2. FDM-Werkstatt | 18.-20.03.2024 | IT Center, RWTH Aachen University

The 2. FDM-Werkstatt was organized by fdm.nrw in cooperation with the IT Center of RWTH Aachen University and DKZ.2R.

Program

Monday, 18 March 2024

- 3 PM: Arrival & Registration
- 3:45 PM: Official Welcome with Prof. Dr. rer. nat. Matthias Müller (IT Center RWTH Aachen University, DKZ.2R)
- 4 PM: Tour of CAVE & HPC
- 5 PM: Pizza party

Tuesday, 19 March 2024

- 8:30 AM: Arrival & Registration
- 9 AM – 12 PM:
    - Session 1: ReSeeD - Hands-On (*This workshop starts at 10 AM*)
    - Session 2: Research Data Management in large heterogeneous Collaborations with OpenBIS
    - Session 3: Build your personal/projectile Data Management System (and document it)
- 12 PM – 1 PM: Lunch
- 1 PM – 4 PM:
    - Session 1: Nutzung großer Datenbestände und Konsolidierung von fragmentierten, historischen Daten
    - Session 2: Automation in eLabFTW
    - Session 3: Python in the browser: Bring your tools to the web with Pyodide
- 4.15 PM – 18 PM:
    - Session 1: Meet Jarves - The Joint Assistant for Research in Versatile Engineering Sciences
    - Session 2: RDM and no idea where to start? From a Data Inventory to a Data Management Plan
    - Session 3: Code Sprint – Metadata for Pandas
- 7:30 PM | Self-Pay Dinner at Whitehouse Aachen

Wednesday, 20 March 2024

- 8:30 AM: Arrival & Registration
- 9 AM – 12 PM:
    - Session 1: Creating an Interactive Workshop Using LiaScript
    - Session 2: Connect Datalad to Coscine Vol. 2
- 12 PM – 1 PM: Lunch
- 1 PM – 4 PM:
    - Discussion 1: Coscine Technical Ask Me Anything
    - Discussion 2: DKZ.2R Diskussionsrunde

Abstracts

Build your personal/projectile Data Management System (and document it)

The session "Build your personal/projectile Data Management System (and document it)" enables you to actively work on the structural management of your very own research data on your

computer. In this workshop you will learn about different data documenting best practices, starting from the individual file and then widen the focus. You will get the chance to apply the defined structure of the *Johnny.Decimal-system* by re-organizing and remodeling these concepts to your (research) data. Additionally, we will explore the benefits of having directory level, machine readable metadata files, and a potential method for organizing and automatically curating these objects.

The content of these files forms the basis for the concept of the "self documenting data management plan" (*sddmp*) - a python project which produces a precise overview of your research data on a local webpage. The core principle of *sddmp* is that researchers can easily and adjust meta information at any time to their project in yaml files, quickly regenerating a detailed overview of their directory without tedious auditing and metadata transcribing. As *sddmp* is currently being developed, we will provide a brief demo, and open up discussion / working time so that participants can contribute with feedback and/or code to the project.

Target Group

- Everyone

Requirements

- Python installed on device, if possible

Person/s in charge

- Lukas C. Bossert (RWTH Aachen University)
- Jonathan Hartman (RWTH Aachen University)

Research Data Management in large Heterogeneous Collaborations with OpenBIS

The Collaborative Research Centre (CRC/SFB) 1394 is dedicated to a comprehensive exploration of defects, their thermodynamic stability in materials and their consequential influence on material properties.

The construction of defect phase diagram as a new materials design tool presents a significant challenge, requiring the integration of an extensive array of experimental data with corresponding outcomes from theoretical models and simulations. On the experimental front, over 10 research groups are devoted to specific facets, employing a diverse range of scientific instruments spanning from sample creation, thermal and mechanical treatments, mechanical testing, scanning and transmission electron microscopy, and atom probe tomography. The ensuing challenge lies in harmonizing data generated by these instruments, often in bespoke and proprietary formats, and capturing both the data and associated metadata efficiently. The RDM system that is being established within CRC 1394 is built on the laboratory information system and electronic lab notebook “openBIS”.

In this workshop, we will focus on implementing a metadata schema focused on experimental materials science in openBIS, devise a set of “standard operating procedures”, and capture experimental results both using the ELN web-interface, as well as extracting metadata from files automatically. Furthermore, we will discuss how to use hierarchical relationships to model experimental design and results.

Target Group

- Everyone

Requirements

- None

Person/s in charge

- Ulrich Kerzel (RWTH Aachen University)

ReSeeD - Hands-On

Repositories play a key role for the long-term access and publication of research data. However, there are cultural barriers for the use of research data repositories, especially for data which are not (yet) intended for public access. In certain scientific disciplines there are renowned repositories available, in other disciplines there is a lack of research data management infrastructure. For the latter, institutional infrastructures come into play. We report a case study of a research data repository (ReSeeD) operated by Ruhr University Bochum (Germany) and based on the Hyrax platform of the Samvera community. Adaptations to the open-source platform were made to lower the cultural barrier for data deposit and direct linking with relevant metadata: A fine-grained role and rights management enables tailored visibility of data. A multi-step and tiered review workflow are prerequisite for 10 years of preservation or publication with DOI. Easy authentication even beyond institutional affiliation is mediated via ORCID. The implementation was carried out in close collaboration between central IT, library and a research use cases. In a first stage, ReSeeD was brought into service late 2023 and is intended to serve as infrastructure for research data management for the about 6000 researchers of the university and their external project partners.

In this workshop, you will have the opportunity to try out the RUB RDM System and experience its basic functions firsthand:

- Create and label datasets
- Save data in datasets
- Share datasets
- Submit datasets for publication or archiving
- Searching for datasets
- Organizing datasets

In the second part, we look forward to getting in touch with the participants. We are open to discussing or deepening specific topics such as

- Requirements, dependencies and installation prerequisites of ReSeeD
- Exchange and development of connections to other tools and infrastructures

For more information about ReSeeD: https://datarepository.ruhr-uni-bochum.de/en

Target Group

- Everyone who is interested in the new research data management system of the Ruhr-Universität Bochum. In the practical part we will cover topics relevant for (potential) repository users, managers and administrators. In the optional part we will discuss technical or repository development aspects.

Requirements

- Laptop with Browser
- An ORCID-account

- A ReSeeD guest account

Person/s in charge

- Johannes Frenzel (Ruhr-Universität Bochum)

Automation in eLabFTW

Using the eLabFTW electronic lab notebook API, data can be synchronized or exchanged between eLabFTW and other applications. Various automation options can be implemented, such as creating, updating or querying laboratory notebook entries or database resources. One use case would be the automated collection of elements from a database, csv or xlsx files containing detailed information about chemicals, including their structure, identifiers, manufacturer and other properties. In the workshop, this will be tried out using reusable scripts.

Target Group

- Everyone

Requirements

- Basic knowledge of eLabFTW, Python and Jupyter Notebook
- Python (any os as long you are familiar with it), Access to JupyterNotebook (Python)
- Text editor installed
- Willingness, to install and use a test-version of eLabFTW (anonymous user with provided password, API-key)
- Willingness to watch video tutorials on elabFTW before the course, effort: approx. 1 hour

Person/s in charge

- Alexander Haller (Universität Heidelberg)
- Adienne Karsten (Universität Münster)
- Hüseyin Uzun (Universität Frankfurt)
- Mahadi Xion (Katholische Universität Eichstätt-Ingolstadt and Ruprecht-Karls-Universität Heidelberg)

Using large datasets and consolidation of fragmented, historical data

Many scientific experiments that are conducted over years acquire data from different versions of sensing and measuring devices. Similarly, historical data from past measurements have to be taken into account. All these data are fragmented, and might have different storage formats, data layouts, and data formats.

The challenge for researchers presents as: How can we bring all these data in a unique format, building a consolidated database that can be filtered or searched for specific information?

An additional question is: Which of these data can be stored and/or processed on a local machine and what are the factors that can make the move to a more centralized server based or cluster based infrastructure necessary?

In this workshop we explore together ways to tackle these challenges by analyzing real world examples from our daily work. We present two examples, one from environmental sensing for monitoring (not so smart) buildings, and one from a research project from technical textile

production. The former involves not only lots of data points but also multiple sensing devices and data paths, the latter generates high volumes of image data, that is not suited to be transported over wide area network connections.

We will provide jupyter-notebooks and datasets with which the workshop participants can try out different approaches, like data merging, re-formatting and filtering of large datasets. Additionally, we show our infrastructure for long term storage and fast access to large datasets.

Target Group

- Individuals and research groups that need to acquire, collect, consolidate and maintain digital data from multiple experiments

Requirements

- Tools used: Excel, Python
- Participants should have fundamental experiences in working with excel
- Participants should have fundamental experiences in programming (python preferred, but not required)

Person/s in charge

- Ingo Elsen (FH Aachen)
- Marcel Remmy (FH Aachen)
- Alicia Janz (FZ Jülich)

Python in the browser: Bring your tools to the web with Pyodide

The Python ecosystem offers many great tools to support scientific computing, data wrangling and data management tasks in general. However, using this elaborated toolbox usually requires the installation of a Python interpreter and project dependencies, which is often not feasible for non-expert users. On the other hand, interactive web applications based on Python code have become feasible with the help of the Pyodide runtime environment in all major web browsers. Such web apps are typically delivered as static content and do not require any backend infrastructure such as JupyterHub or Flask. The Pyodide runtime is in particular useful for legacy software projects that cannot afford transformation to JavaScript to run in web apps.

Topics:

Introduction to WebAssembly and Pyodide; Getting started with Pyodide; DOM manipulation in Python; Type translations between Python and JavaScript; Loading dependencies; Introduction to PyScript; Hackathon project

Ideas for the hackathon project:

We will create a web application that interacts with the Coscine RDM platform using the coscine-python-sdk package and automatically extracts metadata from uploaded files.

Do you want to suggest other Python-based tools to be used in the hackathon? Please get in contact with Frank Lange by 04.03.24.

Target Group

- Everyone

Requirements

- Participants should have basic skills in JavaScript and Python coding and in web page design.
- Coscine account and API token

Person/s in charge

- Frank Lange (RWTH Aachen University)

RDM and no idea where to start? From a Data Inventory to a Data Management Plan

In the Data Management Plan (DMP) workshop, participants will gain an overview of the objectives of a DMP. They will learn how to approach a research project while considering research data management and the available infrastructure at their institution before initiating a new project. This approach is practical and can be applied directly:

> Step 1 involves creating a Data Inventory of the existing IT infrastructure of the institution and/or the specific field of study.
>
> Step 2 consists of translating this Data Inventory into a DMP that also addresses the specific requirements of the respective project. We will introduce the DMP tool RDMO as an example since it is used at RWTH. However, the approach using the Data Inventory and the tool is agnostic and can be adapted to different systems.

Target Group

- Young researchers who just begun with their work and want to gain an efficient overview about the RDM infrastructure for their project
- people who got appointed to an „RDM manager“ in their research group, but have no background in it
- people who are just interested in a structured overview of their given RDM infrastructure

Requirements

- None

Person/s in charge

- Katharina M. E. Grünwald (RWTH Aachen University)

Meet Jarves - The Joint Assistant for Research in Versatile Engineering Sciences

Jarves (Joint Assistant for Research in Versatile Engineering Sciences) is a digital research data management (RDM) assistant. It is designed to help researchers to manage their research data effectively along their research processes. These guided processes are the foundation on which Jarves adds management of RDM policies and corresponding decision support. In this way, researchers are immediately supported based on their project and receive tailored information. In addition, researchers are also provided with training materials for their current research step and the tools used. Lastly, Jarves provides a partial automation to other tools, to reduce the effort needed for good RDM. Tools already connected are RDMO and Coscine.

In our session (2 hours) we would like to present the concept of the tool, give a demonstration of current functionalities, and collect feedback from the participants. Participants will be able access Jarves with their ORCiD, create projects and collaborate with each other in the tool.

In this way, we would like to enhance Jarves for its launch into a pilot phase later 2024. Desired results of the session are feedback on the workflows in Jarves, the functionalities included and planned as well as potential for further tools connected to Jarves.

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

Target Group

- Mostly engineers, however, all technical areas or even people beyond could benefit from Jarves

Requirements

- ORCiD Account
- As it is a web tool, no installation is required

Person/s in charge

- Tobias Hamann (NFDI4Ing), Jonas Werheid (NFDI4Ing)

Code Sprint - Metadata for Pandas

Although Python libraries like Polars are getting traction, I think it is reasonable to state that the Pandas library is the de facto standard for data manipulation and analysis in Python. Last month, PyPi recorded 150 million downloads of the Pandas library.

Due to this widespread use, the comprehensive documentation and the bundance of tutorials, it also is a common choice among researchers of all disciplines. A researcher's work is not finished when all data is processed and analyzed. As Open Science and FAIR data principles become prevailing, the processing and analysis is followed by various RDM tasks such as a thorough description of the produced data and its publication in research data repositories. For tabular data, as it is used and produced by Pandas, xlsx and csv are common publication formats. Although the spreadsheet format xlsx has several positives, such as the preservation of column data types, its proprietary condition is not ideal. The csv format otherwise, excels with openness and easy human-readability, but lacks seriously in formalization. Additionally, both formats do not offer standardized methods to incorporate metadata – including the vital information about the table's columns, such as descriptions, units, data formats (e.g. for datetime strings), etc.

The W3C recommendation CSV on the Web (CSVW1) overcomes these limitations by introducing a method to produce csv-accompanying json documents that contain this missing information. Although several packages that work with or support the creation of csvw-json files are existing, as far as I know, none of them tightly couples the production of a csvw-json to the csv-export of a Pandas data frame.

For the duration of the workshop, I envision a code sprint or mini-hackathon that explores the possibilities of such a coupling by developing a prototypical extension for Pandas. For example, a Pandas data frame already contains datatype information of its columns that vanish when exporting

a csv. The export of a preconfigured csvw-json can take away tedious work from the researchers while simultaneously supporting the integrity of the resulting csvw-json. I can imagine a class derived from the Pandas data frame class, that is enriched with capabilities to store metadata and that overrides the import and export methods for csv-files. During the workshop, we will discuss this idea and its potential, start coding and find a mode for future collaboration.

Target Group

- Everyone

Requirements

- None

Person/s in charge

- Arne Rümmler (Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden)

Creating an Interactive Workshop Using LiaScript

LiaScript is a simple and extendable Markdown dialect that allows anyone to create, collaborate, and share professional looking and interactive online courses, textbooks, and workshops. LiaScript aligns well with some of the key FAIR principles, offering a plain text, collaborative, and interoperable platform for creating and sharing content related to research data management. Together we will delve into the intricacies of LiaScript, exploring its functionalities, unique features, and collaborative possibilities.

The workshop is split into two sections. In the first hour, we will cover how to create and deploy LiaScript content, as well as what makes LiaScript different from standard Markdown. In addition, we will cover a few tricks for going beyond the basic functionality of the language to create interesting and engaging content.

In the remaining time, we encourage participants to bring their own workshop content or ideas, and participants can work with a facilitator support to translate these to a LiaScript format. We will provide guidance, answer any questions, and assist with implementing or exploring any ideas which might arise that fall outside the scope of the initial presentation.

To conclude, we will spend a few minutes reflecting on how participants experienced the process of using LiaScript and how they might use it in the future.

Target Group

- Everyone

Requirements

- An account on GitHub/GitLab/Bitbucket or any other code hosting service is preferable, but not required.

Person/s in charge

- Jonathan Hartman (RWTH Aachen University)

Connect Datalad to Coscine Vol. 2

Datalad is Git for Data and tracks metadata while Coscine provides data storage as well as metadata annotation. Connecting the two data management tools could prove powerful to users and fits Datalad's portfolio well, as connections to various storage solution have been implemented. On Coscine's end this could assist in working with the stored data while tracking changes to the metadata in an automated fashion.

This session will continue on the session during the previous FDM Werkstatt in 2023. In the previous session, collaboration began between members of the Datalad developers and the RWTH IT Center. A general outline for how such a connection could look was established.

We would like to take the opportunity in 2024 to continue this work. With a more powerful API, it is now easier to interact with Coscine. Furthermore, Coscine developers will be present to actively assist on this feature ad provide technical input.

Target Group

- Datalad team, Coscine Team, researchers or data stewards in projects working with or looking to work with both systems.

Requirements

- Knowledge of at least one of the systems (Datalad, Coscine)
- Python programming is a plus

Person/s in charge

- Nikki Parks (RWTH Aachen University) and *tba*

Coscine Technical Ask Me Anything

This session, hosted by some of Coscine's developers, provides technical answers for data stewards, users, admins, and anyone else. It provides the opportunity, first and foremost, for the user base to become more acquainted with the system on a technical level, but also for the developers to engage with the user base, gain insight on how it is being used, and gather inspiration for improvement.

This also provides the opportunity for quick demos, be it new features provided by APIv2, the AIMS application profile generator, or general usage. For questions that may require some further preparation on the developers' end (such as specific demos), please email the event organizers beforehand so we can prepare accordingly.

Target Group

- Anyone with (technical) questions about Coscine

Requirements

- None

Person/s in charge

- Nikki Parks (RWTH Aachen University)