Research Data Management with ckan and Semantic Mediawiki

RDM Service Development at Lab Linked Scientific Knowledge

The Lab Linked Scientific Knowledge's (**LSK**) main objective is to investigate, facilitate and build Research Data Management (**RDM**) in different research areas such as Chemistry and Engineering. The main objective is to make Research Data **FAIR** that stands for:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

The Lab activities are:

- Research and Develop Semantic Artifacts such as Ontologies and Vocabularies
- Research and Develop Terminology Services for making the Semantic Artifact FAIR
- Research and Develop Research Data Management Systems (**RDM**)

Projects

At the moment, LSK is responsible for providing the infrastructure and investigating the bests practices for two CRC (Collaborative Research Center) projects:

CRC1153: Vocabulary-oriented research data management for tailored forming process chains (https://www.sfb1153.uni-hannover.de/en/)

CRC1368: research data management for Oxygen-Free production (https://www.sfb1368.uni-hannover.de/en/sfb-1368/profile/)

The RDM system for the mentioned projects is aimed to:

- Semantically describe the Research data
- Create a semantic profile for the domain-specific metadata
- Make the RDM FAIR by benefiting from Knowledge management systems and Data repositories
- Make the resulting **Linked Data** accessible to researchers through a **Knowledge Graph**

Technology Stack

The base platforms for these RDMs are:

- Semantic Media Wiki (**SMW**) is the Knowledge Management system. https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki
- The Comprehensive Knowledge Archive Network (**CKAN**) is the data repository. https://ckan.org/

RDM characteristics and methods

Although using CKAN and SMW is a big step toward FAIR data, it is often **not** enough to have a functioning FAIR Research Data Management system.

To make these systems suitable for RDM, these extra steps need to be taken:

SMW:

- We need to **identify** our research data **context** in our target scientific **domain**. By context, we mean all the related machines, samples, lab protocols, and generally everything that is related to the experiment that is not reflected in the data itself.
  - Example: John Doe performs an experiment in the lab and generates some data and uploads them to the data repository in CKAN. However, we have no idea what was the process by which this data is generated, As a result, the data is not understandable.
- The next step is to **semantically describe** the identified contextual data in the previous step. Here basically we model our data and annotate them to be **machine-actionable** also.

  - **Why annotated?** Humans are not the only data users. As matter a of fact, soon mostly machines (**AI** for instance) are supposed to perform **actions** on data. Without annotation, the machine's precision and comprehension weaken.
  - **How to annotate?** Ontologies and Vocabularies are rich sources to find domain-specific annotation.
  - **Where to find these annotations?** Terminology Services. TIB already has one: https://terminology.tib.eu/ts
  - **What if I cannot find a proper annotation?** Contribution. You can develop your own vocabulary for your specific domain. Benefit? people in your domain will use it.
- RDM has to serve **Linked Data.** The last step is to **Link** your contextual metadata and data. This means you need an actual link that connects your SMW graph to the CKAN graph. For example a link from a **Sample** page in SMW to the related **dataset** in CKAN.

CKAN:

- CKAN is a ready-to-use data repository. You can just install it and manage your datasets. However, **not enough** for RDM. The general extra steps to be taken are:
  - We need to **link** our data to the contextual concepts that we implemented in **SMW** before. Like linking your dataset to the corresponding machine/device that generated it. Example extension developed by Lab LSK: ckanext-Semantic-Media-Wiki
    - Note: you do not have to make a link from CKAN to SMW and vice-versa at the same time. A one-way link is enough. The direction is your choice.
  - I**dentifying other contextual metadata** for your dataset in CKAN. It is true that we have linked our data to the context in SMW. But there are some other contextual metadata that are not in SMW (They should not be).
    - For example, what is the **publication** related to this dataset? For instance, Lab LSK developed a plugin for linking publication(s) to your dataset: ckanext-Dataset-Reference
    - You can also define some custom metadata for your dataset in ckan by extending the CKAN schema. Example from Lab LSK (plugin **crc1153_dcat_ap**): https://github.com/TIBHannover/ckanext-crc1153
      - **Scenario**: let's say your dataset has domain-specific metadata named 'Temperature'. CKAN does not support this and as a result, you need to extend the CKAN schema.
  - Here we also need to **annotate** our data. CKAN provides the possibility to export your dataset metadata in RDF format based on **DCAT** https://github.com/ckan/ckanext-dcat
  - However, just exporting in DCAT format is **not** enough!
    - DCAT only describes your dataset in **a general context**. What about those contexts that you added in the **previous steps?**
    - **Solution**: Extend DCAT to add your domain metadata (your custom **annotation**). An example from lab LSK (plugin **crc1153_dcat_ap**): https://github.com/TIBHannover/ckanext-crc1153
    - **Where to find these annotations?** Terminology Services. TIB already has one: https://terminology.tib.eu/ts
    - **What if I cannot find a proper annotation?** Contribution. You can develop your own vocabulary for your specific domain. Benefit? people in your domain will use it.

Here are presentation slides that describe the overall approach and the System design

SFB CKAN 26.06.2023.pptx

A short Demo of the system (Focus on the Data repository)

ckan.mp4

Source Code for Plugins

Available on Lab LSK github in TIB Hannover.

https://github.com/orgs/TIBHannover/teams/lab-linked-scientific-knowledge/repositories

Plugin	Extension Repository	Published to CKAN extension registry	pypi	Description
`multiuploader`	ckanext-multiuploader	Requested https://github.com/ckan/extensions.ckan.org/issues/150	Yes Link	ables users to upload multiple resources at once with drag&drop.
`tif-imageview`	ckanext-tif-imageview	Requested https://github.com/ckan/extensions.ckan.org/issues/149	Yes pypi link	converts the target tif image to the jpeg format for showing the preview of the image. It does not change or replace the original tif image.
`email-notification`	ckanext-email-notification	No		sends user registration e-mail to system admins
`feature_image`	ckanext-feature-image	Requested https://github.com/ckan/extensions.ckan.org/issues/151		ables system admins to upload a featured image and caption for the CKAN homepage.
`media_wiki`	ckanext-Semantic-Media-Wiki	No		ables users to link Equipment on semantic MediaWiki to resources/datasets in ckan.
organization_group	ckanext-organization-group	No		The plugin adds an extra step for adding the dataset to a group in ckan and set the owner organization while creating a dataset.
dataset_reference	ckanext-Dataset-Reference	Requested https://github.com/ckan/extensions.ckan.org/issues/153	Yes Link	The plugin ables users to link reference(s) (like a publication or another dataset) to a dataset in CKAN.
cancel-dataset-creation	Ckanext-Cancel-Dataset-Creation	No		The plugin ables users to delete a draft dataset + cancel and delete a dataset during dataset creation process.
`data_comparision`	ckanext-data-comparision	No		The ckan plugin for comparing/visulaization data resources in ckan.
`custom_dataset_type` CRC 1153	ckanext-crc1153	No		Added custom dataset types to ckan dataset. Also, add a new facet section for filtering based on dataset type.
`sfb_search` CRC 1368	ckanext-sfb-search-extension	No		Search in datasets based on the column name in their data resource table (csv, xlsx) Search in datasets based on the linked samples

sample_link	ckanext-Semantic-Media-Wiki	No		Link the data resources in ckan to samples on SMW.
Extend search CRC 1153	ckanext-crc1153	No		Search in datasets based on the column name in their data resource table (csv, xlsx) Search in datasets based on the linked samples and publications Search based the CRC-specific metadata

Publications

- 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

- Sheveleva, Tatyana, Max Leo Wawer, Pooya Oladazimi, Oliver Koepler, Florian Nürnberger, Roland Lachmayer, Sören Auer, and Iryna Mozgova. "**Creation of a Knowledge Space by Semantically Linking Data Repository and Knowledge Management System-a Use Case from Production Engineering**." *IFAC-PapersOnLine* 55, no. 10 (2022): 2030-2035.

- Mozgova, I., O. Altun, T. Sheveleva, A. Castro, P. Oladazimi, O. Koepler, R. Lachmayer, and S. Auer. "**Knowledge Annotation within Research Data Management System for Oxygen-Free Production Technologies.**" *Proceedings of the Design Society* 2 (2022): 525-532.

- Altun, Osman, Tatyana Sheveleva, André Castro, Pooya Oladazimi, Oliver Koepler, Iryna Mozgova, Roland Lachmayer, and Sören Auer. "**Integration eines digitalen Maschinenparks in ein Forschungsdatenmanagementsystem**." In *DS 111: Proceedings of the 32nd Symposium Design for X (DFX2021)*, pp. 1-10. 2021.

- Wawer, Max Leo, Tatyana Sheveleva, Oliver Koepler, Florian Nürnberger, Iryna Mozgova, Roland Lachmayer, and Sören Auer. "**Parametrization of a Hybrid Component Production Process Chain Based on Semantically Annotated Data.**" In *Proceedings of the 10th International Conference on Mass Customization and Personalization-Community of Europe (MCP-CE 2022)*, pp. 200-207. Novi Sad: University of Novi Sad, 2022.

- Mozgova, Iryna, Oliver Koepler, Angelina Kraft, Roland Lachmayer, and Sören Auer. "**Research data management system for a large collaborative project.**" *DS 101: Proceedings of NordDesign 2020, Lyngby, Denmark, 12th-14th August 2020* (2020): 1-12.

- Sheveleva, Tatyana, Kevin Herrmann, Max Leo Wawer, Christoph Kahra, Florian Nürnberger, Oliver Koepler, Iryna Mozgova, Roland Lachmayer, and Sören Auer. "**Ontology-Based Documentation of Quality Assurance Measures Using the Example of a Visual Inspection.**" In *International Conference on System-Integrated Intelligence*, pp. 415-424. Cham: Springer International Publishing, 2022.

- Sheveleva, Tatyana, Oliver Koepler, Iryna Mozgova, Roland Lachmayer, and Sören Auer. "**Development of a domain-specific ontology to support research data management for the tailored forming technology.**" *Procedia Manufacturing* 52 (2020): 107-112.
- Schröder, Max, Susanne Staehlke, Paul Groth, J. Barbara Nebe, Sascha Spors, and Frank Krüger. "**Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation.**" *Journal of Biomedical Semantics* 13, no. 1 (2022): 1-22.