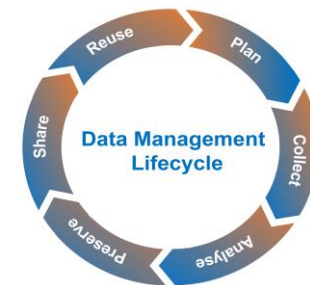


Research Data Management at GSI/FAIR

Andrew Mistry

GSI/FAIR Open Science Workshop 2023

19.10.23



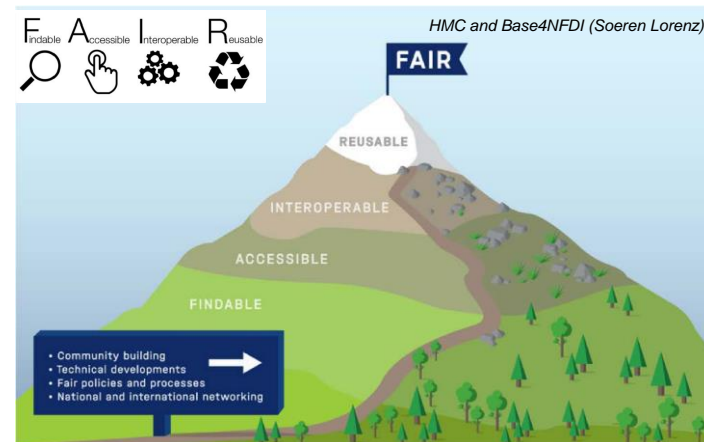
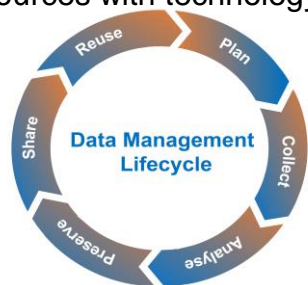
Research Data Management encompasses all aspects of handling research data, from planning, its generation and processing to publication, long- term archiving, and eventual deletion, while adhering to the principles of good scientific practice.

One of the crucial philosophies of RDM are the F.A.I.R principles

We follow: “as open as possible, as closed as necessary”

FAIR Data is not an end goal

-> continual process of improving practices and adapting research resources with technology innovations



Good scientific data management aids in knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.

Going from good research data management to data publication is not a big jump!

Slides and information available on indico <https://indico.gsi.de/event/14680/>

=> **70 participants** in hybrid meeting, variety of themes discussed:

- **Inform and make researchers aware of RDM and efforts in this direction**
- **Open science projects:** GSI/FAIR involvement
- Talks from **individual research groups, Grant office, Helmholtz Open Science Office**
- Forum of **open discussion** and ideas sharing
- **RDM FAQ list** drawn up and points addressed
- **Future plans** and updates for RDM



11. Has your department used an local (internal) metadata scheme or a standardized (external) one for any of your research projects? (e.g. see <https://www.dcc.ac.uk/guidance/standards/metadata>).

If a standard external schema please specify which.

10 responses

No

No standard scheme is used in the heavy-ion community, we are just adhering to some output format conventions (OSCAR and HepMC)

no

not even a concept :-(

We use local files/ directory naming + PI name for long term storage data structure

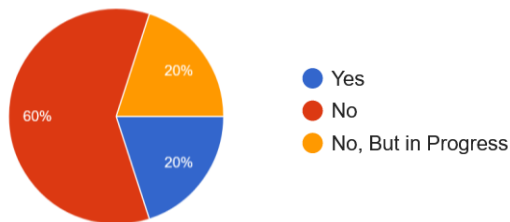
None

we follow policies by HADES and PANDA collaborations within their computing models; no local scheme presently foreseen.

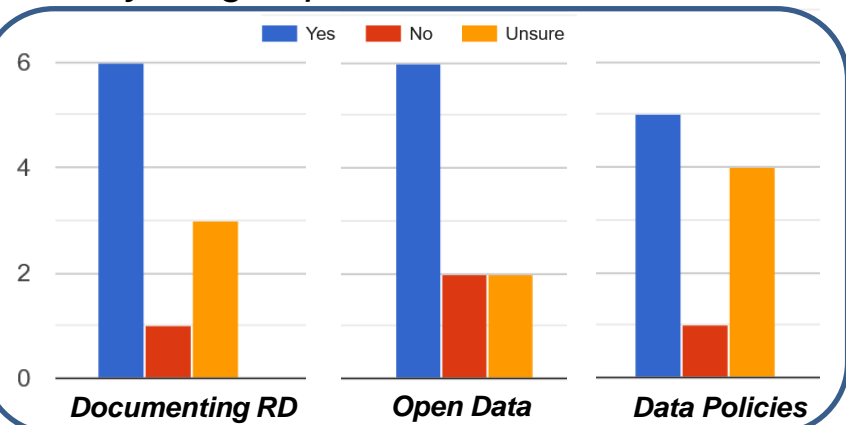
?

10 responses from various groups

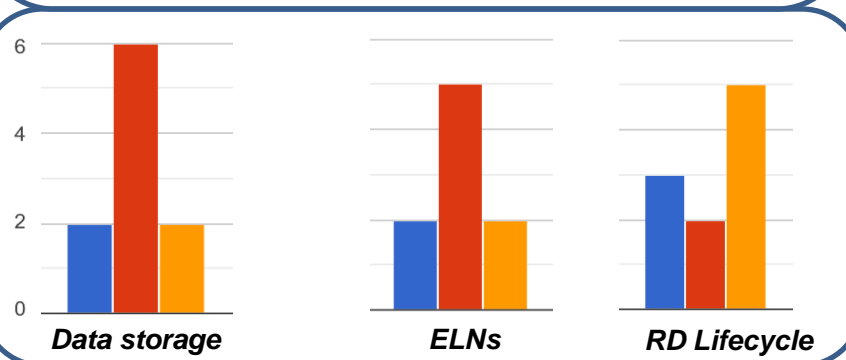
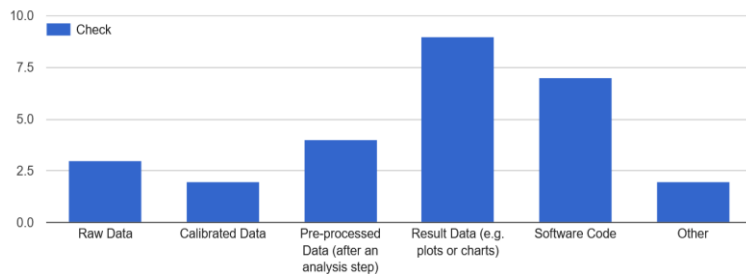
Prepared a data management plan?



Does your group need more information on:



What RD would your group consider making open?



- How should researchers be rewarded for publishing data?
- How can the time constraints on researchers (e.g. PhD students) be effectively managed?
- Where can I find information on RDM at GSI?
- What is the best way to inform and teach researchers and collaboration partners on RDM and steps to take?

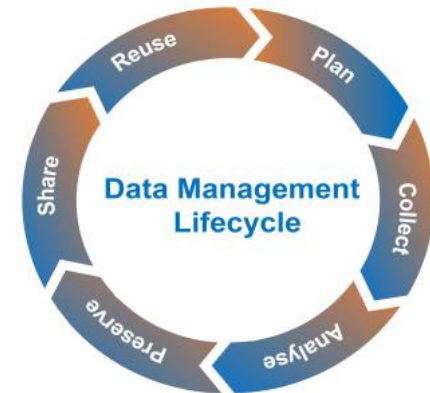
***Now: work on discussions and integration of RDM practices (F.A.I.R. –principles)
with researchers***

Goals:

- To ensure **good RDM (SDM) practices** at GSI/FAIR and emphasize the benefits;
- Promote practices and assist researchers in **publishing data/software**;
- To aim (as best as reasonably possible) that data/software is **published** according to the Findable Accessible Interoperable and Reusable (**F.A.I.R.**) **principles**;
- Develop the **tools and infrastructure** needed to do this.

Development of strategy with Open science working group that includes:

- Policy and guidelines **publication**;
- **Tools**: testing and implementation;
- Collaboration with **external partners** on RDM;
- **Communication** and outreach through workshops, events and teaching;
- Practice and evaluate **'interim' strategies for RDM**, data publication;
- Create and develop **use cases** within each research group



- Published May 2023: Applies to all research data generated at GSI/FAIR
https://repository.gsi.de/record/339448/files/C-VA-RED-en-Research_Data_Management_Policy.pdf
- Introduce RDM and define policy points that should be adhered to for data generated at GSI/FAIR
- Developed in collaboration with Open Science WG
- Aligns with Ethics and Good Scientific Practice Policy

	Document type: Procedure	Date: 10.05.2023
		Page 1 of 5

Title:	Research Data Management (RDM) Policy
Responsible unit	RED
Scope:	GSI & FAIR
Release	This document was endorsed by the GSI/FAIR management on 23.04.2023 * - See glossary

2. Policy Points

- GSI/FAIR strongly advises that a Data Management Plan (DMP)* is prepared at the start of each research project to describe the procedures for the collection, processing, storage and long-term archiving of research data. This aids in determining responsibilities, access, and facilitating the reuse and reproducibility of the research data. The research data manager can be consulted when developing a DMP.
- The principal investigator of the research project holds primary responsibility for the research data, and they should be listed in the DMP. In essence, the principal investigator is responsible for the research data throughout the management lifecycle, compliant within the subject-specific standards. To assist in this task, the principal investigator may delegate some responsibilities to other collaborating members of the project team.
- GSI/FAIR will advise researchers, and provide necessary documentation on the planning and implementation of research data management. This also includes support in the access and use of suitable repositories*, data formats, access to software, and tools for processing. In addition, GSI/FAIR will provide necessary storage solutions for research data, as well as required infrastructure and regulated access. It must be specified in the Data Management Plan (DMP) where the research data will be stored, how it will be backed up, and how it can be accessed. Research data must be stored and safeguarded for a minimum period of 10 years. Longer or shorter retention periods prevail in accordance to legal regulations, funders' and other contractual requirements.
- Whenever possible, research data should be made openly available by the responsible scientists. Suitable repositories should be used for research data accessibility, and the tools and formats used for data collection and analysis must be well-documented using comprehensive metadata*. In addition, the research data generation and any analysis processing steps must be carefully documented (e.g. in electronic logbooks/notebooks). When making research data available, it should be licensed such that it remains accessible, i.e. by choosing a permissive license such as Creative Commons Attribution 4.0 International (CC BY 4.0) [3] if possible, GSI/FAIR retains all rights for access to the research data.
- The research data may be subject to an initial embargo period, with access restricted to the collaboration or persons defined by the principal investigator. Data protection laws, patent laws, economic and contractual framework must be adhered to. Open access of research data (i.e. under a creative commons license) with regards to Technology Transfer, and/or dual usage must be taken into consideration.



Prepare Data management plans



PI is responsible for research data (or someone they assign)



GSI will provide support and infrastructure for RDM. Data should be held for at least 10 years



Whenever possible data should be published in suitable repositories CC BY 4.0 is a suitable licence.



Data may be placed temporarily under embargo. Data protection laws must be adhered to.



RDM Guidelines: In preparation

GSI/FAIR Guidelines on Research Data Management v.4.1
June 2023

GSI/FAIR Guidelines on Research Data Management v.4.1 01/06/2023

Table of Contents

1. Preamble	3
2. Responsibilities	4
2.1 Researchers	4
2.2 Principal Investigators	4
2.3 GSI/FAIR	5
3. Research Data Planning	6
4. Managing Research Data	6
4.1 Documenting research data and Metadata	6
4.2 Data Storage	7
4.3 Publishing Research Data	7
5. Examples of Data Publication	8
5.1 <i>Example 1 Materials Science</i>	8
5.2 <i>Example 2 Large dataset</i>	9
5.3 <i>Example 3 PHELIX</i>	9
5.4 <i>Example 4 ESR</i>	9
6. Data access and licensing	10
7. Jurisdiction	11
8. Glossary	11
Bibliography	13



Broader guide on RDM for researchers

Defines roles and responsibilities in more detail

Explains documentation, publication and planning

Gives explicit examples from researchers

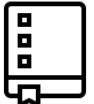
Includes legal and transfer issues



Large volume data storage + archiving (Lustre/LTSM/FSQ) 



Electronic Logbooks (ELog) 



Internal publication repository (JOIN2) 



Code management system (GitLab) 

Coming soon...



Data Management Planning Software (RDMO)



New GSI repository



Metadata schema



Instrument PIDs (all instruments at GSI to have a PID)



Ecosystems e.g. Heliport, PUNCH4NFDI, Eurolabs...

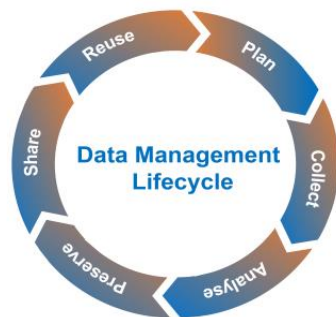
- Task groups in Helmholtz Open Science dedicated to this
- POF IV funding period -> count published research data/software items!
- Spring 2024: entry-level indicator and quality indicator investigative phase
- Indicator should be part of PoF indicator starting 2025 (reporting year 2024)



- A **Data Management Plan** is a research project document that aids in the process of **ensuring that research data is handled correctly**. Describes the Data lifecycle of the project
- **Living document** -> Should be filled out at the start of the project and continuously updated throughout
- Reluctance to filling out Data Management Plans! Seen as just more paperwork...
- **BUT**: Useful for researchers (present and future), needed to enable F.A.I.R data, many funding agencies now require a DMP to be prepared at the start of the project etc.

Data management plans aid:

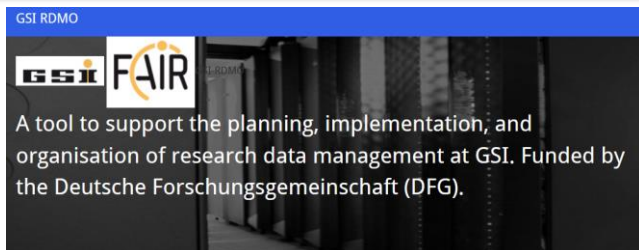
- Communication tool for researchers
- RDM project internal management
- Future reuse and project planning
- Useful for IT/Resource Cost estimates
- Funding body requirements



Contents can include:

- General info (Project name/PI)
- Data info (Data type, size, generation method)
- Overhead (Data protection, personnel costs...)

Goal: Make it easy and encourage to prepare these



Welcome to GSI Research Data Management Tool

The aim of this website is to provide a tool to organise data management plans in an easy way. This is a prototype and is based on the software RDMO. The aim of the RDMO project is to deliver a web application to assist structured planning, implementation and administration of the data in a scientific project. Additionally, the gathered information can be cast into textual forms suitable for funding agencies requirements or for reports.

- Data Management with easy to use tool: ->**RDMO**
- Currently running test version at GSI: Catalogue setup and feature testing -> **if you are interested in testing please contact me!**
- Plan to go live in Autumn/Winter 2023
- *Plan to employ in combination with GATE proposal submission system 2024+*

My Projects / GSI/FAIR Data Management Plan / Data Set Description

Data set Collection

Please fill in the form for each dataset. The different datasets will be referred to in following questions. You can add a new dataset using the green button. Once created, you can edit or delete datasets using the buttons in the top right corner.

5452 [Add dataset](#)

What types of data will be generated?

e.g. Experimental, Simulated, calibrated, transformation/analysis of other data...

What format(s) will the data have?

lmd, root, ascii, png ...

Please enter the items line by line. You can add items using the green button and remove them using the blue cross (x).

[Add item](#)

How large is the raw data set anticipated to be?

Please give the expected value in GB (e.g. for 7TB type 7000)

How large is the processed data set size anticipated to be?

The data set size after processing (not including the raw data). Please give the expected value in GB (e.g. for 7TB type 7000)

Overview

Project: GSI/FAIR Data Management Plan

Catalog: GSI/FAIR Data Management Plan

[Back to my projects](#)

Progress

[Back](#)

[Skip](#)

Navigation

Please note that using the navigation will discard any unsaved input.

Entries with @ might be skipped based on your input.

General

[Data Set Description](#)

[→ Data set Collection](#)

[Data Publication and Access](#)

[Data Findability and Metadata](#)

[Data Interoperability](#)

[Data Resuability](#)

[Ethics and Legal issues](#)

[Additional Notes and Information](#)

[Associated Costs](#)

Metadata essential in Research Data Management and Open Science to enable FAIR data (and code)

More info: [Guide to HMC Better Metadata Booklet](#)

From NeXuS (neutron, X-ray and muon experiments description):

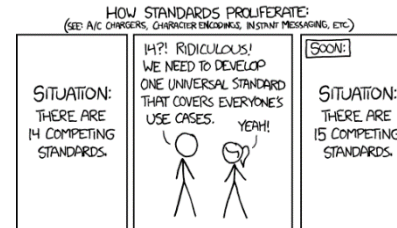
“As instrumentation becomes more complex and data visualization becomes more challenging, individual scientists, or even institutions, find it difficult to keep up with new developments. A common data format makes it easier, both to exchange experimental results and to exchange ideas”

When publishing data, also publish machine readable metadata

- Allows datasets to be searched for and found
- Enables interoperability between datasets
- Enables reprocessing of data: transparency and integrity
- Efficient use of resources

However: No common schema existing between accelerator/nuclear physics experiments ->This is needed

Caution: Don't reinvent the wheel!



By Randall Munroe: <https://xkcd.com/927>

Metadata for nuclear physics experiments

-> See talk by I. Knezevic

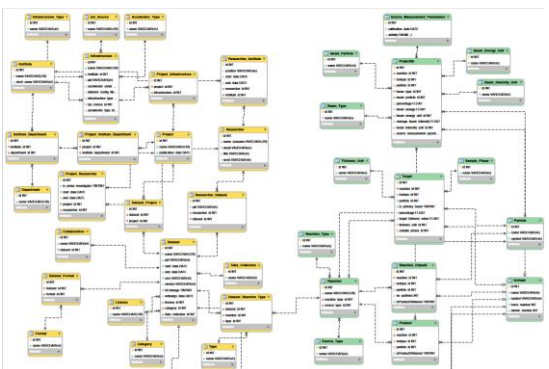
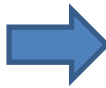


- **No common schema existing between accelerator/nuclear physics experiments**
- **European (and beyond) wide strategy** -> commonalities, leverage open science projects
- Start simple and built up: *Spreadsheet-> ER Diagram-> Database development->Front end web*
- End user should be able to **input project metadata to simple online form and then export (machine readable) file**





Suggestions, recommendations and collaborations welcome; Please contact me

Cardinality	Metadata item	Notes	Field type
General Information			
1	Project Name		String
1	Resource Type	Dataset	String
1	Publication Date	Date of Publication	ISO Date/Time
0/1	Project ID	Official experiment number obtained	String
1+	Principal Investigator	Responsible person. Can include full name	String
	Principal Investigator Email		
	Principal Investigator PID		
Facility/Institute Information			
1+	Facility/Institute of Data Generation	Of data generation	String
1+	Facility/Institute ROR		String
1+	Infrastructure	Overlaps with detector/apparatus? - in	String
1+	Department/Division	The department/division associated with	String
0+	GSI/FAIR pillar	("APPA", "CBM", "NUSTAR"...) can ap	String
0+	POF:	HGF. Nested	String
0+	POFIV Topic		String
0+	POF ID		int
0+	POF Period		int

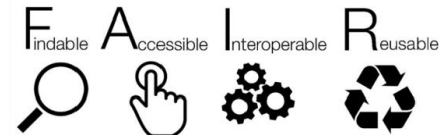


A. Mistry, I. Knezevic, C. Hornung, A. Matta, A. Lemasson, G. Günther, J. Isaak, DESY + HZDR+ PUNCH4NFDI

Compared to data management -> maintenance and versioning required, licensing considerations and applications

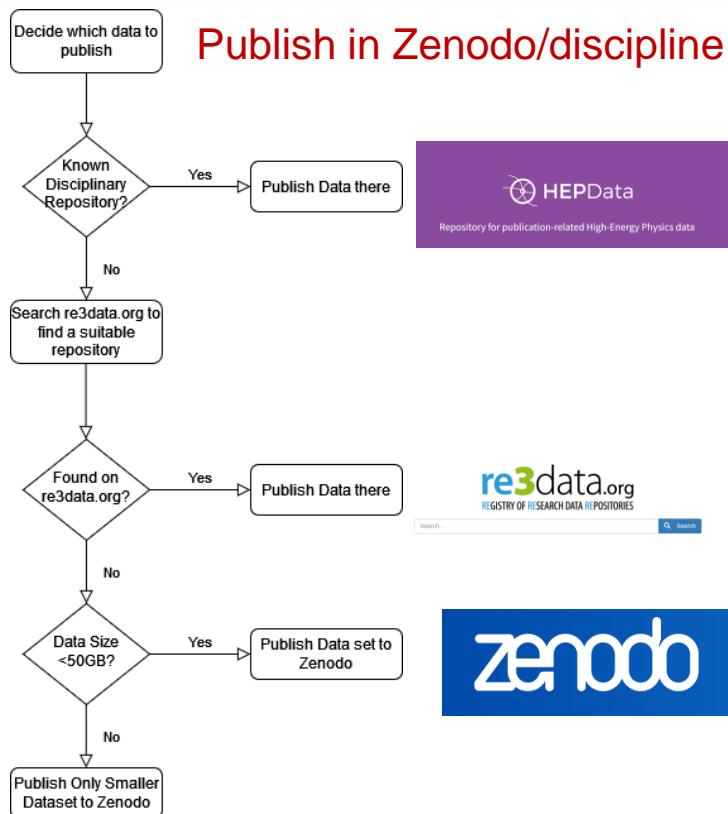
- ✓ **Publication instructions** available here: <https://doi.org/10.5281/zenodo.7628019>
- ✓ **Software licensing guidelines** available here: https://www.gsi.de/fileadmin/Forschung/C-VA-RED-en-Open_source_software_license_at_GSI_FAIR.pdf
- ✓ **Code management system** (GSI Gitlab) available 
- ✓ GSI Supporter of “**Public Money Public Code**” Campaign 
- ✓ **Software can be onboarded** to repositories: curation and promotion. e.g. ESCAPE OSSR <https://projectescape.eu/ossr> -> *See talk by C. Tacke* + Helmholtz Research Software repo
- **Research software guidelines** in preparation

Barker, M., Chue Hong, N.P., Katz, D.S. *et al.* **Introducing the FAIR Principles for research software.** *Sci Data* **9**, 622 (2022). <https://doi.org/10.1038/s41597-022-01710-x>



How To: Data publication (interim solution)

Publish in Zenodo/discipline specific repos



The screenshot shows a Zenodo publication page for the document 'Instructions for uploading and linking research data/software at GSI' by Andrew Kishor Mistry, dated May 2, 2023. The page includes a title, author, and a detailed description of the document's content. It also features a 'Preview' section showing the document's title and version information (v. 3.0, May 2023). The page is indexed in OpenAIRE and has 185 views and 193 downloads. The DOI is 10.5281/zenodo.7628019. The page also lists the publication date, DOI, and keywords, along with the license (Creative Commons Attribution 4.0 International) and a list of versions (v3.0, v2.4, v2.3).

<https://doi.org/10.5281/zenodo.7628019>

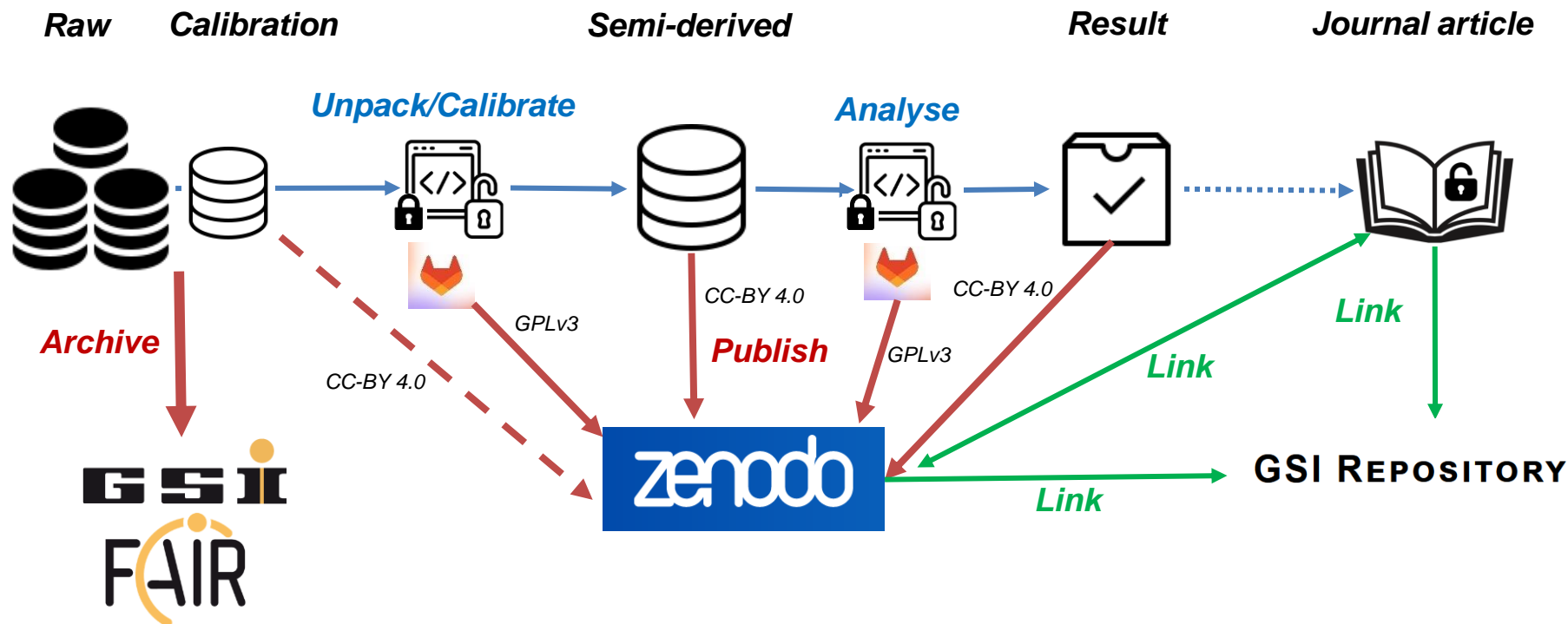
Instructions available on how to publish data at GSI

- Plans for **GSI data repository** underway -> **talk by V. Schulte**
- Longer term goals of projects like ESCAPE, PUNCH4NFDI and collab within Helmholtz for more complex access to datasets and workflows -> reproduce results
- In the meantime using **Zenodo/discipline specific repos for data**
- Interim solutions e.g.: **Semi-derived (pre-processed) data**. Publish also **“result” data** (e.g. from plots) -> Up to the PI/responsible what to publish
- Record then generated in repository (or portal) with a **PID and link to the data**

Notes -> defined in RDM policy/guidelines:

- Aim to **publish data at end of project** i.e. after article publication(s);
- Data may also be placed under **embargo** or only available for specific users...
- **Software may be made open access** during the course of the project, decision by the researchers + dual use/tech transfer should be considered
- Software and Data should have an **appropriate licence**

Publish: semi-derived data, software codes, result data and accompanying metadata after project completion. (Calibration data if useful for result replication!)



How To: Interim Data Publication strategy

Dataset

zenodo Search Upload Communities a.k.mistry@gsi.de

November 2, 2022 Dataset Closed Access DOI New version

GSI Test Dataset

Andrew Kishor Mistry

Here, the dataset should be described in as much detail as possible. Metadata and other data structure should be given. If needed, a separate document describing the dataset in advanced detail can be uploaded.

Software

Search or jump to

amist88 / GSI_TestSoftware_RDM

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

Go to file Add file + Code + About

amist88 Update README.md 20 days ago

README.md

GSI_TestSoftware_RDM

zenodo Search Upload Communities

November 3, 2022 Software Open Access

amist88/GSI_TestSoftware_RDM: GSI Test Code Release 1

DOI

DOI

GSI Publications Repository

Journal Article 931-2022-0011

GSI Test Journal Article for Research Data Management

Mistry, A. K. (Corresponding author)

2022

Nature Pub Group London [u.a.]

Nature - London, 605, 1 (2022)

Abstract This is a test record to describe how to link research data from an external repository to the GSI publications repository

Classification:

- 08-500
- Contributing Institute(s)
- 1 Bioethik & Dokumentation (BUD)
- Research Program(s)
- 1 G12_Cosmic Matter in the Laboratory (POF4-612) (POF4-612)
- Experiment(s)
- 1 (Abstract therefore no facility)

Database coverage:

PubMed® - BIOSIS Previews - Biological Abstracts - Chemical Reactions - Cleanvate Analytics Master Journal List - Current Contents - Agriculture, Biology and Environmental Sciences - Current Contents - Life Sciences - Current Contents - Physical, Chemical and Earth Sciences - Elsevier Academic Search - Essential Science Indicators - IF >= 40 - Index Chemicus - JCR - National Library of Medicine - SCOPUS - Science Citation Index Expanded - Web of Science Core Collection - Zoological Record

The record appears in these collections:

- Private Institute collections > »WGF > »RED > BUD
- Document types > Articles > Journal Article
- Institutes > Library & Documentation
- Workflow collections > Public records
- Publications database

Linked articles:

- Software
- Mistry, A. K. (Corresponding author)
- amist88/GSI_TestSoftware_RDM: GSI Test Code Release 1
- (10.5281/ZENODO.7277784)
- Dataset
- Mistry, A. K. (Corresponding author)
- GSI Test Dataset for Research Data Management
- (10.5281/ZENODO.7274415)

Record created 2022-11-02, last modified 2022-11-03

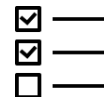
Article -> Dataset -> Software linking

Datasets/software records in GSI publications repository count towards HGF POF IV

RDM: How to easy example (can be expanded)

1. Start of project/shortly before data taking

1. *Prepare Data Management Plan* for the project -> Templates available; tool available soon



2. During Data Generation

1. Ensure *data is described correctly* during data taking ->Electronic Logbooks and encouragement!
2. Ensure *data is archived* -> Speak to IT if no solution currently in place



3. After Data Generation

1. *Return to data management plan* and update with any changes
2. Ensure that *metadata of the project is complete* and up to date -> RDM Team can help



4. During Analysis

1. *Organise processed data* in a way you would treat raw data: labelled and described digitally
2. Ensure that *analysis codes are placed in a code repository* and maintained



5. After analysis

1. *Finalise which data should be published* in order to reproduce the results
2. *Publish data with a PID in a suitable repository, plus associated software+licences*, at the same time as (the last) publication
3. *Link everything* to the GSI publications repository



- **Contact with specific needs; e.g. metadata, assistance with Data management plans**
- **We are looking for data/software publication use cases (at least 1 per group). Please get in touch!**
- **We are also looking for more complex cases involving interoperability of data sets. Any idea on this please also get in touch -> FAIR connection to EOSC**
- **Contact for recommendations on idea, tools and developments**

Contact: open-science@gsi.de
Website Open Science @ GSI/FAIR: <https://www.gsi.de/open-science>

- Research data management is the term to describe the handling of research data
- Practicing good RDM is a great asset
- Developments ongoing at GSI/FAIR to enable best RDM practices
- Data management planning -> tool in preparation
- Metadata and project description -> Preparing schema for research areas
- Data publishing: Interim solution, new repository in development
- Dissemination of information: RDM team can assist!

We are looking for data/software publication use cases (1 per group). Please get in touch!

Contact: open-science@gsi.de

Website Open Science @ GSI/FAIR: <https://www.gsi.de/open-science>

Thanks for your attention!