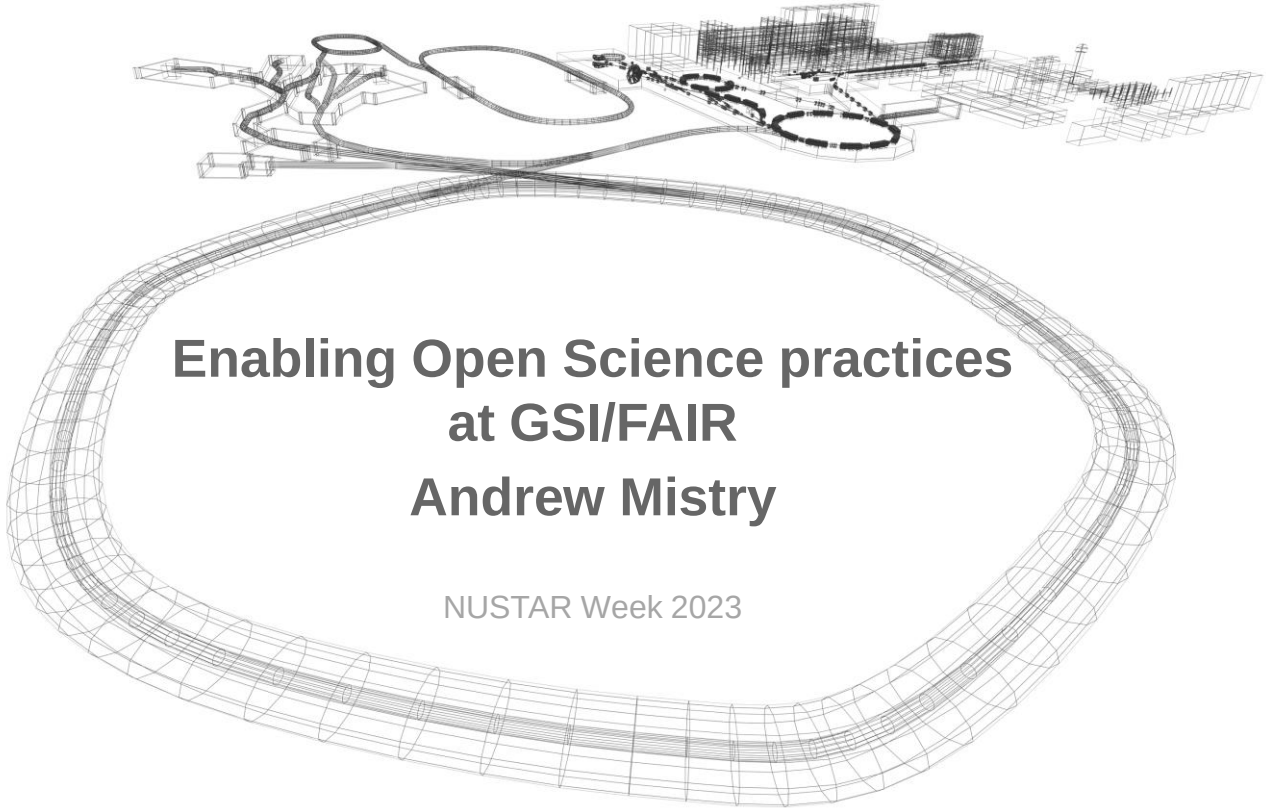
HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

FAIR GSI



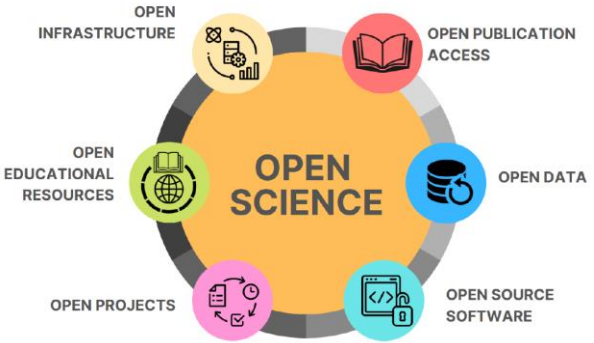
Enabling Open Science practices at GSI/FAIR

Andrew Mistry

NUSTAR Week 2023

What is Open Science?

- Science lives on the **open** exchange of knowledge
- ***Openness*** -> offers new prospects in the entire scientific research cycle and ***enables research outputs to be made openly accessible*** *and* ***broadly reusable*** (in sustainable infrastructures).
- This ***open culture*** of scientific endeavor is captured by the term "***Open Science***," Defined by the concepts of: ***transparency, sustainability, transfer, collaboration and sharing***
- *In essence:* Make **research outputs** (+ infrastructure) **publicly** available to science, industry, and society for **reuse** with ***as few barriers as possible***
- How to define and shape Open Science practices, tools and dissemination in a way that maximises the rewards and benefits? Important to consider what is ***useful for researchers!***
- **Open Science requires a shift in research culture**: it takes additional work and resources to practice open science: **Support needed from leadership**



Why is Open Science important?

Accelerates **knowledge transfer** by breaking down access barriers to research outputs.

Fosters **collaboration** within and across disciplines, leading to quality improvements and new solutions.

Open Science promotes **transparency**, building **trust** among researchers and the public.

Attracting **future researchers**: Open Science signals inclusivity and appeals to diverse talent.

Sustainability improves as resource sharing reduces repetitiveness

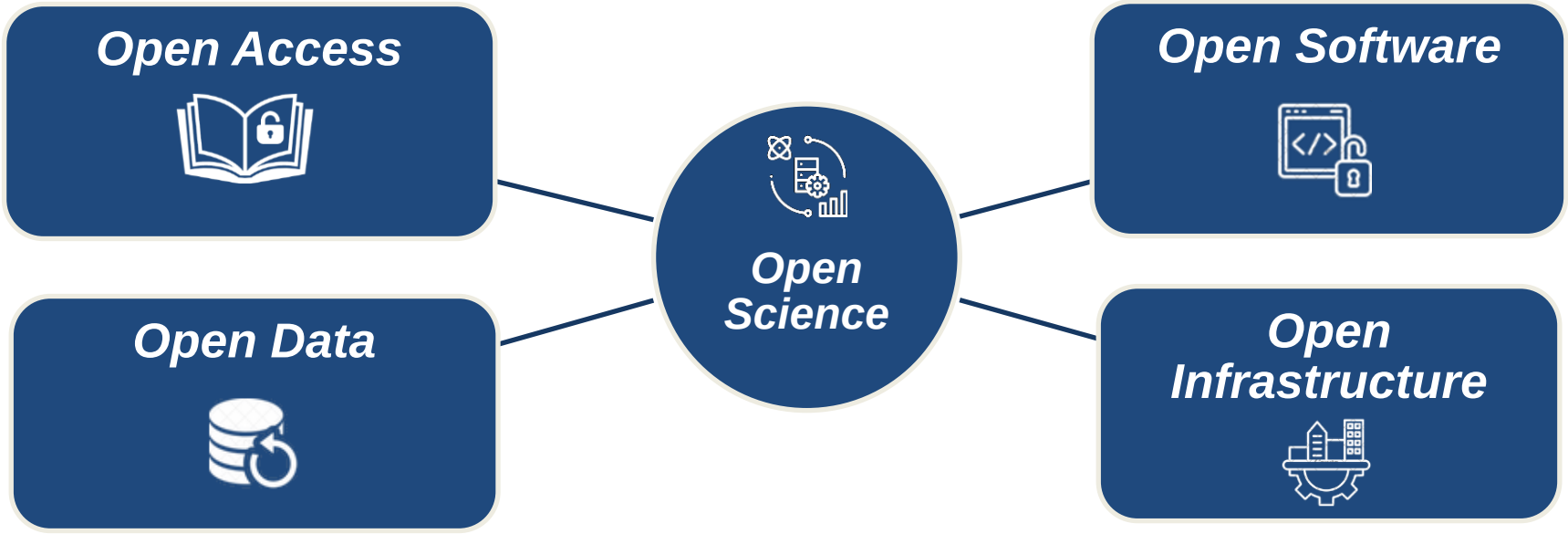
Offers a **new metric** for research assessment to remove the outdated dependence on e.g. h-index

Strengthens **technology transfer** with industry partners

On the agenda of many **governments and funding bodies** (**Helmholtz POF IV+** requires it)

Ultimately, the researcher should benefit from making research outputs open

The Pillars of Open Science



Open Science Policies/Statements

Number of Open Science polices as

broader statements

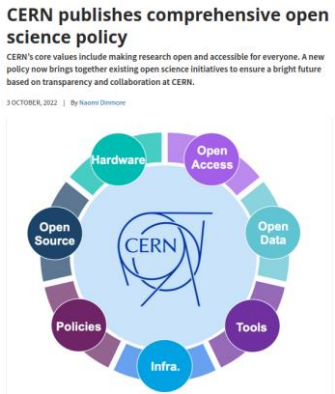
- Can be broad OS policy or separate Policy for each pillar

Institutional

Inter/national

Chapter on OS will be included in

NuPECC LRP



Literature: Moradi & Abdi., **Open Science-related policies in Europe**, *Science and Public Policy,* (2023) https://doi.org/10.1093/scipol/scac082

F.A.I.R Principles outlines (briefly for info)

Findable

- Centrally orchestrated storage and access of data essential to enable the data/software to be findable.
- Usage of Persistent IDentifiers (PID), Digital Object Identifiers (DOI) -> Guarantee access to Digital Research Objects.
- Generation of ‘data record’ in discipline specific repositories

Accessible

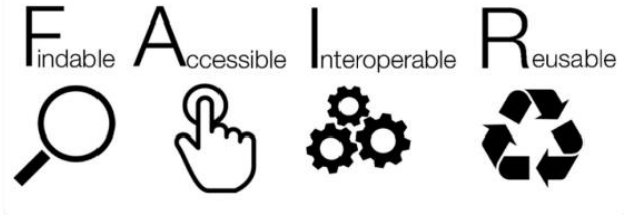
- Data and software produced/dedicated for F.A.I.R communities and publications centrally stored.
- Common & “user-friendly” interface to store and retrieve data

Interoperable

- Common metadata formats
- F.A.I.R-produced data operable with other datasets

Reusable

- Ensure (as reasonably as possible) data stored long term
- Metadata should be retained indefinitely

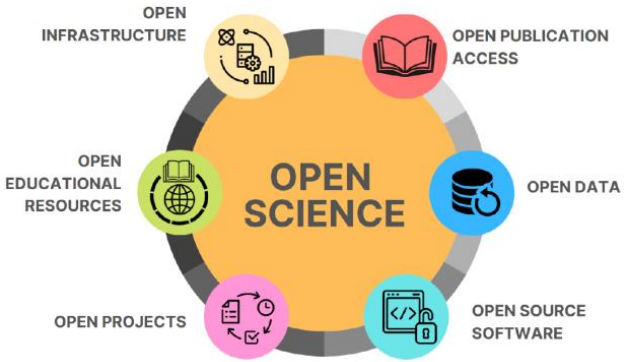


Wilkinson *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

https://www.go-fair.org/fair-principles/ -> See here for more detailed info

Open Science @ GSI/FAIR: What we want to achieve

- **Open Access Publications** –> Mandatory publication of Open Access articles
- **Research Data** –> Publish research data in suitable repositories (F.A.I.R. Data)
- **Open Software** –> Make open whenever possible (F.A.I.R. Software)
- **Open Infrastructure** –> Open Projects in research and industry
- Develop an **Open Science Ecosystem** to combine everything

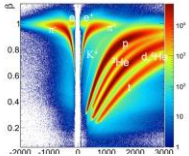


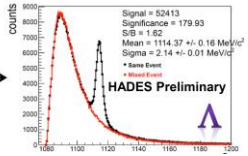
Considerations:

- The steps and processes to achieve this are complex… Start smaller and work up
- Do not make it too ‘general’ needs finer granularity and use-cases
- Aim to address **all researchers who use GSI/FAIR**: Students, Postdocs, PI’s, Group leaders…

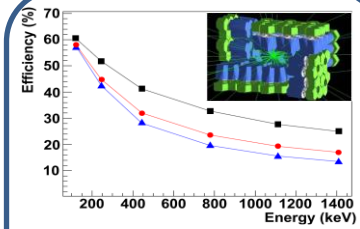
Research Data and Software at GSI: Rich and varied

Experimental Data:



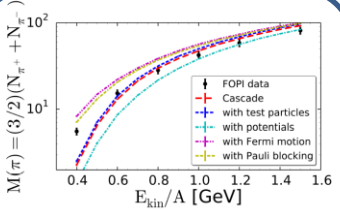


Raw *Pre-processed* *Result*

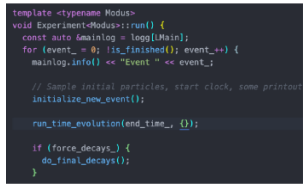


A.K.Mistry et al. NIM A 1033, 166662, (2022)

Simulation

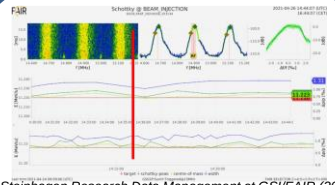


Theory



J. Staudenmaier Research Data Management at GSI/FAIR (2022)

Software



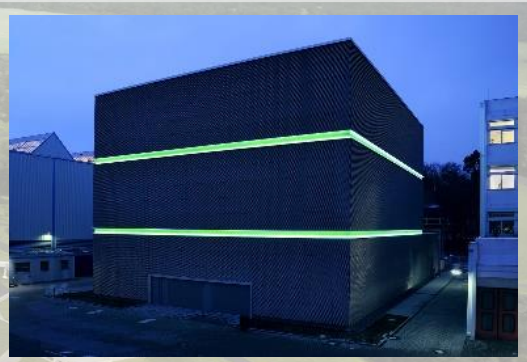
R. Steinhagen Research Data Management at GSI/FAIR (2022)

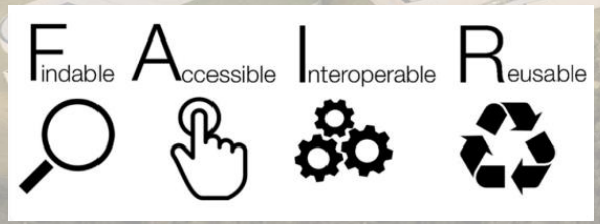
Machine Parameters

Research Area	Exp. Run Time	Raw Data	Calibrated Data	Simulations	Final Datasize
Heavy-ion reactions	3 weeks	130TB	300TB	150TB	<1TB
Materials Science	~Minutes	~MB - GB	-	-	10MB

FAIR Goes F.A.I.R

- **'FAIR goes F.A.I.R**': commitment to open science
- Towards the next generation "**data challenge**"
- **~TB/s data rates**, online processing, 105 -106 cores
- **Distributed computing** with a large user community
- **Data preservation** and **accessibility** key to success
- **HPC** Green cube expansion



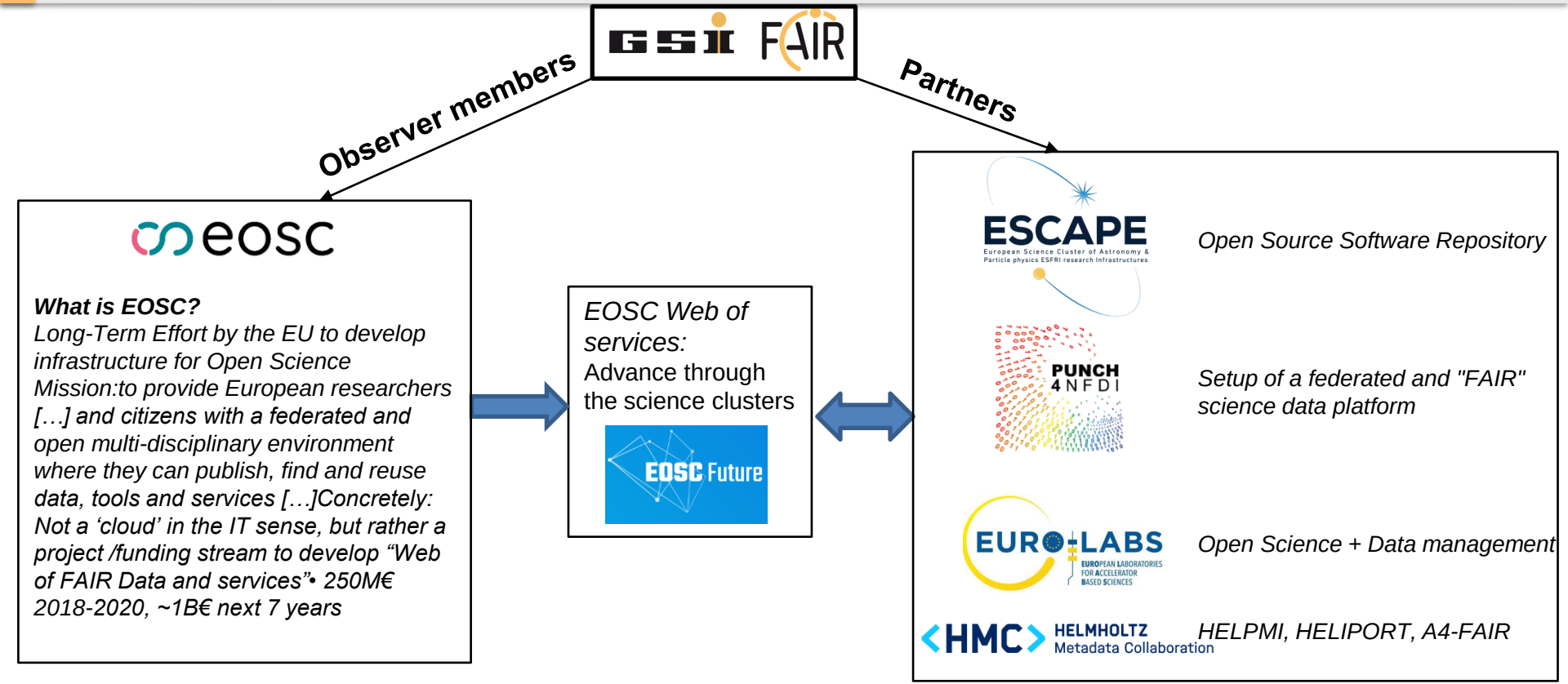


GSI/FAIR Involvement in External OS projects

GSI/FAIR Involvement in External OS projects

ESCAPE	*European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures:* Open Source Software Repository (OSSR) developer and maintainer
NuPECC	*Nuclear Physics European Collaboration Committee:* Participation and writing Open Science section of the LRP 2024
eosc	*European Open Science Cloud:* GSI/FAIR both observer members, contribution and suggestions for EOSC Future
EURO-LABS	*EuroLabs:* Work Package on Open, diverse and inclusive Science (C. Hornung)
PUNCH4NFDI	*Particles, Universe, NuClei and Hadrons for the NFDI:* Two Task areas; Developments on data portal, AAI, data lake and other infrastructure from GSI IT department and Research division
MT DMA	*Matter and Technology, Data Management and Analysis:* IT contributions
HELMHOLTZ Open Science	*HGF Open Science:* Members of the OS, software and POF IV indicators working groups
HMC HELMHOLTZ Metadata Collaboration	*Helmholtz Metadata Collaboration:* Participation in HMC funded projects, links and connections to Matter division
ELEMENTS	*Exploring the Universe from Microscopic to Macroscopic Scales:* Supporting Open Science area of the project (as well as other direct research areas)

EOSC + PUNCH + ESCAPE + …



Developments: Open Science working group @GSI/FAIR

Monthly meetings, started in April 2022

Members Researchers, IT department, library and documentation, accelerator division, scientific council representation, legal department, technology transfer department

Plan and discuss open science national/international initiatives, OS progress at GSI(FAIR), sharing of ideas, new tools.

Expected Outcomes Implement OS policies and guidelines, best practices, test new tools and promotion/communication facility-wide

Endorsed by the GSI/FAIR management: Sept 2023

FAIR GSI	Document type: **Terms of reference**	Date: 08.09.2023
		Page 1 of 3

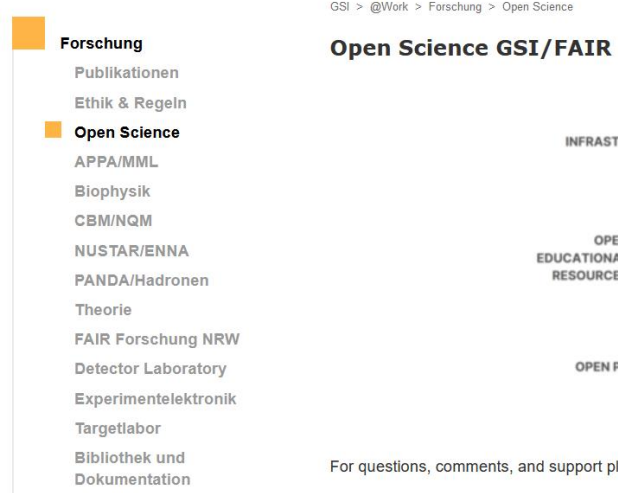
1 GSI/FAIR Open Science Working Group (OSWG) – Terms of reference

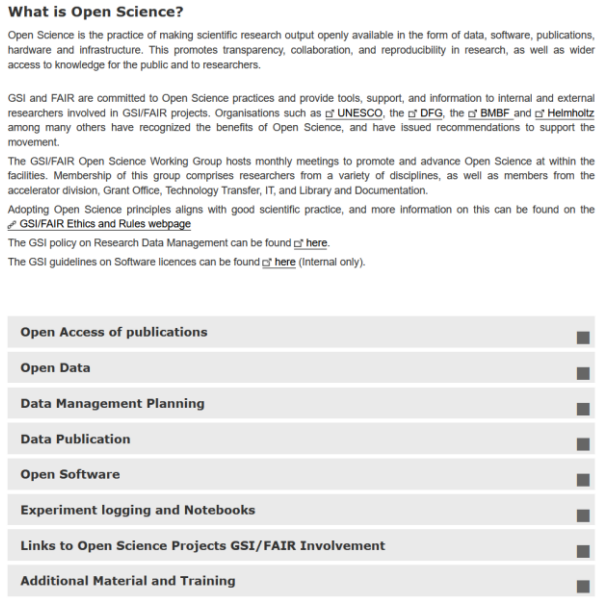
Open Science embodies the principle of making research output openly accessible and widely reusable through sustainable infrastructures, with as few barriers as possible. This can include open access publications, data, software, and hardware. Applying Open Science practices at GSI/FAIR will bring forth significant advantages through the dissemination of research output including networking and collaboration for science, industry, and society. Ongoing digitisation has opened doors to the development of essential infrastructure and tools necessary for enabling Open Science. The Open Science working group can help to develop and strategise Open Science at GSI and FAIR, as outlined in this document.
This ToR document defines the purpose, scope, objectives, membership and expected outcomes of the Open Science Working Group (OSWG) at GSI and FAIR. It serves as a mandate for the group to make recommendations (where necessary) to the GSI/FAIR management related to open science initiatives, and is officially endorsed by the current OSWG and the GSI/FAIR management. This ToR is a living document that can be updated as needed to reflect changes in the group's focus or activities.

Developments: GSI/FAIR Open Science Website

Email address: open-science@gsi.de

GSI/FAIR Open Science Webpage: https://www.gsi.de/open-science





Open Access Publication @ GSI

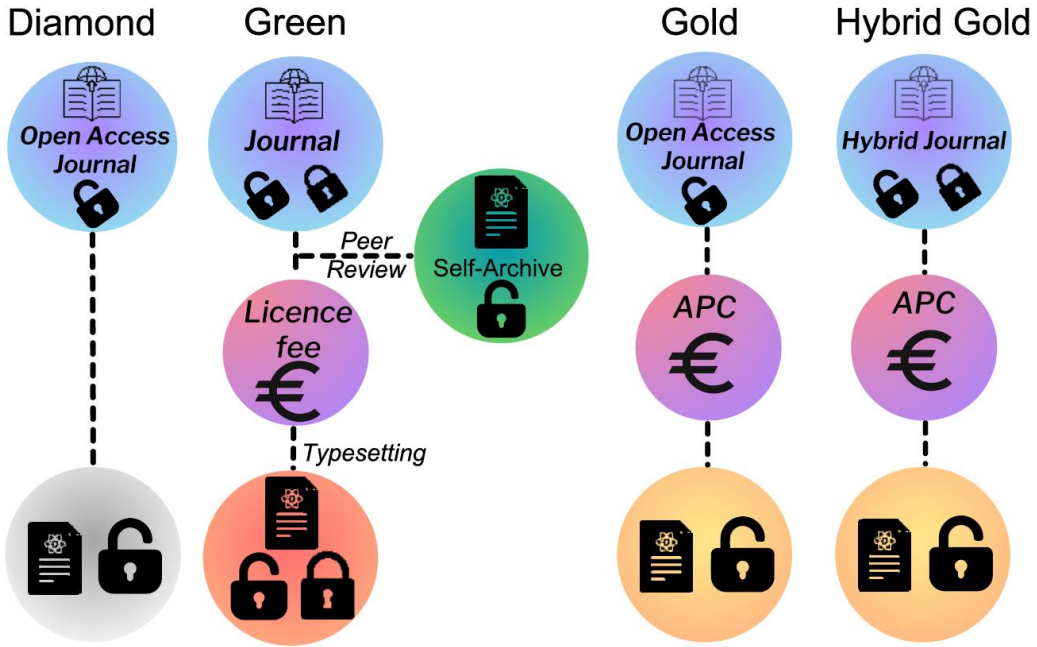
- **Open access publications 100% are now mandatory at GSI**
- **Publications Guidelines in preparation**

European Commission Policy on Open-Access to Scientific Publications and Research Data in Horizon 2020

David Guedj (✉), Celina Ramjoué

European Commission, Directorate General for Communications Networks, Content and Technology
Brussels, Belgium
http://ec.europa.eu/dgs/connect/en/content/dg-connect

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities

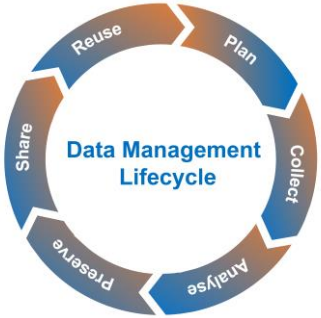


https://www.gsi.de/en/bibliothekunddokumentation/open_access_gsi

Open Research Data and Management

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, which follow "as open as possible, as closed as necessary"

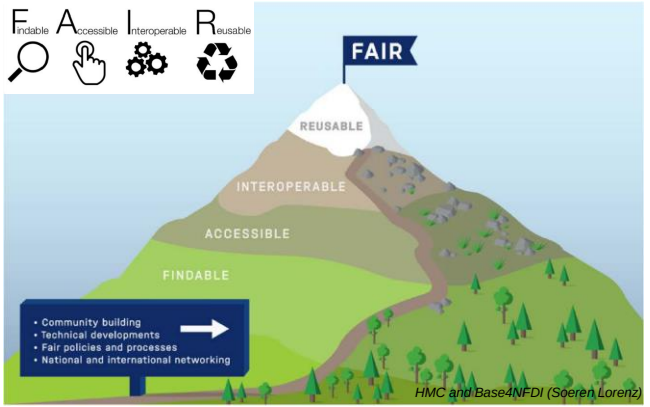


FAIR Data is not an end goal

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

Goals:

- to ensure good RDM practices at GSI/FAIR;
- promote and assist researchers in publishing data;
- to aim (as best as reasonably possible) that data is published according to the Findable Accessible Interoperable and Reusable principles;
- develop the tools and infrastructure needed to do this (repositories, Electronic logbooks …)



Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* **The FAIR Guiding Principles for scientific data management and stewardship**. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

GSI/FAIR Research Data Management Policy

- **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

FAIR GSI	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

(a) 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.

(b) 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.

(c) 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.

(d) 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.

(e) 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 Status @ GSI/FAIR: Guidelines

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 *Example 1 Materials Science* 8
5.2 *Example 2 Large dataset* 9
5.3 *Example 3 PHELIX* 9
5.4 *Example 4 ESR* 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

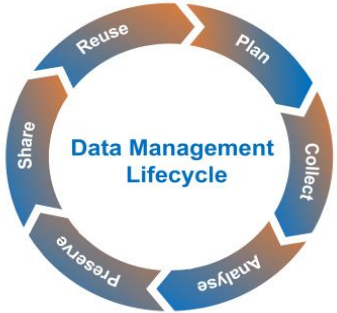
Includes legal and transfer issues

RDM Status @ GSI/FAIR:Data Management Plans

- 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

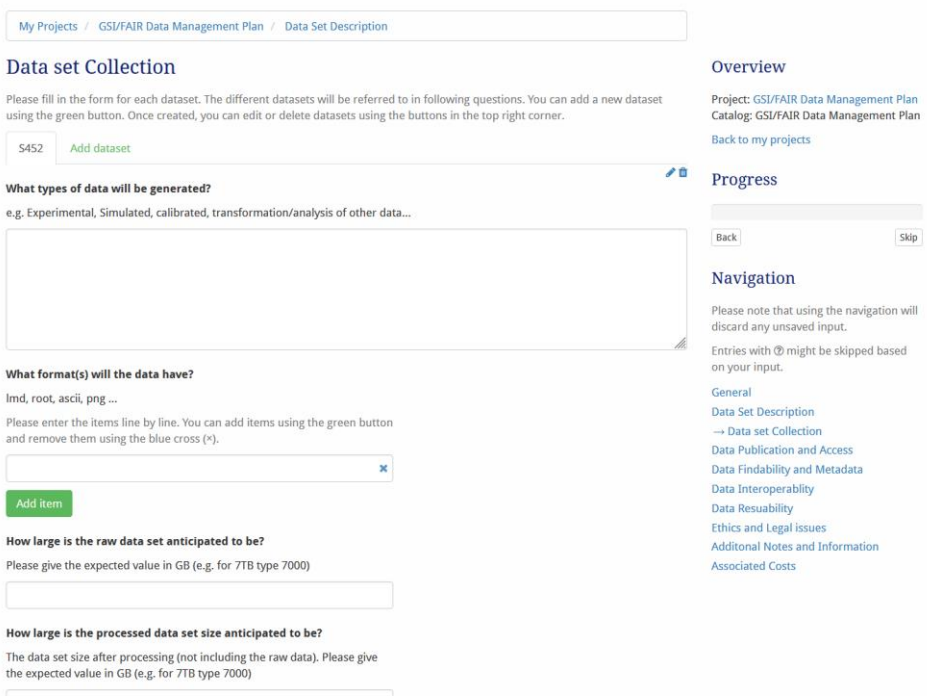


Contents can include:

- 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

RDM Status @ GSI/FAIR: Data Management Plans



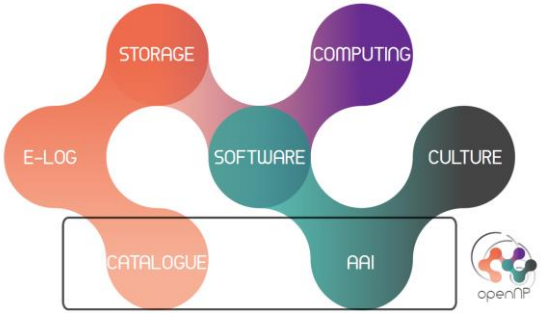
- Data Management with easy to use tool: ->**RDMO**
- Currently running test version at GSI: Catalogue setup and feature testing
- Plan to go live in Autumn/Winter 2023
- *Plan to employ in combination with GATE proposal submission system 2024+*

Example Project: EURO-LABS WP 5.2: Open Science and Data Management

- *Participants:* CSIC, GANIL (Leading partners), INFN, CNRS, IJCLab, GSI
- Task leader: A. Lemasson (Ganil); subtask leader: C. Hornung (GSI), A. Matta (CNRS), M. Jouvin (IJCLab)

Goals:

- Bringing the nuclear physics community into the EOSC (European Open Science Cloud) framework
- Developing services to enhance FAIR (Findable, Accessible, Interoperable and Reusable) data principles
- Integration of Nuclear Physics community to existing infrastructures/services of EOSC environment - using present experience from ESCAPE/HEP physics community
- Task 1 : Data Management Plan (A. Lemasson -- GANIL)
- Task 2 : openNP Research Data catalog (A. Matta -- LPC Caen)
- Task 3 : New Services for improved access to data (and shared services)
  - 3.1 : Authentification and Authorization Infrastructure (M. Jouvin -- IJC Lab)
  - 3.2 : Prototype of data access platform (C. Hornung -- GSI)

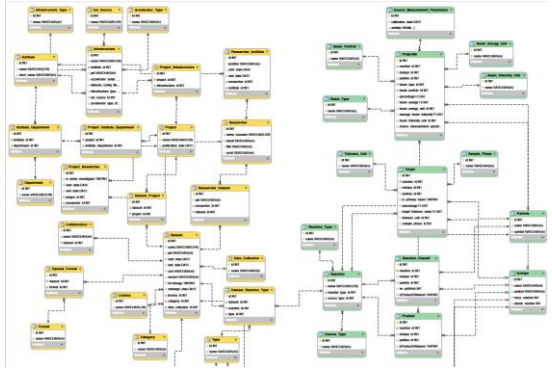


Metadata for nuclear physics experiments

- ***When publishing data, also publish machine readable metadata***
  - Allows datasets to be found
  - Enables interoperability between datasets
  - Enables reprocessing of data: transparency and integrity
  - Efficient use of resources
- **No common schema existing between accelerator/nuclear physics experiments**
- **European (and beyond) wide strategy** -> commonalities, leverage open science projects

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 fiel	String
	Principal Investigator Email		
	Principal Investigator PID		
	Facility/Institute Information		
1+	Facility/Institute of Data Generation	Of data generation	String
1+	Facility/Institute ROR		
1+	Infrastructure	Overlaps with detector/apparatus? - ir	String
1+	Department/Division	The department/division associated w	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. Gunther +EURO-LABS, PUNCH4NFDI

Open Research Software and Management

Different to data management -> maintainance 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
- ➢ **Research software guidelines** under preparation
- ➢ **Software can be onboarded** to repositories: curation and promotion. e.g. ESCAPE OSSR *https://projectescape.eu/ossr*

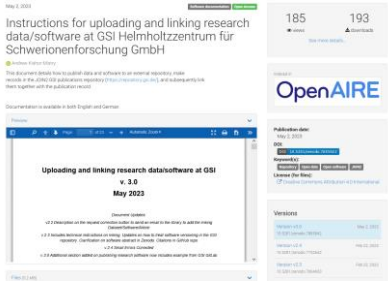
Barker, M., Chue Hong, N.P., Katz, D.S. *et al.* **Introducing the FAIR Principles for research software**. *Sci Data* **9**, 622 (2022). https://doi.org/10.1038/s41597-022-01710-x

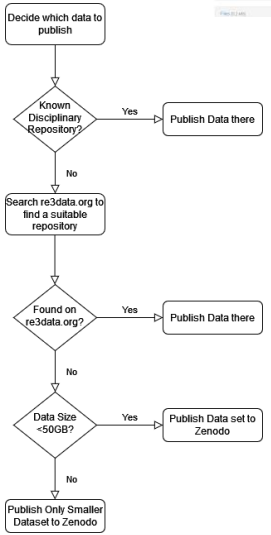
Findable Accessible Interoperable Reusable

How To: Data/Software publication (interim solution)

- Plans for ***GSI data repository*** underway
- In the meantime using **Zenodo/discipline specific repos for data**
- **Instructions available on how to publish data at GSI**
- Data sets can be very large in NuSTAR! Complications on publishing (too large, huge compute + personnel resources needed to process data from the raw stage)
- Possible solutions: **Semi-derived (pre-processed) data** could be published. 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**
- Longer term goals of projects like ESCAPE, PUNCH4NFDI for more complex datasets and workflows

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





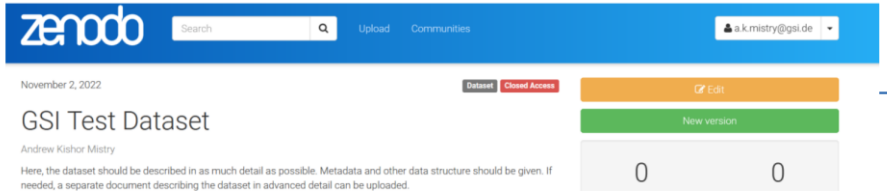
HEPData
Repository for publication-related High-Energy Physics data

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

zenodo

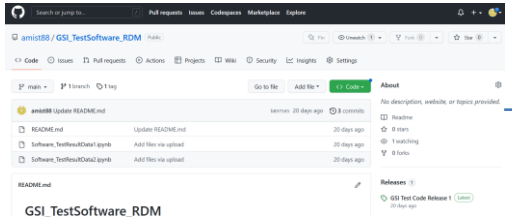
How To: Interim Data Publication strategy

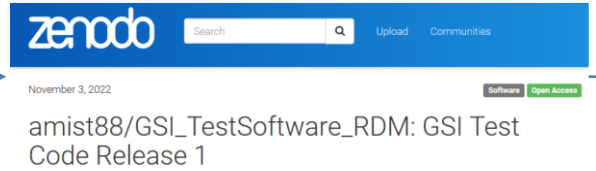
Dataset



DOI

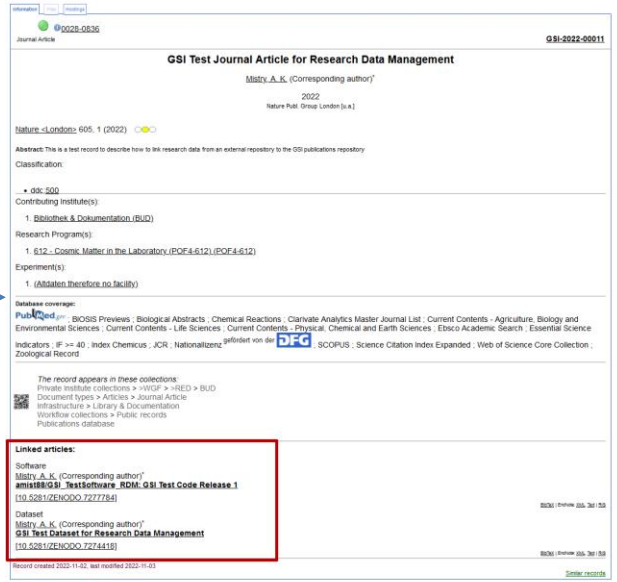
Software





DOI

GSI Publications Repository



Article -> Dataset -> Software linking

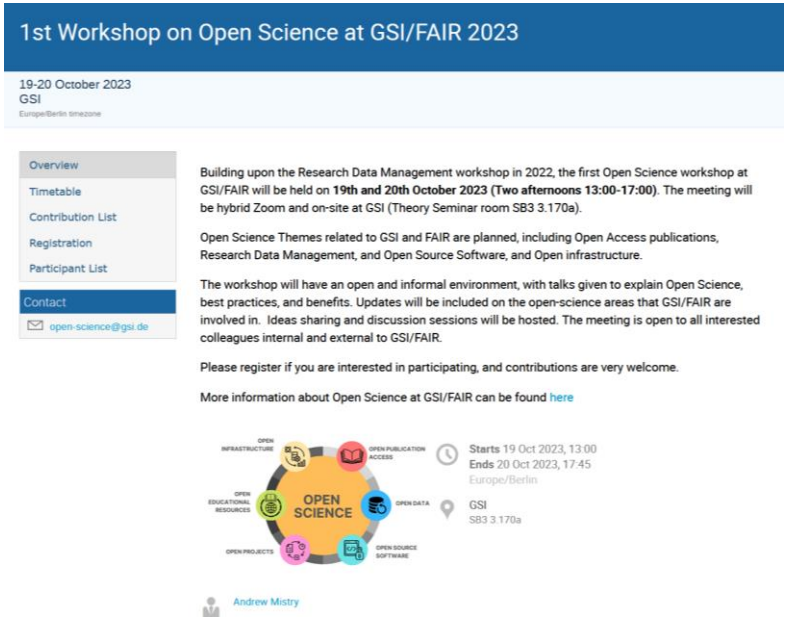
Datasets/software records in GSI publications repository count towards POF IV

Open Science Workshops @ GSI/FAIR

Research Data Management workshop 2022: Slides and information available on indico
https://indico.gsi.de/event/14680/

⇒ **Follow up workshop on Open Science** (Hybrid event) 19th + 20th October (2 afternoons) **https://indico.gsi.de/event/17498/**

⇒ If you're interested, please feel free to register!



Open Science @ GSI/FAIR

Open Science cultural change under development at GSI/FAIR, a step-wise iterative approach is underway to best suit the research areas

Open Science Working Group: tools, strategies and developments ongoing internally and externally

Continued contribution to OS projects (e.g. PUNCH4NFDI, ELEMENTS, Eurolabs, EOSC…) and external communities and sharing of ideas and infrastructure developments

Communication with Researchers feedback, information distribution and developments

We are looking for use cases on data/software publication – please get in touch!

Open Science related questions or Comments: open-science@gsi.de
Website Open Science @ GSI/FAIR: https://www.gsi.de/open-science
(Hybrid) Open Science Workshop @ GSI 19/20th October https://indico.gsi.de/event/17498/

Thanks for your attention…