INTERNATIONAL DESIGN CONFERENCE – DESIGN 2022
https://doi.org/10.1017/pds.2022.54

Knowledge Annotation within Research Data Management System for Oxygen-Free Production Technologies

I. Mozgova [1,✉], O. Altun [1], T. Sheveleva [2], A. Castro [2], P. Oladazimi [2], O. Koepler [2], R. Lachmayer [1] and S. Auer [2]

[1] Leibniz Universität Hannover, Germany,
[2] Leibniz Information Centre of Science and Technology University Library, Germany

✉ mozgova@ipeg.uni-hannover.de

Abstract

The comprehensive implementation of digital technologies in product manufacturing leads to changes in engineering processes and requires new approaches to data management. An important role belongs to the processes of organizing the collection, storage and reuse of research data obtained and used in the process of product, system or technology development, taking into account the FAIR data principles. This article describes a Research Data Management System for the organization of documentation and measurement requests in the research and development of new oxygen-free production technologies.

Keywords: research data management, knowledge management, semantic modelling, FAIR data principles, design support system

1. Introduction

In large collaborative research projects, the establishment and maintenance of an efficient research data management that supports the research processes and facilitates data flow between sub-projects is crucial. Modern information technologies enable efficient collection and accumulation of a large amount of heterogeneous data in engineering. Especially in large interdisciplinary projects with several sub-projects, the use of Research Data Management (RDM) is essential. This requires a comprehensive analysis of the process and software involved in data generation as well as the information flows and data exchanges between sub-projects.

Due to the increasing digitization in the manufacturing industry, it can be assumed that successful concepts from the IT sector are also likely to become established in daily operations in industry. One example is inner-sourcing, in which open-source concepts, network-based collaboration processes and tools are used in intra-organizational projects. By establishing such procedures also in research projects, the exchange of ideas and knowledge within collaborative projects and thus its innovation potential can be strengthened. Data repositories serve as RDM systems for archiving and accessing data. They are comprehensively described by metadata. Knowledge Management Systems (KMS) complement the documentation of data provenance. The linkage and combined use of such systems generates added value for researches, especially in identifying contexts.

In (Amorim et al., 2017) well-known RDM systems are described that can be used by organizations to support the RDM workflow. They are important tools in large collaborative projects to implement the FAIR (Findable, Accessible, Interoperable, Reusable) data principles according to (Wilkinson et al., 2016). These principles describe the requirements for annotating data with detailed metadata and therefore to provide information about the context of a data set and ensuring its correct interpretation.

In addition to the FAIR data principles, other objectives must be considered when developing RDM systems. A general description of the goals of RDM systems is summarized in (Mozgova et al., 2020). The digital representation of materials, processes, resources and research results is an important aspect of joint organization of work in large collaborative projects and helps to generate information and knowledge from available data. To do so several issues have to be considered and the following questions appear:

- How to implement intuitive RDM tools to support research data processing between different projects or in collaborative projects?
- How can be organized the process of documenting and monitoring measurements of research objects between sub-projects?
- How can data in RDM systems be interlinked with each other in order to generate new findings?

This paper describes the concept of a semantically linked system consisting of a data repository and a KMS. An overview of current data repositories, approaches for handling metadata, and the creation of domain-specific controlled vocabularies and ontologies is given. The general concept of the RDM system is described and the implementation on an example of a Measurement Request (MR) system into the RDM system of the Collaborative Research Center (CRC) 1368 Oxygen-free production is shown.

2. Systems to support the research data management

In many cases, there are no common standards for describing data and documenting data generation. In this context, documentation and timely accessibility of data across several interrelated sub-projects enables all participants to process and analyse data more efficiently. This chapter provides an overview of various approaches to organizing RDM systems.

2.1. Research data management systems within large research projects

There are some CRCs that are running an RDM system in different disciplines. One of the examples from mechanical engineering and production technology is the CRC1194 focusing on the analysis of the interaction between transport and wetting processes when heat and mass transport processes occur in parallel and when complex fluids and surfaces are applied. Research data gained within research activities is stored in an internal institutional repository based on open-source software DSpace (Smith et al., 2003) enabling metadata-based data description, assignment of DOIs, access rights management, versioning, linking to the relevant publications, persons, and third-party projects.

Another example is the CRC 985 Functional Microgels and Microgel Systems focuses on microgels research by using a comprehensive approach that comprises individual particles as well as technical-scale production and formulation processes (Claus et al., 2019). The RDM system comprises a web-based platform for the internal collaboration and communication within the whole CRC as well as sample management system based on SharePoint. The sample management system captures information describing sample characteristics, e.g. their global unique persistent identifiers, interlinks the information about initial samples, which have a precursor relation with each other, and refers the samples to the appropriate publications. Additionally, the system stores administrative sample information about e.g. responsible persons and sub-projects. An experiment related system workspace includes information about experiments procedures and results.

The CRC/TRR 270 aims at the development of new magnetic materials for efficient energy technologies (Grönewald et al., 2020). Its RDM supports the development processes and provision of tools (e. g. electronic lab books and repositories) for organization and storage of research data and information. It facilitates the further data usage in the context of machine learning, which shall be applied on the processing of material, characterization of its structural and magnetic properties in order to predict the process flows and behavior of material not realized yet.

2.2. Collaborative Knowledge Management Systems

KMSs are widely used in different knowledge areas and aim different goals. Semantic MediaWiki (SMW) (Krötzsch et al., 2006), an open-source extension of the MediaWiki software running Wikipedia is a prominent example used for KMS. It is widely used by institutions and applied to different use cases.

https://doi.org/10.1017/pds.2022.54 Published online by Cambridge University Press

In the field of research, SMW is used as a collaborative KMS in different research domains. An example from the field of geography, geology and archelogy is the CRC 806, which has developed the knowledge base Afriki on SMW (Willmes et al., 2018). Within the biology and medicine domain SMW has been used to develop databases like SNPedia (Cariaso et al., 2012), Gene Wiki (Huss et al., 2008) and MetaBase (Bolser et al., 2012).
Furthermore, SMW is used for server and infrastructure documentation and for the management of internal administrative workflows by the Karlsruher Research Center for Information and Technologies has established a knowledge management system supporting internal activities of employment process and training process for new employees (Herzig et al., 2010).
Wikibase is open-source knowledge base software developed by Wikimedia for driving Wikidata (Vrandečić and Krötzsch, 2014). It consists of a set of MediaWiki extensions allowing Wikibase the usage a data model. The knowledge base that collects basic elements called items, that are described using statements. In contrast to SMW it consists of the Wikibase repository for data storage and management, and of Wikibase client realizing retrieve of stored data. Compared to SMW, Wikibase is not able to display the data within the wiki itself. Data can be visualized either via querying service, external tools or via the MediaWiki extensions Graph or LinkedWiki. Cargo is another extension of MediaWiki. It can be seen as an alternative to Semantic MediaWiki, but provides more lightweight way for storage and querying data. Cargo provides both exploration and search functions (MediaWiki contributors, accessed 2021).

2.3. Data Repository Platforms as Data Management Systems

Among data repository platforms, Comprehensive Knowledge Archive Network (CKAN, ckan.org), DSpace (Smith et al., 2003), EPrints (Tansley and Harnad, 2020) and Zenodo (zenodo.org) are the most used ones by research and governmental institutions (Amorim et al., 2017). CKAN is a data publishing platform that is widely used by governmental institutions. CKAN does not store the metadata in the Resource Description Framework (RDF) format. However, there is an official extension ckanext-dcat (Ckanext-dcat), that maps the metadata into the RDF format based on the DCAT standard. DSpace is a publication and data publishing platform that stores the metadata in the RDF format. Zenodo is an open-access repository for the publication of various data types. The platform allows metadata harvesting based on OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol. EPrints is an open source software that provides open access for publications and research data. Like Zenodo, EPrints also allows data harvesting based on OAI-PMH protocol (Lagoze and Sompel, 2002).
The overview of knowledge management systems and data repository plattforms allowed to determine the concept of a RDMS, for which among the main requirements should be noted the freedom of choice in the creation of semantic annotation, user-friendly visualization of the context, easy extensibility of the basic functionalities of the basic tools, realization of collaborative work and the need for basic tools that allow their independent maintenance. The chosen system architecture is described in the next section.

3. The Concept of a Joint Research Data Management System

In production engineering oxygen is often a significant disturbance factor in many processes. In the CRC 1368 around 40 researches from different disciplines are working together in several sub-projects on the fundamentals of oxygen-free production technologies in order to have new, energy efficient and resource saving production processes (Maier et al., 2020). The research data within the CRC is heterogeneous and ranges from measurement data, simulation data to images and videos, etc. Findability, accessibility, interoperability and reusability of data play a crucial role in the CRC s where data is exchanged between projects or along a process chain.
Methods and tools for the systematic storage, semantic representation, processing and analysis of information in scientific processes of production technologies need to be applied to ensure the quality of research data management with a reasonable amount of effort in terms of time and resources. In order to develop a sustainable RDM system, all research activities, data structures, equipment as well as user requirements for components of the RDM system were determined through surveys with the researchers of the sub projects (Altun et al., 2021). With regards to the data repository, requirements included granular access rights managements for datasets, the needs to collaboratively work on datasets,

https://doi.org/10.1017/pds.2022.54 Published online by Cambridge University Press

versioning of datasets, and the application of established metadata standards. An open-source solution seemed to be the best fit to implement the additional requirements of the CRC into an existing solution.

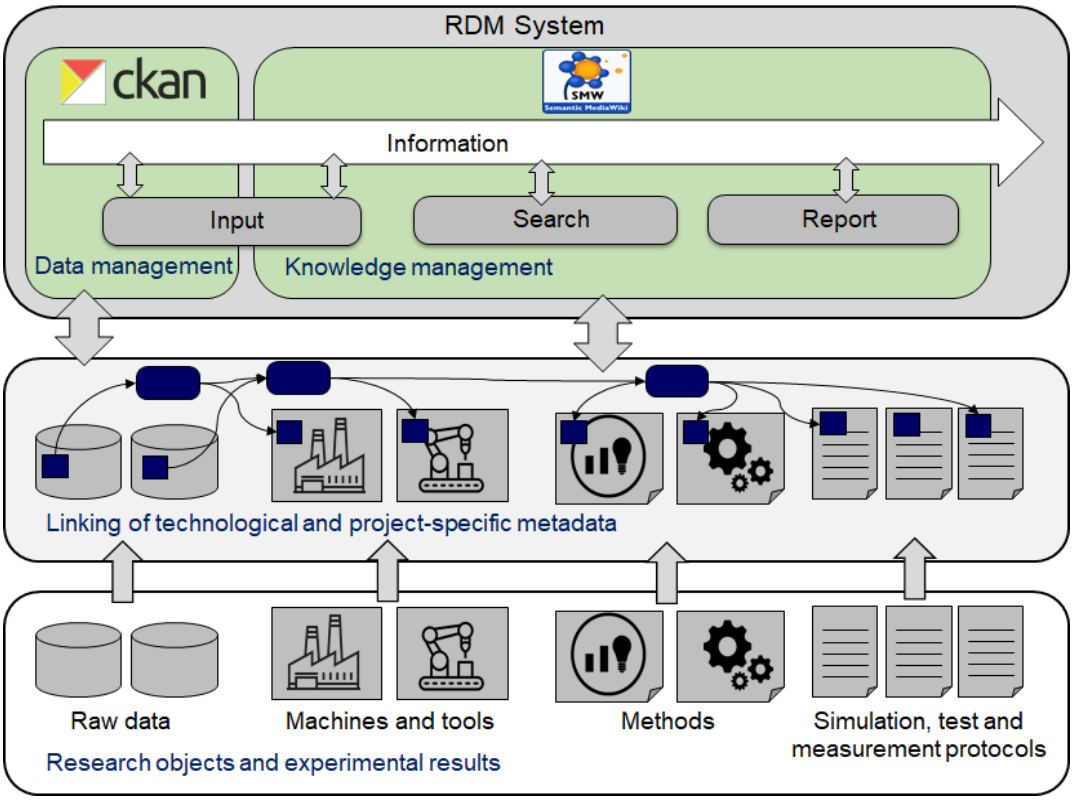

Figure 1. The Concept of research data management system

The proposed RDMS within this article provides a central data management system consists of the two sub-systems: a Knowledge Management System (KMS) based on Semantic MediaWiki (SMW) and the Data Management System (DMS) based on the CKAN software. KMS the CKAN software. Both sub-systems are connected with each other based on semantic annotation of research activities documented in form of protocols and a set of metadata applied for description of datasets gained within research activities. As a result, a common understanding of a domain-specific content within the whole CRC is provided and a mutual access to the DMS and KMS is promoted.

4. Use Case - Measurement Request System

The purpose of this example is to demonstrate how the joint RDM System support the research and development processes. Within the CRC 1368 in order to automate the collection and monitoring of the location and status of the material samples transported, prepared or transferred for the surface analytics, the project team analysed requirements for the collection and storage of the necessary data as well as for appropriate reporting and visualization of information flows.

4.1. Identification of the researchers requirements

The central analytic sub-project (Szafarska, 2021) is responsible for performing various analysis measurements for all the other sub-projects in the CRC. Researchers have to file specimen measurement request, measurement times with different devices have to be coordinated, and results and measurement data returned to the sub-projects.

The former process of the Measurements Request (MR) was based on PDF-forms that were emailed by the researcher requesting the measure of a specimen to the researcher responsible for making the measurement. The measurement and the results were documented with an Electronic Lab Notebook (ELN) used by the sub-project and emailed back to the requesting researcher. Although this system worked, as a

https://doi.org/10.1017/pds.2022.54 Published online by Cambridge University Press

way to request a measurement and receive back the results for specimens, it presented some significant shortcomings. The lack of a structured way to store the requests and their results, which made it impossible query or organize the requests according to their parameters. The data resulting from the MR, specimen information, measurement settings, measurement results and corresponding data files, were also dispersed over disconnected information spaces, such as emails, PDF forms, and ELN pages. These limitations of the previous MR system, but also its relevance to the researchers, arguing for its inclusion in the new CRC1368 RDM environment. For this use case, the structures, procedures, and interfaces need to be developed to allow researchers to create, track, search and associate MRs to other related data structures and processes within the SMW and CKAN, which form the core of CRC1368 RDM environment.

4.2. Specification of Semantics within the Research Data Management Systems

In order to realized the documentation of a MR within the RDMS, the an appropriate entities for semantic annotation is to be created for it. Thereby, the key semantic elements in form of categories and their properties are significant. Thus, the relevant key knowledge objects relevant for a measurement request are identified and the appropriate categories are created: *Institute*, *Projects*, *Specimen*, *Equipment*, and *Measurement request*.

In the next step, created categories are characterized by specific sets of properties: The category *Institute* does not uses any specified properties and is instead has been instantiated by several individuals representing research institutes participating in the CRC1368. This individuals are used to create initial wiki pages representing research institutes and therefore designated according to the official name of these institutes, like *Clausthal Centre for Material Technology*.

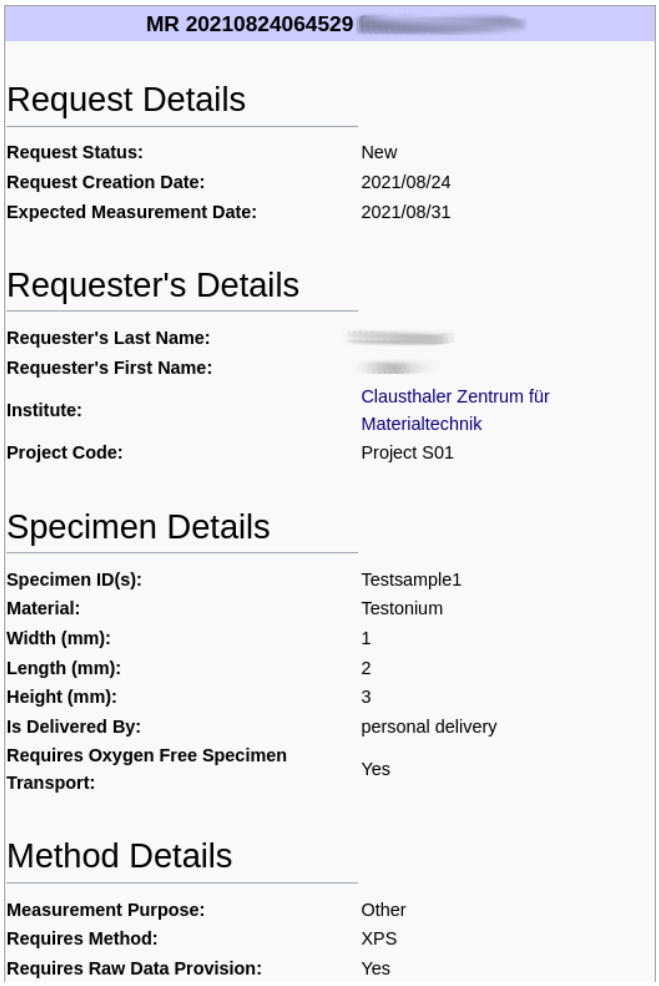

Figure 2. Measurement request instance‘s details

https://doi.org/10.1017/pds.2022.54 Published online by Cambridge University Press

The category *Projects* is characterized by the administrative property *hasProjectTitle*, *hasProjectCode*, *hasProjectLeader*, *HasProjectManager*, and *belogsToProjectArea*. The values of the last properties use a list of the predefined values representing sub-projects within the CRC1368, like *A* or *B*.
The category *Specimen* is specified by the two type of properties. The first set is geometry-relevant as *material*, *width*, *length*, and *heigh*. The second property set involves administrative properties as *hasSpecimenID*, *hasImage*, *requiresOxygenFreeSpecimenTransport*, and *isDeliveredBy*. In order to simplify the filling in of the request for the researchers as much as possible, the property *requiresOxygenFreeSpecimenTransport* has been created with the predefined value of type Boolean. The property *isDeliveredBy* is featured with predefined datatype values *postal delivery* and *personal delivery* implemented as option for selection in a drop-down list.
The category *Equipment* is characterized by using *hasModel*, *hasManufacturer*, *depiction*, *atLocation*, *employedIn*, *and allowsForMethod*. This properties are part of the Machine and Tool Ontology (MATO) and used within the implementation of a digital machine park in RDMS of the CRC1368, as described in (Altun et al. 2021). For the property *allowsForMethod*, a list of the predefined allowed values relevant for the MR is created: *AES*, *EBSD*, *ECX*, *Phase analysis*, *REM*, *STEM*, *XPS*, *residual austenite determination*, and *residual stress measurement*.
The category *Measurement request* is specified via the two sets of properties. The first set is administrative-related, where *creationDate*, and *deadlineDate* are properties of datatype Date, and *status* with the predefined values *new*, *in progress* and *done* appearing as portion for selection in a drop-down list. The property *hasInstitute* is implemented of the datatype Page, whose predefined values are represented using the named individual of the category *Institute*. Adding of the institute page as property value enables semantic interlinking so that by clicking of such value, one is redirected to the appropriate institute wiki page. The second property focuses on the method to be applied: *requiresMethod*, that uses the list of the predefined allowed values of methods mentioned above in association with the machine-related property *allowsForMethod*. The property *purpose* uses predefined values *Publication*, *Thesis*, *Proposal*, and *Other*. The property *requiresRawDataProvision* uses Boolean as a value.

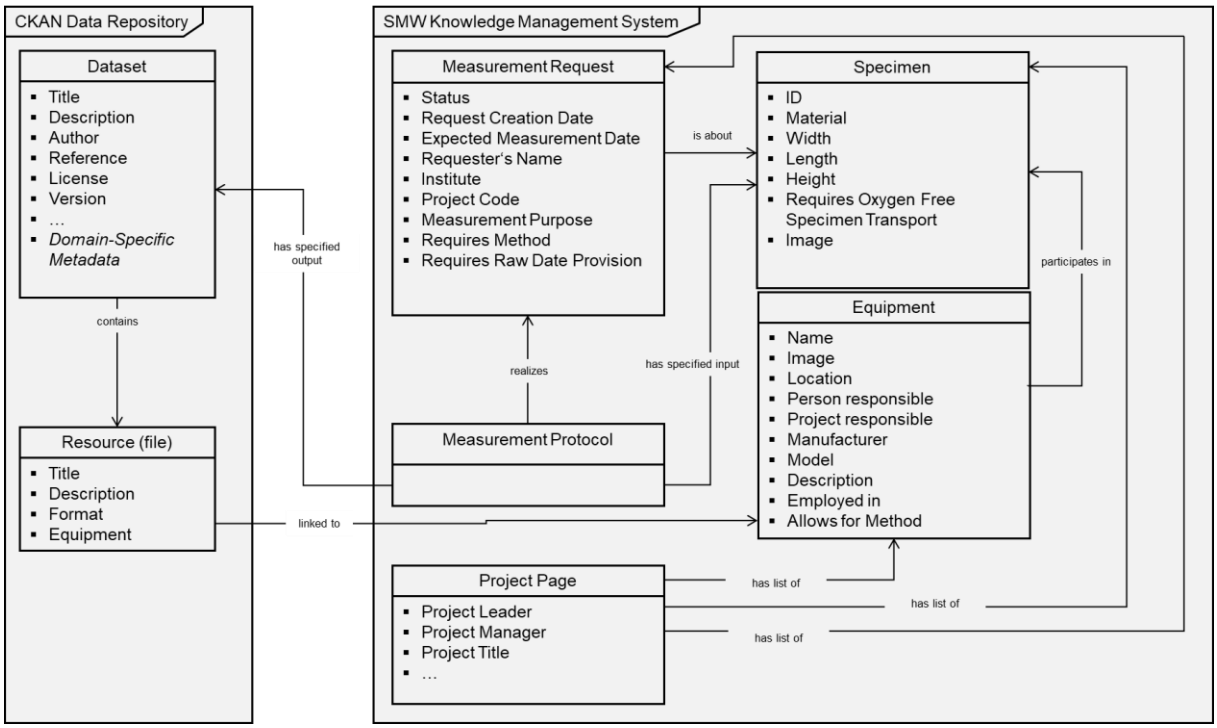

Figure 3. Relations between sub-systems in measurement request system

After the key entities are identified and their descriptions are completed, they are used in the next step of the stabishing of the Measurment Request System within the RDMS. First, a template for a MR is created is created within the SMW as represented in Figure 2. It partly involves cateries and their

https://doi.org/10.1017/pds.2022.54 Published online by Cambridge University Press

properties mentined belove devided in four theme-specific sections: The first section, Request Details, capture administrative information about the MR. The second section, Requesters Details, the information about which researcher and which research affiliation requested the measurement. The third section, Specimen Details, incorporates the description of specimen as an object of measurement to be performed. This section is implemented into the MR template based on a predefined query, that allows to automatically visualize information about specimen in form of a separate table on the relevant MR page after the MR is filled out and saved. The fourth section, Method Details, captures information about the method to be applied for the given MR.
When concidering the RDMS from a global point of view, the MR protocol implemented in SMW builds only a single part of the whole MR system and only realizes an indirect semantic connection of the sub-systems SMW and CKAN. Direct linking is provided between the data resources stored in CKAN and machines documented in SMW (Figure 3) through the usage of the plug-in Ckanext-Semantic-Media-Wiki (TIB, 2021). By storing dataset resources in CKAN, there ones is able to select for each of them a relevant equipment available in the SMW. After the dataset has been stored, the URL of the equipment is visualized on the relevant data resource page in CKAN. This URL serves as a redirection to the relevant machine page within SMW.

5. Conclusion and future work

This article presents a concept for implementing an RDM system according to the FAIR data principles for large collaborative projects. For the implementation, open source systems and their adaptation to project-specific requirements with the help of semantic annotation are proposed. The use of the RDMS has been presented here on the basis of a global measurement request system for the CRC1368. Actually, the expantion of the MR systems is planned, e.g. by semantic connection of the both sub-systems from the SMW side to the CKAN side. To do this, a new category, *Measurement Protocol*, has to be created, characterized, and interlinked to a CKAN dataset using the property *has specified output*. Via this category MR is then inderectly linked to all other knowledge elements within SMW, as well as to the appropriate datasets stored in CKAN, as it shown in Figure 3.
The advantage of presented procedure is its universality. The described use case can be generalized and mapped to other research domains with respect to the representation of samples, protocols, and equipment. So far, it has been successfully implemented in another collaborative project CRC1153 Tailored Forming for the integration of a digital machine park into an RDMS. In future contributions, the realization of further components within the RDM system is to be investigated and a methodical procedure for the implementation of the proposed RDM concept is to be provided.

Acknowledgement

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

References

Altun, O., Scheveleva, T., Castro, A. et al. (2021), "Integration eines digitalen Maschinenparks in ein Forschungsdatenmanagementsystem", Proceedings of the 32nd Symposium Design for X (DFX2021). DOI https://doi.org/10.35199/dfx2021.23.

Amorim R.C., Castro J.A., Rocha da Silva J. et al. (2017), "A comparison of research datamanagement platforms: architecture, flexible metadata and interoperability", Univ. Access Inf Soc, Vol. 16:851. DOI https://doi.org/10.1007/s10209-016-0475-y.

Bolser, D.M. et al. (2012), "MetaBase - the Wiki-Database of Biological Databases", Nucleic Acids Research, 40:D1250–D1254. DOI https://doi.org/10.1093/nar/gkr1099.

Cariaso, M., Lennon, G. (2012), "SNPedia: a wiki supporting personal genome annotation, interpretation and analysis", Nucleic Acids Research 40, D1:D1308–D1312. DOI https://doi.org/10.1093/nar/gkr798.

Ckanext-dcat [online], Available at: https://github.com/ckan/ckanext-dcat (accessed 15.11.2021).

Claus, F., Kirchmeyer, S., Müller, M.S., Richtering, W. (2019), "Das INF-Projekt Im SFB 985. Funktionelle Mikrogele Und Mikrogelsysteme", Bausteine Forschungsdatenmanagement, No. 2; pp. 104-111. DOI https://doi.org/10.17192/bfdm.2019.2.8097.

https://doi.org/10.1017/pds.2022.54 Published online by Cambridge University Press

Grönewald, M., Niekamp, R. (2020), "CRC/TRR 270 Z-INF - Inside a multidisciplinary joint project (1.0)", HeFDI Plenary 2020, Philipps-Universität Marburg. Zenodo https://doi.org/10.5281/zenodo.4808439.
Herzig, D.M., Ell, B. (2010), "Semantic MediaWiki in Operation: Experiences with Building a Semantic Portal", In: Patel-Schneider P.F. et al., editors. The Semantic Web – ISWC 2010. ISWC 2010. Lecture Notes in Computer Science 6497, Springer, Berlin, Heidelberg, pp. 114-128. DOI https://doi.org/10.1007/978-3-642-17749-1_8.
Huss, J.W., et al. (2008), "A Gene Wiki for Community Annotation of Gene Function", PLOS Biologue Vol. 6, No. 7, pp. 1398-1402. https://dx.doi.org/10.1371/journal.pbio.0060175.
Krötzsch, M., Vrandečić, D., Völkel, M. (2006), "Semantic MediaWiki", In: Cruz I. et al., editors. The Semantic Web - ISWC 2006. ISWC 2006. Lecture Notes in Computer Science 4273. Berlin, Springer, Heidelberg, pp. 935-942. DOI https://doi.org/10.1007/11926078_68.
Lagoze, C. and Sompel, H.V. (2002), "The Open Archives Initiative Protocol for Metadata Harvesting Protocol", Computer Science.
Maier, H.J. et al. (2020), "Towards Dry Machining of Titanium-Based Alloys: A New Approach Using an Oxygen-Free Environment", Metals Vol. 10, p. 1161. https://dx.doi.org/10.3390/met10091161. 2020.
MediaWiki contributors. Extension:Graph, [online], Available at: https://www.mediawiki.org/w/index.php?title=Extension:Graph&oldid=4764057 (accessed 15.11.2021).
MediaWiki contributors. Extension:LinkedWiki., [online], Available at: https://www.mediawiki.org/w/index.php?title=Extension:LinkedWiki&oldid=4669297(accessed 15.11.201).
MediaWiki contributors. Extention:Cargo. [online], Available at: https://www.mediawiki.org/w/index.php?title=Extension:Cargo&oldid=4690024 (accessed 15.11.2021).
Mozgova, I., Koepler, O., Kraft, A., Lachmayer, R., Auer, S. (2020), "Research Data Management System for a large Collaborative Project" DS 101: Proceedings of NordDesign 2020, Lyngby, Denmark, 12th-14th August, 12 pages. DOI https://doi.org/10.35199/NORDDESIGN2020.48.
Smith, M., et al. (2003), "DSpace: An Open Source Dynamic Digital Repository", D-Lib Magazine, Vol. 9, No. 1. DOI https://doi.org/10.1045/january2003-smith.
Szafarska, M., Gustus, R., Maus-Friedrichs, W. (2021), "Sauerstofffreier Transport, Präparation und Transfer von Materialproben für die Oberflächenanalytik", In: Clausthaler Zentrum für Materialtechnik (Hg.): Tagungsband 4 . Symposium Materialtechnik. Düren: Shaker Verlag, pp. 829–839.
Tansley, R., Harnad, S. (2000), "Eprints.org Software for Creating Institutional and Individual Open Archives", D-Lib Magazine, Vol. 6, No. 10, [online], Available at: https://www.dlib.org/dlib/october00/10inbrief.html#HARNAD (accesses 15.11.2021).
Technische Informationsbibliothek (TIB). (2021), Ckanext-Semantic Media Wiki. URL: https://github.com/TIBHannover/ckanext-Semantic-Media-Wiki (accessed 24.01.2022).
Vrandečić, D., Krötzsch, M. (2014), "Wikidata: a free collaborative knowledgebase", Communications of the ACM, Vol. 57, No. 10, pp. 78–85. DOI https://doi.org/10.1145/2629489.
Willmes, C., Viehberg, F., Lopez, S.E., Bareth, G. (2018), "CRC806-KB: A Semantic MediaWiki Based Collaborative Knowledge Base for an Interdisciplinary Research Project", Data Vol. 3. No. 4, 44. DOI https://doi.org/10.3390/data3040044.
Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", Sci Data 3, 160018. DOI https://doi.org/10.1038/sdata.2016.18.

https://doi.org/10.1017/pds.2022.54 Published online by Cambridge University Press