

Creation of a Knowledge Space by Semantically Linking Data Repository and Knowledge Management System - a Use Case from Production Engineering

Tatyana Sheveleva*, Max Leo Wawer, Pooya Oladazimi*, Oliver Koepler*, Florian Nürnberger***, Roland Lachmayer**, Sören Auer*, Iryna Mozgova****

**TIB - Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30167 Hanover, Germany (Tel: 511-762-18838; e-mail: tatyana.sheveleva@tib.eu).*

*** Institute of Product Development, Leibniz University Hannover, An der Universität 1, 30823 Garbsen, Germany (e-mail: mozgova@ipeg.uni-hannover.de)*

****Institut für Werkstoffkunde (Materials Science), Leibniz University Hannover, An der Universität 2, 30823 Garbsen, Germany (e-mail: nuernberger@iw.uni-hannover.de)*

Abstract: The seamless documentation of research data flows from generation, processing, analysis, publication, and reuse is of utmost importance when dealing with large amounts of data. Semantic linking of process documentation and gathered data creates a knowledge space enabling the discovery of relations between steps of process chains. This paper shows the design of two systems for data deposit and for process documentation using semantic annotations and linking on a use case of a process chain step of the Tailored Forming Technology.

Keywords: Knowledge Management in Production, Semantic Annotation, Research Data Management.

1. INTRODUCTION

The need to organize processes to improve the collection, processing, storage, and reuse of scientific data as a valuable asset, although still not part of the scientific culture, is becoming more evident in the scientific community. The existence of initiatives such as the German National Research Data Initiative (NFDI) indicates the importance of data management relevance and governance in research projects (Hartl et. al., 2021).

Heterogeneity of data types and formats, incomplete or missing data description, duplication, gone or inconsistent data (Effertz, 2010), the lack of standardized processes and common platforms for working with research data are some of the most common problems of research data utilization. These problems are specific to individual research projects and become especially noticeable at the level of collaborative projects where different organizational units are involved.

In Collaborative Research Centers (CRCs), research projects from various sub-disciplines work together on a mutual research problem (Mozgova et al., 2020). Each project generates, processes and analyzes data using a variety of procedures and methods (Sandfeld et al., 2018). The consolidation and harmonization of this data and the documentation of the data generation for an overall view and analysis is a critical success factor of collaborative research. The FAIR data principles (findability, accessibility, interoperability, and reusability) (Wilkinson et al., 2016) describe how data can be improved to meet such challenges. Research Data Management Systems (RDMSs) support research projects in processing, managing and depositing data collections (Mozgova et al., 2022), applying semantic annotations and enabling comprehensive descriptions of data, thus improving the overall quality of data management.

This paper presents the design and implementation of semantic annotations for the documentation of

10.1016/j.ifacol.2022.10.006

experiments of a production process step in a RDMS. It shows an approach to provide contextual information about datasets by linking the data generation process and the data itself. A use case from the CRC 1153 “Process Chain for Manufacturing Hybrid High Performance Components by Tailored Forming” (Siqueira et al., 2019) is provided. Taking the example of documenting a joining process within the RDMS, the approach for applying an appropriate semantic annotation in the RDMS is presented. In section 2 the RDMS and its components are described. Section 3 is dedicated to the use case in detail, starting with the system requirements from the users’ and developers’ perspectives, the identification and creation of the entities for semantic annotation, and its implementation into the system to enable the researchers to document their research processes. Section 4, summarizes the created knowledge space by linking research of the semantic annotation.

2. RESEARCH DATA MANAGEMENT SYSTEM

The designated RDMS addresses two major aspects of research data management (Fig. 1). These are, on the one hand, the research data itself and, on the other hand, all data and information necessary to comprehensively describe and document the context of research data generation. In particular, the descriptive data (metadata) are of outstanding importance when it comes to the reproducibility of research data across research projects reusing data from various sources. The research data in the CRC have a wide range in terms of data formats used, size and descriptive metadata. Data may originate from experiments or simulations. In most cases the generation of research data involves the use of machines and tools. The descriptive data for these machines and tools, in particular setting parameters, are important metadata to document the research data generation. The contextualization of research data generation is captured in protocols from experiments, tests, measurements, or simulations. These documents contain detailed information on samples and methods used and the by whom, when, with what and why. Protocols therefore provide all contextual data and information needed to fully understand the idea for an experiment, its design, realization, data captured, and its interpretation. The designed RDMS consists of a data repository (DMS) to deposit research data and a Knowledge Management System (KMS) to manage contextual data or information in the form of protocols. On top of this, a detailed cross-linking of all entities like data, protocols, samples, machines, and tools used between the DMS and the KMS is implemented. However, entity linking is not limited to a single experiment. In the CRC, interconnected steps within process chains that interact with each other are studied. Identifying hidden dependencies and correlations between experimental conditions of these process steps offers huge potential to increase the efficiency of the manufacturing process. Semantic linking of protocols, data, and all entities involved across DMS and KMS enables the creation of a knowledge space for a given research question in the end.

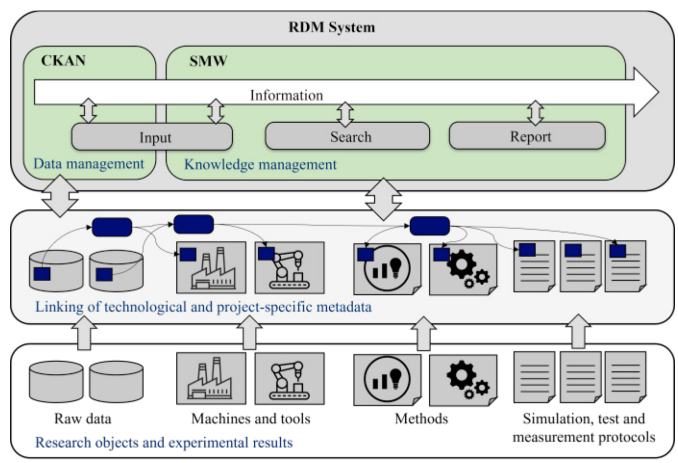

Figure 1. The Concept of RDMS.

The RDMS of the CRC 1153 takes into account the specific needs for the manufacturing processes within the Tailored Forming Technology. The typical manufacturing process consists of three or more process steps: 1) joining of two or more mono-materials to one hybrid workpiece, 2) its further preforming to a required shape, and 3) the subsequent post-processing. In such process chains, communication and exchange of data between all participants is essential to achieve most efficient manufacturing results. For this reason, the description of applied processes and their setting parameters, the machines and tools, the properties of the materials used, as well as the properties of the samples should be available to all those involved in the process chain in order to increase the reproducibility of data.

With the implementation of the RDMS, both the deposit of research data using upload forms of the DMS and the documentation of data generation using protocols is more structured and standardized. For this reason, a machine-interpretable controlled vocabulary expressing the concepts of the Tailored Forming Technology is created. This will later on enable the semantic linking of the different entities.

The KMS captures the description of entities related to research activities like protocols, samples, processes, machines, and tools. The KMS is built on Semantic MediaWiki (SMW) (Semantic-mediawiki.org contributors, 2022) and enables contextualizing, storing as well as maintaining unstructured data and information in a structured way using machine-interpretable semantics (Krötsch, et al., 2006). Samples, protocols, machines, and tools are entered into the KMS by applying normalized templates which use the controlled vocabulary in the background. The DMS builds on the open data portal software CKAN (CKAN Association, 2022). CKAN can store various data formats, provides data preview features and comprehensive search functionalities. It further enables semantic annotations of datasets using controlled vocabularies.

Both KMS and DMS are linked based on the shared semantics used for annotation of protocols in SMW and datasets in CKAN. This promotes mutual access to all related entities, as well as a common understanding of the data across all sub-projects. Using controlled vocabularies within KMS and DMS is a well-known approach in different research disciplines (Wellner, 2013). An example use case is the quality assess of the production chain results already in the early steps (Brockmöller et al., 2017).

3. USE CASE - DOCUMENTATION OF EXPERIMENTAL PROTOCOLS

3.1 Identification of researchers' requirements

Within the sub-project *Extrusion*, the researchers are responsible for the production of composite profiles by applying a novel joining process Lateral Angular Co-Extrusion (LACE) (Thürer et al., 2020). Since LACE is an experimental manufacturing process that aims to identify the best condition for forming a stable bonding zone for a set of given materials to be joined, it is performed numerous times under variation of materials' batches and process configurations. To keep a full overview of the experiments performed, the researchers face major challenges: Firstly, administrative information as well as information about applied machines, tools, and their settings are to be documented. Secondly, applied joining materials and their characteristics are to be considered, because alloys from different batches are used. They may feature different material characteristics like melting point, strength, or plasticity. Such differences in characteristics might influence the experiment performance and result both positively and negatively. Thirdly, the documented information shall be accessible and traceable for all CRC members involved into the relevant process chain, as LACE forms the first step in to enable a more time- and resource-efficient process design. This also concerns data generated.

To meet the researchers' needs for the protocol documentation, the following requirements are identified in order to create an appropriate semantic annotation:

R1: The protocol structure shall be universal to enable reuse for the documentation of other experiment types within the CRC;

R2: Entities that represent samples, projects, protocols, and devices shall be semantically interlinked;

R3: The protocol shall visualize the general characteristics of the samples assigned to this protocol;

R4: When filling out the protocol, only machines and tools including their properties relevant to the given process shall be available for selection by the user;

R5: It shall be traceable which data resource is gained in the given experiment;

R6: Data resource shall be semantically interlinked with the machine and tool that produces this data resource;

R7: The DMS and KMS shall be semantically interlinked.

3.2 Creation of semantic entities

When developing semantic annotation, first key research objects appropriate for LACE are identified and relevant categories are created: *Sample*, *Protocol*, *Project*, and *Device*. Hereby, the *Device* category is divided into *Machine*, which can be used independently to realize a specific function, and *Tool*, which cannot act or be used independently and needs to be combined with a machine. This allows the implementation of templates to document the same experiment repeated several times using one machine in combination with different tools. To realize individual wiki pages represented by individuals of the given categories unambiguously, several kinds of unique identifiers are created: The *Sample Unique Identifiers* are built by using the letter *A* representing the project *A1* and a four-digit number, (*A0001*). The *Unique Protocol Identifier*s consist of the *P* that stands for a protocol, *A1* - for the project, and *LAE* - for the process acronym. (In order to simplify the processing of the identifier by SMW, the internally used acronym LACE has been shortened to *LAE*). The four-digit number represents an individual protocol, (*PA1LAE0000*). *Machine* is designated using three elements: machine name, type designation, and manufacturer name, (*Extrusion press 10 MN SMS Meer*). *Project* is represented via internal designation, (*Project A1*).

Next, properties specifying identified categories are created. The category *Sample* is characterized by the set of properties *HasUniqueSampleIdentifier, HasBatchNumber, RepresentsDemonstrato*r, and *HasMaterialCombination*. The category *Machine* is characterized using two property sets: The first property set is only valid for the LACE-specific experiment setting: *HasContainerTemperature, AppliedRamSpeed*, *AppliedToolTemperature*, and *AppliedExtrusionRatio*. Here, the standardized designations (DIN EN 12258-1, 2012) are used in order to avoid possible misunderstandings between researchers. The second property set involves generalized properties applying to all machine types specified within KMS: *HasPhysicalLocation, HasManufacturer*, *HasType*, *HasImage*, *UsedBySubproject,* and *UsedByProtocolType*. The values of the two last properties are specified as pages, which allow redirection to the relevant sub-project and protocol pages. *Protocol* is specified using the properties *HasDate*, *AppliedByPerson*, *CreatedSample,* and *AppliedBySubproject*. The last two properties have values as pages, which enable redirection to the sample and sub-project pages.

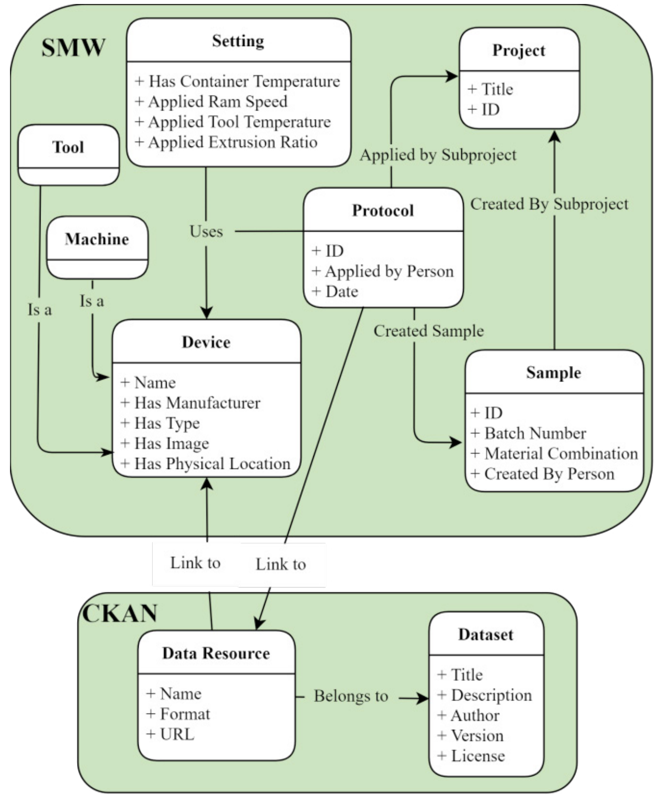

Figure 2. Entities and their relations describing a LACE protocol

After the category descriptions have been completed, the appropriate templates and forms are created for them. The LACE-relevant entities, their relations, and their interlinking with datasets in the DMS of the RDMS are illustrated in (Fig. 2). These are further used to build a complex protocol template for the LACE protocol in SMW.

3.3 Representation of a protocol in Semantic MediaWiki

Considering the requirements from 3.1, the needed protocol structure is realized as an SMW page with several embedded templates of the above categories (see Fig. 3). These categories are visualized as separate sections. The first section, General Protocol Information, is based on the template combining a set of protocol-related administrative properties mentioned in 3.2. The second section, Sample Info, is based on a sample-related template, that is built using a query calling the information stored on an individual sample page, (unique sample identifier, material combination, batch number, and the relevant demonstrator type). The third section, Machines & Tools - Settings, is created through the integration of a complex machine-related template. It incorporates two separate auxiliary templates. The first auxiliary template involves settings associated with the machine that are used within the LACE. The second auxiliary template uses a query and enables visualization of settings on the protocol page, once the relevant machine is selected. The fourth section, Linked Data Resources in CKAN, is realized by the integration of a template enabling to add the URL of the relevant data resource stored in CKAN.

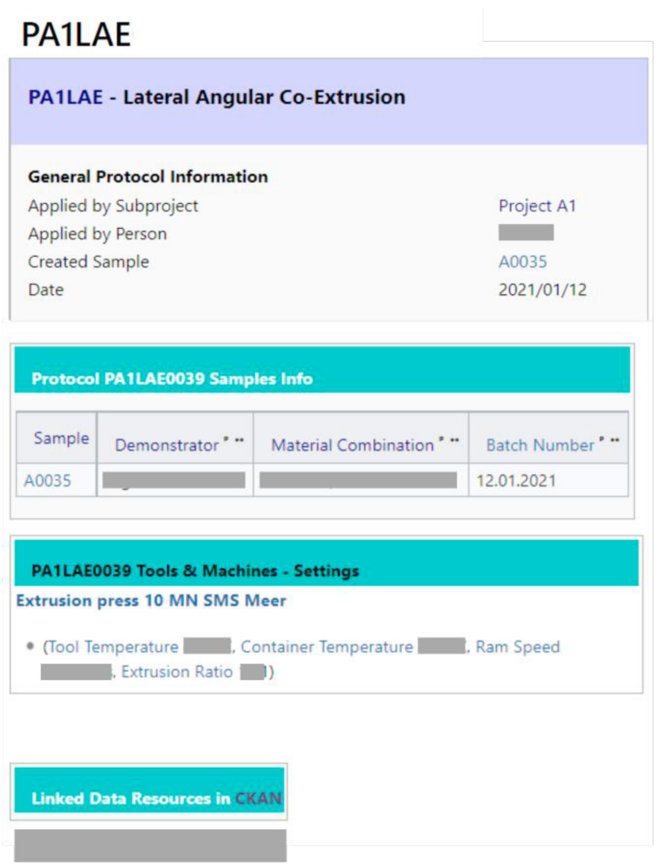

Figure 3. Visualization of a LACE protocol in SMW

Each data resource in a dataset in CKAN can be linked to a device in SMW by utilizing the plug-in Ckanext-Semantic-Media-Wiki (TIB, 2021). After storing the dataset, the device URL is shown on the target data resource page in CKAN. Each protocol in SMW is linked to a data resource in CKAN. The data resource URL is shown on the target protocol page in SMW. Since protocols and devices are connected to each other in SMW, the link between CKAN data resources, protocols, and devices can be queried and realized.

4. CONCLUSIONS

This paper presents an RDMS consisting of a data repository and knowledge management system for handling both generated research data and the contextual information on data generation in the form of samples, protocols, machines and tools. The semantic annotation of these research artifacts enables their connection across the system, thus providing a comprehensive understanding of both data and process design. It further enables researchers to seamlessly reproduce research results even along a sequence of experiments along the process chain. This approach increases the FAIRness of the research data in the RDMS. The specific use case of a given process step in the Tailored Forming Technology can be generalized and mapped to other research domains with respect to the representation of samples, protocols, tools, and machines. So far, it has been successfully implemented in another collaborative project (Altun et al., 2021) for the integration of a digital machine park into a RDMS. Although parts of the semantic annotation for the LACE protocol are very specific, one can see that the structure remains very flexible. It can be modified by adding or removing entities to make the protocol suitable for other applications. In the future, the addition, semantic annotation, and linking of further research artifacts in the RDMS is planned. Furthermore, the granularity in describing research objects like *Sample* will be increased by adding their input/output characteristics for a given experiment.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support from the CRC 1153 Process Chain for Manufacturing Hybrid High Performance Components by Tailored Forming, Project number 252662854 (INF), funded by the German Research Foundation (DFG).

REFERENCES

Altun, O., Sheveleva, T., Castro, A., Oladazimi, P., Koepler, O., Mozgova, I., Lachmayer, R., Auer, S. (2021). Integration eines digitalen Maschinenparks in ein Forschungs-datenmanagementsystem. *Proceedings of the 32nd Symposium Design for X* (DFX2021), 32 (23). DOI: 10.35199/dfx2021.23.

Brockmöller, T., Mozgova, I., Lachmayer, R. (2017). An Approach to Analyse the Potential of Tailored Forming by TRIZ Reverse. *Proceedings of the 21st International Conference on Engineering Design* (ICED 17), 4: Design Methods and Tools, pp. 445-452, Vancouver, Canada, 21-25 August. ISBN: 978-1-904670-92-6.

CKAN Association. CKAN - The Open Source Data Portal Software. https://ckan.org/. Accessed 08.02.2022.

DIN EN 12258-1. (2012). Aluminium und Aluminiumlegierungen - Begriffe - Teil 1: Allgemeine Begriffe; Dreisprachige Fassung EN 12258-1:2012. DOI: 10.31030/1888726.

Effertz, E. (2010). The Funder's Perspective: Data Management in Coordinated Programmes of the German Research Foundation (DFG). In Bareth G., Curdt C. (ed.), *Proceedings of the Data Management Workshop.* Geographisches Institut der Universität zu Köln, Cologne. DOI: 10.5880/TR32DB.KGA90.7.

Hartl, N., Wössner, E., and Sure-Vetter, Y. (2021). Nationale Forschungsdateninfrastruktur (NFDI). *Informatik Spektrum,* 44, pp. 370–37. DOI: 10.1007/s00287-021-01392-6.

Krötsch, M., Vrandečić, D., and Völkel, M. (2006). Semantic MediaWiki. *The Semantic Web* 4273, pp. 935-942. DOI: 10.1007/11926078_68.

Mozgova, I., Koepler, O., Kraft, A., Lachmayer, R., Auer, S. (2020). Research data management system for a large collaborative project. *Proceedings of NordDesign 2020*, Lyngby, Denmark. DOI: 10.35199/NORDDESIGN2020.48.

Mozgova, I., Jagusch, G., Freund, J., Kraft, A., Glück, T., Herrmann, K., Knöchelmann, M., Lachmayer, R. (2022). Product Life Cycle Oriented Data Management Planning with RDMO at the Example of Research Field Data. Heuveline, V., Bisheh, N. (Hrsg.): *E-Science-Tage 2021: Share Your Research Data*, Heidelberg: heiBOOKS, DOI: https://doi.org/10.11588/heibooks.979.

Sandfeld, S., Dahmen, T., Fischer, F. O.R., Eberl, C., Klein, S., Selzer, M., Nestler, B., Möller, J., Mücklich, F., Engstler, M., Diebels, S., Tschuncky, R., Prakash, A., Steinberger, D., Kübel, C., Herrmann, H.-G., Schubotz, R. (2018). Strategiepapier Digitale Transformation in der Materialwissenschaft und Werkstofftechnik. *Deutsche Gessekschaft für Materialkunde.* URL: https://edocs.tib.eu/files/e01fn18/1028913559.pdf. Accessed 22.01.2022.

Semantic-mediawiki.org contributors: Semantic MediaWiki, https://www.semantic-mediawiki.org/w/index.php?title=Semantic_MediaWiki&oldid=76616, Accessed 30.04.2022.

Siqueira, R., Bibani, M., Duran, D., Mozgova, I., Lachmayer, R., Behrens, B.-A. (2019). An Adapted Case-based Reasoning for Design and Manufacturing of Tailored Forming Multimaterial Components. *International Journal for Interactive Design and Manufacturing*. 13 (3) DOI: 10.1007/s12008-019-00566-7.

Technische Informationsbibliothek (TIB). (2021). Ckanext- Semantic Media Wiki. URL: https://github.com/TIBHannover/ckanext-Semantic-Media-Wiki. Accessed 24.01.2022.

Thürer, S., Peddinghaus, J., Heimes, N., Bayram, F., C., Bal, B., Uhe, J., Behrens, B.-A., Maier, H. J., Klose, Ch. (2020) Lateral Angular Co-Extrusion: Geometrical and Mechanical Properties of Compound Profiles. *Metals,* 10(9), 1162. DOI: 10.3390/met10091162.

Wellner K. (2013). B 6 Ontologien. In Laisiepen K. (ed.), Grundlagen der praktischen Information und Dokumentation: Handbuch in die Informationswissenschaft und -praxis, De Gruyter Saur, Berlin, pp. 207-218. DOI: 10.1515/9783110961355.160.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3(160018). DOI: 10.1038/sdata.2016.18.