K. Rial¹, H. Görzig¹, R. Krahl¹, A. Velazquez Sanchez¹

¹⁾ Helmholtz-Zentrum Berlin für Materialien und Energie, Hahn-Meitner-Platz 1, 14109, Berlin

HZB Helmholtz Zentrum Berlin

ROCK-IT

DATA & METADATA IN THE HELMHOLTZ ROCK-IT PROJECT

ABSTRACT: For a ROCK-IT thermal catalysis experiment, the (meta)data lifecycle in the facility begins with the initial proposal and extends through the entire investigation, culminating in the archiving and cataloguing of the (meta)data. At all stages, coherence between data and metadata is vital. At HZB we are using a combination of established systems and new initiatives to achieve a comprehensive framework aligned with FAIR principles.

Introduction

ROCK-IT is a collaborative project between four Helmholtz Institute centres (DESY, HZB, HZDR, KIT) that aims to standardise and automate the process of performing thermal catalysis experiments to allow remote operation regulated by AI/machine learning. The project is split into work packages (WPs). WP5 is responsible for data management.

Metadata requirements

Metadata is vital to give data context. For ROCK-IT, the necessary metadata can be categorised as follows: <u>Anything needed...</u> <u>Groups to specify needs</u> ... to run the experiment WP2, WP3, WP4 ... for analysis User, Beamline Scientist ... for searchability User, Community ... for accountability Registration/User Portal ... for reproducibility User, WP2, WP3, WP4

Common formats



HZB Helmholtz Zentrum Be

been agreed that data will be stored in NeXus files, which uses the HDF5 format. This allows the structure



The four centres already collect data and metadata from experiments, using their own unique setups. ROCK-IT is not proposing to replace these, rather the challenge is to continue using the existing systems whilst standardising the metadata requirements and the data formats used. In doing so, it is important to also conform to FAIR data principles.

These various requirements necessitate the creation and/or introduction of new elements to cover specific areas that are not currently robust. This is largely centre-specific, as it complements the existing infrastructure, yet must align to a universal framework across the project.

This highlights the need for involvement from users and the wider catalysis community in identifying the metadata that should be collected.

This is complicated by the abundance of terms in use. For ROCK-IT, we are building a cross-reference of terms used by resources such as NFDI4(at DAPHNE4NFDI, DataCite and the DAPHNE NFDI4Cat Voc4Cat ontology, as well as those used internally by bluesky, 4NFDI NeXus NXxas and the control systems. **DataCite**

Implementing new systems

Where necessary, ROCK-IT is implementing new tools. In some cases, this is to standardise existing data or procedures. Introducing bluesky as a top-level instrument control software is an example of this; see poster Automatic Export of Data from Catalysis *Experiments to NeXus in ROCK-IT* for more information. (See also **Common Formats** above.)

of the data output to be standardised across the centres, regardless of which formats the various instruments favour. To achieve this, at HZB we have developed a conversion tool called NeXusCreator. See poster NeXusCreator & ICAT - HZB applying FAIR data management for more information.

Similarly, it is helpful if incoming data is standardised. LinkML provides a way to convert data from a wide variety of ELNs (electronic lab notebooks) into a common format that can be used by the ROCK-IT centres. In particular, this will supply sample data, which at HZB will then be imported into SEPIA (see **Implementing new systems** below). At HZB we are testing this with data from NOMAD, which is being widely adopted as a data manager and link ELN in our labs. **L'NOMAD**

this information is vital. To this end we are developing SEPIA, a new database to store sample data that will be linked directly to the corresponding (meta)data in ICAT. This fulfils FAIR data ideals and gives a more complete record of our experiments. SEPIA will also be able to mint a global PID (persistent identifier) through IGSN for each record, so that samples can be tracked through multiple investigations. **IGSN** For ROCK-IT, multiple samples will be used during a single experiment. SEPIA will provide full information about each sample and ensure that each is linked to the correct datasets. Looking forward, it may be possible to adapt SEPIA for the other centres as well.



Here we outline how data management is being handled across ROCK-IT, and consider the specific setup being developed and implemented at Helmholtz-Zentrum Berlin (HZB).

Working with existing systems

The four ROCK-IT centres have all been operational for a number of years, and therefore have a variety of systems in place. ROCK-IT uses the existing infrastructure wherever possible. Many systems cannot be changed; for example the archiving and cataloguing of data is achieved using SciCat at some centres and ICAT at



In other cases, new tools fill a hole in the established systems. At HZB we have not historically stored details



about samples being tested. In order for data to be useful beyond the scope of the original experiments,

Flow of data **ROCK-IT** within a ROCK-IT experiment Operando Proposal Publication system processing Measurement Synthesis

others, and this will not change in the foreseeable future.

To bring all these systems together, we first determined the flow of data through the experiments. We then identified the systems in place at each centre and investigated how similar or different they are. We agreed on common formats that all four centres could use within their existing setup.

SUMMARY

- ROCK-IT requires both global and centre-specific solutions, in order to provide a uniform experience despite each centre using a different underlying framework.
- Existing systems can be aligned to work similarly, enabling a uniform approach at higher levels.
- New tools are required to complement the existing systems. These can either help the established systems to conform to global requirements (e.g. NeXusCreator) or supply additional functionality in order to achieve FAIR data (e.g. SEPIA).
- HZB has developed a number of solutions for its

We also identified gaps in the current setups, either universally or for individual centres. Some processes could not be automated, some information was not stored by the centre (which prevented the data being 'FAIR at the point it leaves the facility'). In particular, top-level systems were introduced that could interface with the established setup at each centre and then provide a uniform experience across the project. These have allowed us to standardise the data collection for ROCK-IT **[**Cat without replacing functional systems.

systems. These are intended to be adaptable, should the other centres have a need for them.

• Metadata requirements can only be established with input from a wide array of stakeholders. Crucially, we are working with catalysis scientists to identify metadata needs

• Metadata terms are not currently unified, although community projects such as Voc4Cat and NXxas are making progress in this area. In the meantime, we need to identify equivalent terms relevant to ROCK-IT.

DAPHNE4NFDI TA1 Face2Face Meeting and Workshop 2024 Kiel University 9-11 December 2024

MORE INFORMATION

Katherine Rial

katherine.rial@helmholtz-berlin.de +49 (30) 8062 - 12540 ORCiD: 0009-0007-7266-485X

Heike Görzig

heike.goerzig@helmholtz-berlin.de +49 (30) 8062 - 42683 ORCiD: 0000-0001-9121-8643