ENHANCED FINDABILITY AND REUSABILITY OF ENGINEERING DATA BY CONTEXTUAL METADATA

Altun, Osman (1);
Oladazimi, Pooya (2);
Wawer, Max Leo (1);
Raumel, Selina (3);
Wurz, Marc (3);
Barienti, Khemais (4);
Nürnberger, Florian (4);
Lachmayer, Roland (1);
Mozgova, Iryna (5);
Koepler, Oliver (2);
Auer, Sören (2)

1: Leibniz University Hannover, Institute of Product Development;
2: Leibniz Information Centre of Science and Technology University Library;
3: Leibniz University Hannover, Institute of Micro Production Technology;
4: Leibniz University Hannover, Institut für Werkstoffkunde (Materials Science);
5: Paderborn University, Data management in mechanical engineering

ABSTRACT

Complex research problems are increasingly addressed by interdisciplinary, collaborate research projects generating large amounts of heterogeneous amounts of data. The overarching processing, analysis and availability of data are critical success factors for these research efforts. Data repositories enable long term availability of such data for the scientific community. The findability and therefore reusability strongly builds on comprehensive annotations of datasets stored in repositories. Often generic metadata schema are used to annotate data. In this publication we describe the implementation of discipline specific metadata into a data repository to provide more contextual information about data. To avoid extra workload for researchers to provide such metadata a workflow with standardised data templates for automated metadata extraction during the ingest process has been developed. The enriched metadata are in the following used in the development of two repository plugins for data comparison and data visualisation. The added values of discipline-specific annotations and derived search features to support matching and reusable data is then demonstrated by use cases of two Collaborative Research Centres (CRC 1368 and CRC 1153).

Keywords: FAIR Data, Research Data Management, Knowledge management, Information management, Project management

Contact:
Altun, Osman
Leibniz University Hannover
Germany
altun@ipeg.uni-hannover.de

Cite this article: Altun, O., Oladazimi, P., Wawer, M. L., Raumel, S., Wurz, M., Barienti, K., Nürnberger, F., Lachmayer, R., Mozgova, I., Koepler, O., Auer, S. (2023) 'Enhanced Findability and Reusability of Engineering Data by Contextual Metadata', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.164

1 INTRODUCTION

In research projects from engineering sciences, large heterogeneous data sets are often generated as outcomes of experiments (e.g. data series, images, videos, etc.) as well as simulations (e.g. finite element analysis, CAD models, etc.). The organization of the resulting data sets is usually done on researcher's computer or institute's server in folder structures, sometimes supplemented by descriptions in so-called "readme-files" in a semi-structured way. This leads to challenges regarding the findability and re-usability of data both within the own organization, and even more for potential external researchers. Public data repositories are a first step in making data findable and available in the long term, but often requires a lot of additional effort by researchers. Data files need to be selected, assembled, described properly, and uploaded. Licenses regarding reuse conditions have to be defined. Furthermore, data repositories and metadata schema are often quite generic and domain-independent, which leaves the domain-specific context of the data generation unclear.

Thus, the need to organize data to ensure the storage, accessibility, and re-use of scientific data as a valuable resource is not yet fully part of the scientific culture, but it is becoming increasingly evident in the scientific community. Complex and interdisciplinary research problems require overarching collaborations and collective study of data from various domains. The efficient management of data and information generated and the extraction of knowledge are critical success factors for large collaborative projects (Kapogiannis 2018). The FAIR data principles (findability, accessibility, interoperability, and reusability) (Wilkinson et al., 2016) describe how data handling can be improved to address such challenges. Research data management (RDM) refers to the entire handling of research-related data, from the planning of data collection to its subsequent use (Büttner et al. 2011). In the past, several infrastructures and services have been developed to establish RDM. The German National Research Data Infrastructure (NFDI) aims to link existing infrastructures and services and close remaining gaps with additional services for all scientific disciplines (Hartl et. al., 2021). Research data management systems (RDMS) support research projects in processing, managing, and depositing data collections (Sheveleva et al., 2022), and the comprehensive descriptions of data, thus improving the overall quality of data management and accessibility. Heterogeneous data types and formats, duplicate, incomplete or missing or inconsistent data or data descriptions (Effertz, 2010), as well as the lack of standardized processes and common platforms for accessing and working with data, are some of the most common problems. These problems are particularly evident at the level of collaborative projects involving different organizational units, as is the case in Collaborative Research Centres (CRCs). In CRCs a large number of projects from different sub-disciplines work together on an overarchring research problem (Mozgova et al., 2020). Each subproject generates, processes, and analyses data using different procedures and methods (Sandfeld et al., 2018). Harmonization and accessibility of data are critical success factors for a comprehensive view on data across projects of collaborative research projects and their subsequent reuse.

In previous publications we introduced an implementation of the open-source data repository Comprehensive Knowledge Archive Network (CKAN) as a part of a research data management system for large collaborative projects (Mozgova et al., 2022; Sheveleva et al., 2022). We use the data repository in CRC 1368 (Collaborative Research Center 1368, Oxygen-free Production) and CRC 1153 (Collaborative Research Center 1153, Tailored Forming), two different collaborative projects with a larger number of sub-projects. In this paper, we present an approach to provide more discipline-specific metadata annotations of data and ease-to-use workflows for the generation of such metadata to increase the FAIRness. The identification of such specific, shared metadata concepts in collaborative projects, the implementation of automatic data annotations as well as the resulting use in terms of advanced search functions, data visualization, and data comparison features are presented in the following sections of this paper.

2 STATE OF THE ART - RESEARCH DATA MANAGEMENT SYSTEMS

In (Amorim et al. 2017) well-known RDM systems that can enable researchers and institutions to distribute data and support the RDM workflows are described. Data repositories play an essential role in any RDM system alongside knowledge management systems. An ideal data repository allows users to publish their data according to the FAIR Data principles where the data is findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). As a prerequisite for any re-use of data, data needs to be findable first. In this case, data and metadata must be found by both humans and machines.

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

Machine-readable metadata is essential for the automatic searchability of data. Accessible means that users need to know how to access the found data. Interoperability refers to the need to integrate data with other data and with other applications and workflows. Reusable refers to the fact that metadata and data should be sufficiently described to be able to replicate or combine them with other settings. Optimizing the reuse of data is seen as the ultimate goal of FAIR. Repositories need to be designed in a FAIR way, enabling and allowing for the execution of the principles in publishing and sharing data (Devaraju et al., 2021). Many data repository platforms have been developed in the past to support researcher to publish their data as FAIR as possible. Based on the usage by researchers and institutions, CKAN (ckan.org), Zenodo (zenodo.org), EPrints, and DSpace are the most notable ones (Amorim et al, 2017). The mentioned platforms allow the users to annotate their data based on a predefined vocabulary that provides a harvesting opportunity for other platforms with the same annotation. For instance, in CKAN and DSpace, one can export metadata in RDF (Resource Description Framework) format and import it in another data repository instance. As mentioned, CRC utilizes CKAN as the data repository (Mozgova et al., 2022; Sheveleva et al., 2022). CKAN is developed by the CKAN Association, a member of the Open Knowledge Foundation, and is in use on over 1000 data portals worldwide, such as data.gov and data.gov.uk. or in terms of research data for example in the Leibniz Data Manager[1] (Beer et al., 2022) or the European Data Portal[2].
The usage of controlled vocabulary for annotation is a core practice of data repository platforms that aim to comply FAIR data principle. Annotation based on a vocabulary makes the data understandable by both humans and machines. Besides, annotation can help to derive hidden relations and meaning between a variety of data gathered from different resources. Multiple data vocabularies have been developed for describing the data in different areas. Data Catalogue Vocabulary (DCAT) (Maali and Erickson, 2020) is one of the noteworthy ones since it allows to description the data in governmental institutions, and recently also in scientific communities. For example, users of CKAN can utilize an extension to export and annotate their data based on DCAT. Within the international repository directory “re3data.org” are over 500 repositories related to engineering sciences listed, approximately 80 of them managed in Germany. These repositories show a high heterogeneity concerning the procedures used for data archiving and publication, metadata standards, and persistent identifier systems used. The published datasets often use generic metadata standards (e.g. Dublin Core (DC), DCAT, DataCite, ISO 19115, DDI, etc.) (Koppe et al., 2015) or metadata standards from other disciplines (e.g. Data Cube Vocabulary, CIF, netCDF). While the flexibility remains high, the use of generic metadata significantly limits the findability of discipline-specific content (Altun et al., 2021).

3 RESEARCH BACKGROUND AND PROBLEM ANALYSIS

As described, various data repositories and data management systems exist, but these often have the shortcomings of not being designed domain-specifically for engineering requirements. Subsequently, the possibilities of annotating the data sets usually do not go beyond generic approaches (e.g. custom key/value-pairs, read-me files, etc.) and have to be carried out manually by the researchers. This leads to usability challenges for researchers, as they are conventionally used to store data without an extended approach in terms of contextualization. As a result, it becomes increasingly challenging to implement RDM according to the FAIR data principles from the user's point of view and to ensure findability, accessibility, interoperability, and reusability for other researchers as well as stakeholders, especially in the case of the generation of large heterogeneous data volumes in collaborative projects.
Our use cases involve CRC projects (CRC 1368 & CRC 1153), where about 80 researchers from various disciplines are involved. In CRC 1368 processes in an oxygen-free atmosphere are being researched to develop sustainable production technologies. In CRC 1153, the design of novel manufacturing process chains using the tailored forming technology to manufacture hybrid solid components with locally adapted properties is the research goal. We can conclude, that both CRCs share common metadata concepts on the level of project representations, data creators in multiple subprojects and the general description of data types and formats. A detailed representation of the context of data generation and processing requires domain-specific vocabularies and metadata for each CRC and the respective subproject.

[1] https://service.tib.eu/ldmservice/service/
[2] europeandataportal.eu

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

Thus, the requirement for supporting contextual engineering metadata in the CKAN data repository is detected in both CRC projects. The requirement is in particular notable due to the system's inability to let users define the contextual metadata for each data resource (file). Contextual metadata is mostly referring to the process parameter(s) describing the condition of data generation. The key usage of these metadata in the CRC project is to increase the searchability of the research data by benefiting from the them in shape of an advance search or facet filters. One can argue that the description field in CKAN can be used for noting the contextual metadata. However, descriptions are text-based metadata, and writing a long text makes it cumbersome to look for data based on a specific contextual metadata field value. Moreover, CKAN allows users to define custom metadata for their dataset in the shape of "key:value" where the key is the metadata's name and the value is the metadata's value. However, this prevents users from defining data-resource-level custom annotations. As a use-case, in the CRC project, a dataset often contains a high number of data resources. Each one of these data resources can have its contextual metadata. Moreover, the key:value feature is completely manual and can be named anything by the researcher who uploads the data. This makes it difficult to search based on the contextual metadata due to name inconsistencies. For example, one researcher can name a contextual metadata "material" and the other can name the same metadata as "material_used". Although both metadata keys are referring to the same concept, it would be challenging to detect this since they have different titles.

Data repositories allow an easy access to a large amount of published data for one's own research. Visualization is a common practice to scan and analysis the data to find the related ones. In particular, data comparison via visualization is a vital practice in research projects and scientific approaches (Gleicher, Michael, et al., 2011). The reason for this, that there are many column data in CSV/XLSX format that researchers need to compare to identify the related data and perform knowledge exchange between different research project. To support visualization, CKAN has a feature that let users visualize one data resource. However, the process is completely manual and requires the researchers' knowledge regarding the context of the data columns. Besides, the comparison between multiple column data resources is not supported. This is in particular notable in numeric-based data resources such as CSV files which researchers utilize to put different variables measurement data. The issue here is that the context of the variables (data columns) is not clear in a way that a machine (software) or a human could visually plot and compare them. For example, what would be the dependent and independent variables? To have a high-quality visual comparison, both human users and machines (in the case of auto-plotting) need to understand the data context. The inability of the metadata in CKAN to explain the data resources (files) context is an issue for data visualization in CKAN. Failing to recognize the content of a data column can interrupt the research activity chain due to human error and ambiguity. Also, it increases the need for human interaction which increases the research project's time and effort. The metadata regarding the column data can be used to compare the data by visualization, enhancing the interoperability and re-usability concerning FAIR.

Based on the described analysis of the current state above and with regard to the enhancement of data management systems in terms of the FAIR data principles, the following research questions have been derived, which will be addressed in the next sections of this paper:

- How can research data be contextualized in an RDM system to enrich the data fairness concerning the FAIR data principles?
- How can researchers be efficiently supported in their search for relevant research data and findings of other researchers?
- How can researchers be efficiently supported in visual comparing of different data resources?

4 ANNOTATING DATA TO ENSURE FAIR-DATA PRINCIPLES - USE-CASE CRC

Both CRCs are from the domain of engineering sciences and have a similar organizational structure. The difference, besides the research questions and goals in the specific context of this work, lies in the definition of vocabulary for contextualized metadata. While in CRC 1368 the metadata "Material Combination, Atmosphere, Data Type, Surface Preparation and Analysis Method" is taken as the basis for further implementations in the RDM System, in CRC 1153 "Material Combination, production process" is used. General metadata, such as project number, project manager, organizational affiliation, etc. are of the same type for both projects. The contextual metadata mentioned above were identified together with the researchers of the subprojects in surveys and workshops. As an example,

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

the development and implementation of the plug-ins for metadata annotation to enable the RDM system in terms of enhanced FAIRness, representative for both CRCs, are presented below using the contextual metadata of CRC 1368. Table 1 shows the mentioned metadata with their description and their potential values in both CRCs.

Table 1. CRC specific contextual metadata

Contextual metadata (CRC 1368)	Examples of value	Contextual metadata (CRC 1153)	Examples of value
Material(-combination)	aluminium, copper, titan	Material(-combination)	steel, aluminium, titan
Atmosphere	argon, argon-silan, air	Production process	impact extrusion, turning, deep rolling
Data Type	mechanical, chemical	Analysis method	metallography, hardness, dilatometer
Surface Preparation	spinn coating, grinding	Demonstrator	shaft, bearing bushing, bearing washer
Analysis Method	x-ray microscopy, tensile test		

As mentioned, the CRC utilizes the CKAN data repository in its RDM system. CKAN's core component is the dataset. For example, the search result in CKAN only contains the datasets that are matched with the search query. A dataset in CKAN is a group of data resources (files). A data resource can be any file format such as CSV, PNG image, PDF, etc. Moreover, CKAN has open-configurable interfaces (APIs), and also extensibility of the service via plugins and domain-specific vocabularies. Thus, the service can be extended with relevant tools to a specific domain that for example allow the run-ability of software codes (e.g. Jupyter notebooks) or the visualization of engineering research data (e.g. AutoCAD .dwg files or other data output and inputs from mechanical experiments and simulations). A visualization tool for example can be used to check data resources preview for their significance to one's research without the necessity to download to the researcher's local computer in advance, which is a time-consuming process. However, the auto- preview option for column-based data is not supported and needs to be carried out manually by the researcher.

4.1 Annotating data resources with contextual metadata

The goal of the approach is to increase the granularity of provided metadata by Domain-specific metadata and also increase usability via supporting the user by providing a data upload standard for the sake of name consistency and automated metadata extraction. To address the lack of context in the data resources, data resource level annotations are adopted to CKAN. The approach focuses on three aspects to increase the granularity and richness of metadata. The first aspect is aimed to extract metadata automatically and directly from a predefine metadata profile. This aspect saves researchers from manually entering the metadata on each data upload. Also, it eliminates the chance of user errors (for example, entering a wrong metadata name). This requires the collection of additional domain-specific metadata based on the predefine metadata profile and controlled vocabulary. This aspect concerns the contextual metadata about a data resource in CKAN obtained from the experiment in the CRC project. Compared to the "key:value" metadata fields in CKAN, this approach guides the user and creates a consistency to name the metadata. Besides, vocabulary support provides a shared understanding of the data among researchers and sub-projects in CRC.

The second aspect concerns the enrichment of data findability via enhancing the data repository search functionality. This aspect benefits from the extracted metadata from the predefine metadata profile to enable researchers to perform an advanced search based the contextual metadata to identify the related data resources. The third aspect is aimed to define X/Y-categories to the data columns in CSV/XLSX data resource formats. The idea is to make the data column understandable for the visualization purpose. In this regard, two annotations are used: X-category and Y-category. X-category indicates that the column is representing an independent variable (usually shown on the abscissa) and Y-category indicates that the column is representing a dependent variable (usually shown on a plot ordinate). The X/Y categories annotation help the researchers to visually compare several data resources by plotting these in one plot. Besides, it enables the machines (for example a CKAN plugin) to automatically plot the data that helps a researcher to quickly get insights from the target data resource. This also saves the researchers from downloading and plotting the data which is time/energy consuming. It is worth mentioning that CKAN provides a plotting tool to visualize one data resource. However, the process is completely manual that requires also a background regarding each column's context.

The mentioned annotations are critical in a research project since a researcher is not necessarily aware of other researchers' data contexts. Using these annotations makes the column data self-explanatory reducing the need for writing extensive and long descriptions and also reducing the need for a human presence to clarify the data.

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

4.2 Implementation into the data repository

To implement the data resource annotation, CSV/XLSX data formats are targeted. For automated processing of data files and metadata extraction, a common standard template for the column data is required. Therefore, a predefine metadata profile has been created and distributed among the researchers that they could use to enter both the numeric data for X/Y categories and set the mentioned metadata in the last section of this paper. Using the predefine metadata profile saves the researchers in CRC from annotating the data manually. Table 2 shows the predefine metadata profile template.

Table 2. Predefine metadata profile for annotation in Project CRC 1368

X-Category	Y-Category	Analysis Method	Material (-combination)	Atmosphere	Data type	Surface Preparation
Caption_unit	Caption_unit	e.g. XPS	e.g. Aluminium	e.g. Air	e.g. Mechanical	e.g. Spin coating
x-Data 1	y-Data 1					
x-Data 2	y-Data 2					
x-Data 3	y-Data 3					
x-Data 4	y-Data 4					
..	..					
..	..				Data Ressource	
x-Data n	y-Data n				Metadata description	

As part of the implementation, metadata is extracted and ingested into CKAN during the upload process. A CKAN plugin has been developed to achieve the goal. The plugin has two main functions. First, it extends the data resource metadata schema stored in the CKAN database. Second, it reads and extracts the metadata automatically and ingest them into CKAN. It is worth mentioning that it is also possible for researchers to manually enter the contextual metadata after data resource upload without using the predefine metadata profile in case they prefer not to use it.

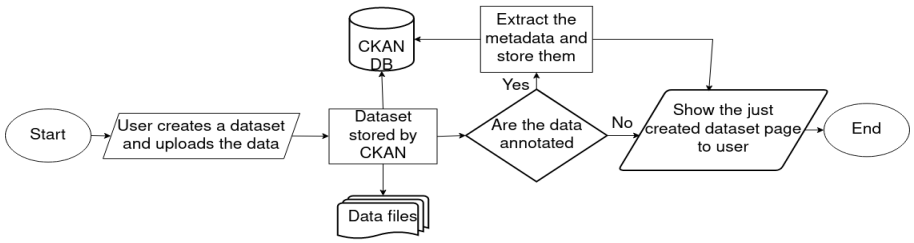

Figure 1. Implemented annotation extraction workflow in CKAN

4.3 Features resulting from the data annotation

The data resource annotation approach is used in two features in CKAN data repository: Search and Data Visualization. The main goal here is to enhance the data repository concerning the FAIR principle to support research project's objectives.

4.3.1 Enhanced findability and accessibility of data resources of interest via extended search functions

For the first application, a CKAN plugin[3] has been developed to enable users to search for datasets and data resources based on the extracted metadata profile explained in table 1. The plugin customizes the CKAN search template view and adds a drop-down menu to it where users can select the target metadata. After a user enters the search term and selects the target metadata, the plugin searches in the target metadata value among existing data resources. Finally, the plugin returns the search result to the user. The search result is a list of datasets whose data resource(s) match the entered search term for the specified contextual metadata. This way a researcher can further narrow down the result to increase the search precision.

As mentioned in section 4, in CKAN, the search result contains the list of target datasets. In CRC projects, like many other research projects, datasets are usually comprised of a number of data resources. As a result, returning only the dataset list for the search result is problematic since the search target here is the data resource in the dataset. Therefore, the plugin also shows the matched data resources for each dataset in the result list. Listing the data resources in the search result saves the

[3] Search Plugin: https://github.com/TIBHannover/CKANext-sfb-search

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

researchers from opening the dataset page in CKAN to manually find the target data resource. Figure 2 shows the design and workflow for the search plugin in CKAN.

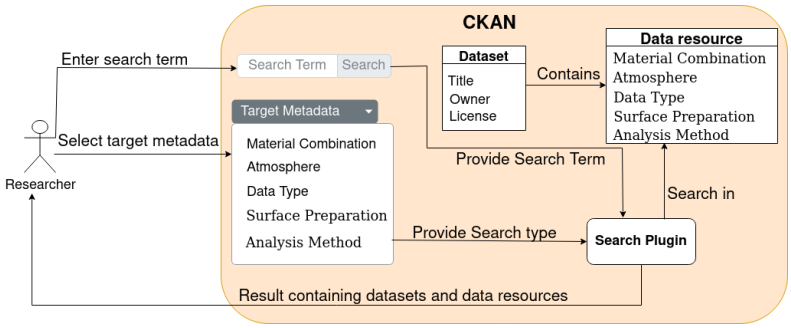

Figure 2. Design and workflow for the implemented search plugin in CKAN

4.3.2 Data visualization and comparison based on metadata annotation

The second plugin[4] developed based on the provided annotation is the Data-Vis. The goal of the Data-Vis plugin is to facilitate data visualization in CKAN by using column annotation in the metadata profile. As mentioned in section 4.2, the user who uploads the data (.csv/.xlsx) annotates the data column with X/Y categories. The Data-Vis plugin utilizes these annotations in two ways: Data preview and Data comparison. The goal here is also to improve the data repository concerning the FAIR principle by increasing the data understandability and readability via visualization for researchers. Users can benefit from the data preview feature in the plugin by opening a data resource page in CKAN. After they open the target page, the plugin checks the annotation existence for the data resource. If the annotation exists, then the plugin draws a plot for that data resource automatically. Compared to the CKAN standard data visualization which is manual, this approach benefits from the existing annotations, provided by the researchers who uploaded the data, to automatically plot the data preview without the researcher intervention. As a new feature for CKAN and data repositories, the Data-vis plugin also enables researchers to compare data resources via visualization. This saves the researchers from downloading a group of data resources and visualize them in extended applications. The data visual comparison acts as an enhanced description for a data resource that increases the data understandability and re-usability in research projects with respect to the FAIR principle. When a user opens a data resource page in CKAN, the plugin provides a button that he/she can use to enter the comparison space. In this space, the target data resource columns are already shown alongside an "import" button that users can utilize to import other column data (within the same dataset or from other datasets) to compare. Users can select the column for the x-axis with double clicks and y-axis with a single click for the comparison plot. If the column data is annotated based on X/Y-categories, then the annotations get shown to users to inform about the context of the target columns (dependent or independent variable). Besides, if all imported column data resources are annotated, the users do not have to choose the X-axis and Y-axis variables from the data and can directly jump to the plot since the plugin knows what goes where by reading the annotation. The implementation design and usage workflow are shown on Figure 3.

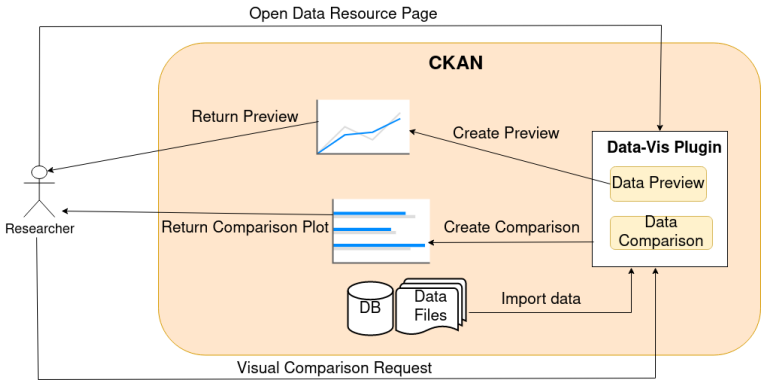

Figure 3. The implementation design and usage workflow for the data-vis plugin in CKAN.

[4] Data-Vis Plugin: https://github.com/TIBHannover/CKANext-data-comparison

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

5 USE-CASE WITHIN CRC 1368

In this section, the above-described data annotation approach and the resulting new features of the repositories are demonstrated with a use-case of two subprojects of CRC 1368. Both sub-projects investigate surface roughness as related to various surface treatments in oxygen-free atmospheres from a different point of view. While one project is more application-oriented and focuses on the topic of roll bonding (Hordych et al., 2021), the other project is foundation-oriented in terms of surface effects (Raumel et al., 2021). In general, the comparison is often difficult and therefore of special interest. The values to be investigated are the normalized deoxidation time and the arithmetic mean surface roughness (R_a). From the perspective of one of the users the *DataVis*-tool can be utilized as following: Subproject 1 (SP1) performed tests with the material copper and obtained R_a over the deoxidation time (table 3). To have a certain arithmetic mean roughness value the samples were initially polished. The roughness values were measured by means of atomic force microscopy. Subsequently, deoxidation was performed by plasma. The data has been uploaded to the repository using table 2 as template.

Table 3. Data resource of SP1: Plasmadeoxidation of copper samples

X-Category	Y-Category	Analysis Method	Material (-combination)	Atmosphere	Data type	Surface Preparation
Norm. deoxidationtime_s	R_a_nm	Atomic force microscopy	Copper	Argon	Mechanical	Plasmadeoxidation
0	20					
0.05	21.30					
0.1	21.80					
0.15	23.20					
0.2	26.40					
0.3	27.50					
0.5	21.40					
1	19.20					

For further viewing and comparison of similar data sets already existing in the system, the researcher can search after contextual metadata of interest. Thus, further data resources from subproject 2 (SP2) appear as shown in tables 4. Table 4 shows the roughness of mechanically deoxidised aluminium and copper samples. These were mechanically deoxidised in a grinding process. HNO_3 and NaOH were used for the chemical surface preparation of copper and aluminium accordingly. For the surface roughness measurement, a confocal 3D laser scanning microscope was used in all data resources of SP2.

Table 4. Data resources of SP2: Chemical deoxidation of aluminium and copper samples as well as mechanical deoxidation of aluminium.

X-Category	Y-Category	Analysis Method	Material (-combination)	Atmosphere	Data type	Surface Preparation
Norm. deoxidationtime_s	R_a_nm	3D laser scanning microscopy	Copper	XHV	Chemical	Chemical HNO_3
0	99					
0.3	101.5					
0.5	110					
1	118					
Norm. deoxidationtime_s	R_a_nm	3D laser scanning microscopy	Aluminium	XHV	Chemical	Chemical NaOH
0	122					
0.17	203					
0.33	257					
1	236					
Norm. Deoxidationtime_s	R_a_nm	3D laser scanning microscopy	Aluminium	XHV	Mechanical	Grinding
0	122					
0.25	1294					
0.5	1845					
0.75	1378					
1	595					

The obtained comparative graphs in CKAN for copper and aluminium samples at different surface preparations are shown in Figure 4. Comparing different possibilities of surface deoxidation offers the researcher insight into the surface characteristics and helps to identify relevant processing parameters for processes such as for example roll bonding where roughness is influencing the quality of a products bonding zone. Thus, the researcher quickly obtains a graphical comparison in the data repository and can identify previously unknown trends. Related research can be found and combined with one's own research in order to increase the domain knowledge.

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

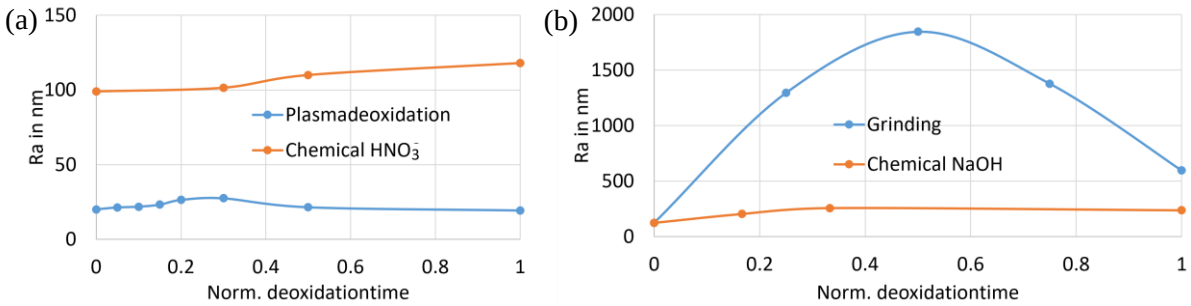

Figure 4. Roughness over deoxidationtime as results of different surface preparations for copper (a) and aluminium (b) samples.

6 CONCLUSION AND FUTURE WORK

Although data repositories are increasingly available to meet the requirements for providing FAIR data we have identified some shortcomings in the generation of comprehensive FAIR data and their searchability from a discipline-specific perspective. The need for contextual metadata has become evident during the implementation of data repositories in two CRC projects. While domain specific metadata enhances the FAIRness of data and enables improved searchability and therefore findability it adversely increases the time and efforts to provide these metadata. In this paper, we present an approach how to extend generic metadata with the contextual, discipline-specific metadata for CKAN data repository.

To minimize researchers' work to provide the extended metadata we have developed a CKAN plugin to support automated metadata extraction from data templates in CSV/XLSX format. The enriched metadata are in the following used by two developed plugins of the data repository to improve the precision of search and secondly to provide a data comparison and data visualization tool. Both features support the researcher in their efforts to find matching data in a large data pool and perform a first evaluation of their reusability. Benefits of the implementations are shown in the use cases of subprojects of a collaborative research centre. Future work will focus on extending contextual metadata to improve data visualization. With regards to linked data and knowledge graphs we will work on a DCAT application profile to include the discipline-specific metadata into the generation of RDF (Resource Description Framework). Furthermore, it could be investigated to what extent AI technologies may further reduce the user's involvement in metadata annotation in order to increase flexibility in data storage. For this purpose, the use of controlled vocabulary in the documentation for corresponding search engines and pattern recognition processes could be considered.

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 394563137 – SFB 1368 and
by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 252662854 – SFB 1153.

REFERENCES

Altun, O., Scheveleva, T., Castro, A., Oladazimi, P., Koepler, O., Mozgova, I., Lachmayer, R., and Auer, S. (2021), "Integration eines digitalen Maschinenparks in ein Forschungsdatenmanagementsystem", *Proceedings of the 32nd Symposium Design for X (DFX2021)*. https://doi.org/10.35199/dfx2021.23

Amorim, R.C., Castro, J.A., Rocha da Silva, J. and Ribeiro, C. (2017), "A comparison of research datamanagement platforms: architecture, flexible metadata and interoperability", *Univ. Access Inf Soc*, Vol. 16, pp.851–862. https://doi.org/10.1007/s10209-016-0475-y

Beer, A., Brunet, M., Srivastava, V. and Vidal, M.E. (2022), "Leibniz Data Manager – A Research Data Management System", In: Groth, P., Rula, A., et al., *The Semantic Web: ESWC 2022 Satellite Events. ESWC* 2022. *Lecture Notes in Computer Science*, Vol. 13384, Springer, Cham, pp. 73–77. https://doi.org/10.1007/978-3-031-11609-4_14

Büttner, S., Hobohm, H.-C. and Müller, L. (2011), *Handbuch Forschungsdatenmanagement*, Bock + Herchen Verlag, Bad Honnef.

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Akerman, V., L'Hours, H., Davidson, J. and Diepenbroek, M. (2021), "From Conceptualization to Implementation: FAIR Assessment of Research Data Objects", *Data Science Journal*, Vol. 20, No. 4, pp. 1–14. https:// doi.org/10.5334/dsj-2021-004

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. http://dx.doi.org/ 10.5880/TR32DB.KGA90.7

Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C.D. and Roberts, J.C. (2011), "Visual comparison for information visualization", *Information Visualization*, Vol. 10, No. 4 pp. 289-309. https://doi.org/ 10.1177/1473871611416549

Hartl, N., Wössner, E. and Sure-Vetter, Y. (2021), "Nationale Forschungsdateninfrastruktur (NFDI)", *Informatik Spektrum*, Vol. 44, No. 5, pp. 370-373, Springer. https:// doi.org/10.1007/s00287-021-01392-6

Hordych, I., Barienti, K., Herbst, S., Maier, H.J. and Nürnberger, F. (2021), "Cold Roll Bonding of Tin-Coated Steel Sheets with Subsequent Heat Treatment", *Metals*, Vol. 11, p. 917. http://dx.doi.org/ 10.3390/met11060917

Kapogiannis, G. and Sherratt, F. (2018), "Impact of integrated collaborative technologies to form a collaborative culture in construction projects", *Built Environment Project and Asset Management*, Vol. 8 No. 1, pp. 24-38. https://doi.org/10.1108/BEPAM-07-2017-0043

Koppe, R., Gerchow, P., Macario, A., Haas, A., Schäfer-Neth C. and Pfeiffenberger, H. (2015), "O2A: A generic framework for enabling the flow of sensor observations to archives and publications," *OCEANS* 2015, Genova, pp. 1-6. https://doi.org/10.1109/OCEANS-Genova.2015.7271657

Lamprecht, A.-L. et al. (2017), "Towards FAIR Principles for Research Software", *Data Science Journal*, Vol.3, No.1, pp. 37-59. http://dx.doi.org/10.3233/DS-190026

Maali, F. and Erickson, J. (2020), *Data Catalog Vocabulary (DCAT) - Version 2*. W3C Recommendation [online], Available at: https://www.w3.org/TR/vocab-dcat-2/ (02.11.2022).

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

Mozgova, I., Altun, O., Sheveleva, T., Castro, A., Oladazimi, P., Lachmayer, R. and Auer, S. (2022), "Knowledge Annotation within Research Data Management System for Oxygen-Free Production Technologies", *Proceeding of Design Society* 2022, Vol. 2, pp. 525-532. http://dx.doi.org/ 10.1017/pds.2022.54

Raumel, S., Barienti, K., Dencker, F., Nürnberger, F. and Wurz, M.C. (2021), "Einfluss von Silan dotierten Umgebungsatmosphären auf die tribologischen Eigenschaften von Titan", *Tribologie und Schmierungstechnik*, Vol. 68, pp. 5–13. http://dx.doi.org/10.24053/tus-2021-0002

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. and Schubotz, R. (2018), *Strategiepapier Digitale Transformation in der Materialwissenschaft und Werkstofftechnik*. Deutsche Gesellschaft für Materialkunde. Available at: https://edocs.tib.eu/ files/e01fn18/1028913559.pdf. (01.11.2022).

Schultes, E. and Wittenburg, P. (2019), "FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure", In: Manolopoulos, Y., Stupnikov, S. (eds) *Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science*, Vol. 1003, Springer, Cham. https://doi.org/10.1007/978-3-030-23584-0_1

Sheveleva, T., Wawer, M.L., Oladazimi, P., Koepler, O., Nürnberger, F., Lachmayer, R., Auer, S., Mozgova, I. (2022), "Creation of a Knowledge Space by Semantically Linking Data Repository and Knowledge Management System - a Use Case from Production Engineering", *IFAC-PapersOnLine*, Vol. 55, No. 10, pp. 2030-2035. https://doi.org/10.1016/j.ifacol.2022.10.006

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, Vol. 3. https://doi.org/10.1038/sdata.2016.18

https://doi.org/10.1017/pds.2023.164 Published online by Cambridge University Press