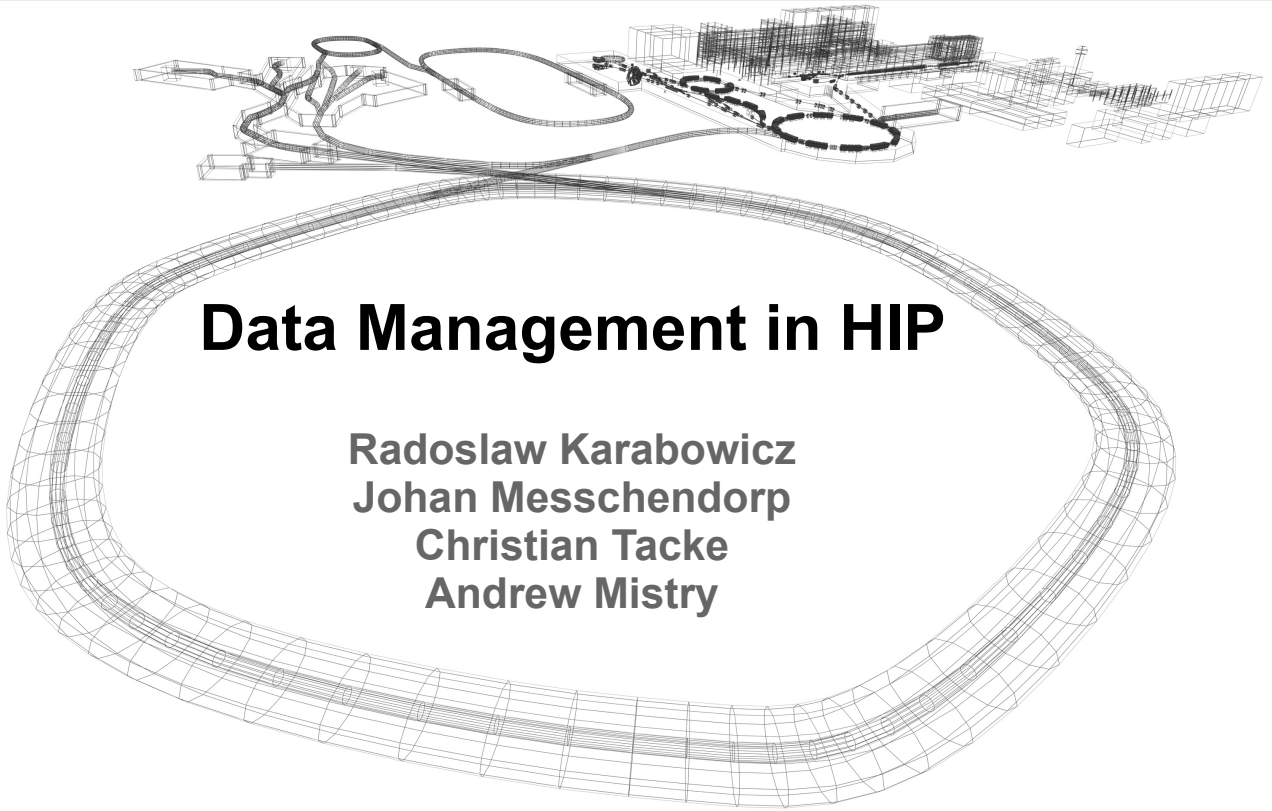
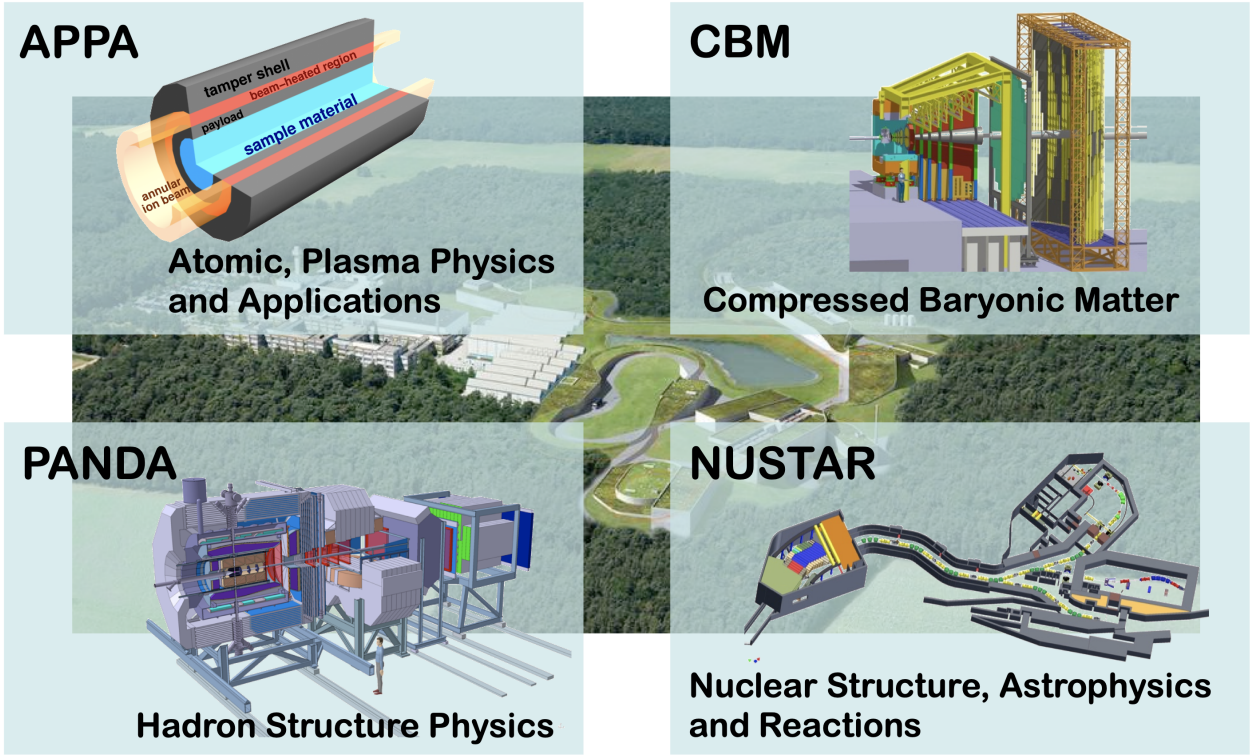


Data Management in HIP

Radoslaw Karabowicz
Johan Messchendorp
Christian Tacke
Andrew Mistry

Four experimental pillars of FAIR



CBM and **PANDA**

- more than 500 members each
- HEP like detectors

APPA and **NUSTAR**

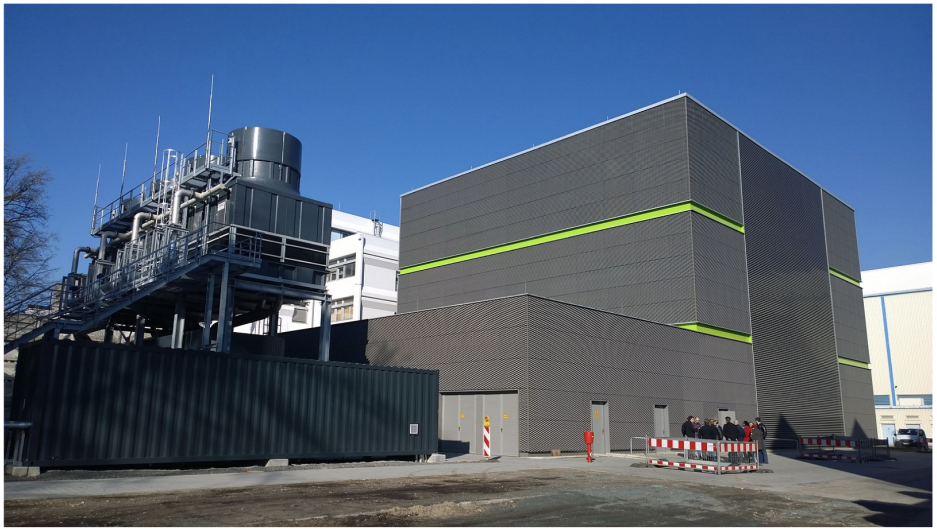
- 700-800 members each
- many small detectors / sub-collaborations

Computing at FAIR

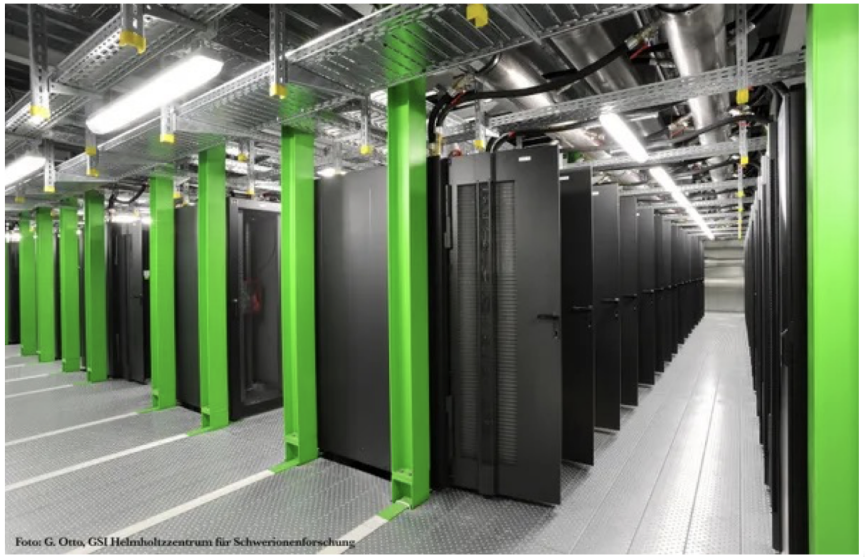
CPU	CBM	PANDA	APPA	NUSTAR
number of cores (for simulations)	45k	68k	11k	9k
number of cores (for online)	45k	34k	-	7k

Storage	CBM	PANDA	APPA	NUSTAR
disk total (TB)	103,000	60,680	7,037	34,250

Green IT Cube



In operation since 2016



PUE < 1.07
4 MW cooling
capacity for 768 racks on 6 floors

Classic approach vs Containerized solution

Kronos – **classic** approach:

in production from 2015 until 2020

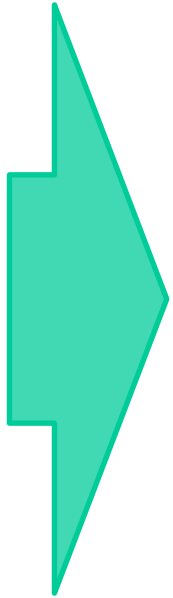
- same operating system on every node,
- more than 4000 software packages installed,
- operated without upgrading underlying OS or Slurm.

Classic approach vs Containerized solution

Kronos – **classic** approach:

in production from 2015 until 2020

- same operating system on every node,
- more than 4000 software packages installed,
- operated without upgrading underlying OS or Slurm.



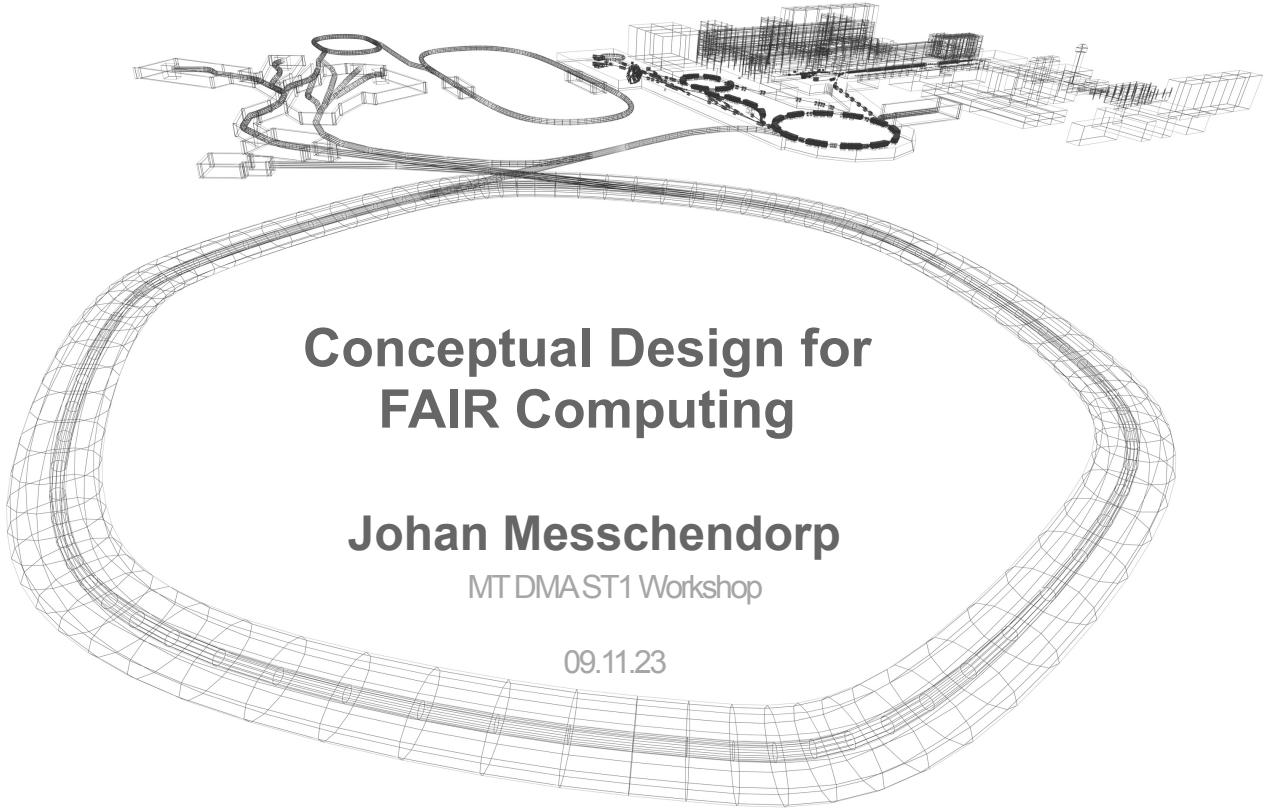
Virgo – **fully containerized** approach:

in production since August 2020

- separates user application space from host system,
- jobs executed in a container,
- minimal host system: Slurm and Apptainer,
- users are free to choose the OS and install software,
- admins are free to upgrade host OS and/or Slurm at any time.

Virtual Application Environment (VAE)

- Ready to use solution provided by GSI-IT:
  - login into container – interactive session on submitter node,
  - submitted jobs run in the same VAE,
  - software stack for VAE installed on CVMFS and mounted on run-time,
  - produced data stored on LUSTRE.
- Containers' definition files available in a git repository:
  - easy start for advanced users to build their own containers.



Conceptual Design for FAIR Computing

Johan Messchendorp

MT DMA ST1 Workshop

09.11.23

Facility for Antiproton and Ion Research - "The Universe in the Laboratory"

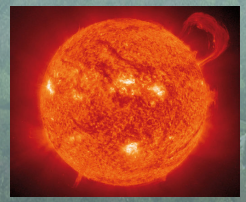
203x

Facility for Antiproton and Ion Research - "The Universe in the Laboratory"

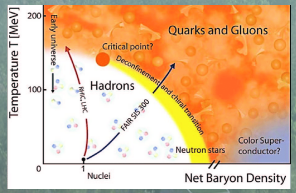
Cutting-edge science and technology

- ESFRI Landmark near Frankfurt, Germany (ESCAPE)
- Top priority for European Nuclear Physics Community
- International: 50 countries, 3000 researchers
- Diverse community from atomic to particle physics
- High intensity+precision+diversity+parallel operation
- Monolitic and modular experimental setups

APPA

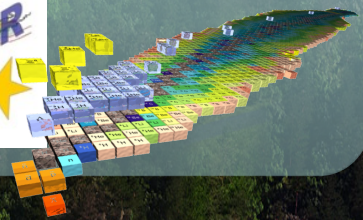


CBM



panda

NUSTAR



+THEORY and BEAM physics

203x

Facility for Antiproton and Ion Research - "The Universe in the Laboratory"

Cutting-edge science and technology

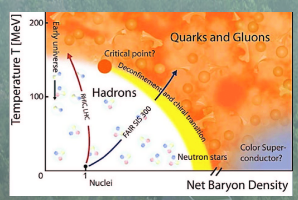
- ESFRI Landmark near Frankfurt, Germany (ESCAPE)
- Top priority for European Nuclear Physics Community
- International: 50 countries, 3000 researchers
- Diverse community from atomic to particle physics
- High intensity+precision+diversity+parallel operation
- Monolitic and modular experimental setups

Towards the next generation "data challenge"

- Volume, Velocity, Veracity, Variety, and Complexity!
- ~TB/s data rates, online processing, ~5×10^5 cores
- Data stored on disk ~35 PB/year
- Distributed computing with a large user community
- Committed to "open-science" (FAIR) concept

APPA

CBM



panda

NUSTAR

+THEORY and BEAM physics 203x

Conceptual Design Report for FAIR Computing?

Why?

- Comprehensive overview of FAIR-research computing needs and plans
- Base document for funding and policies etc.

Who?

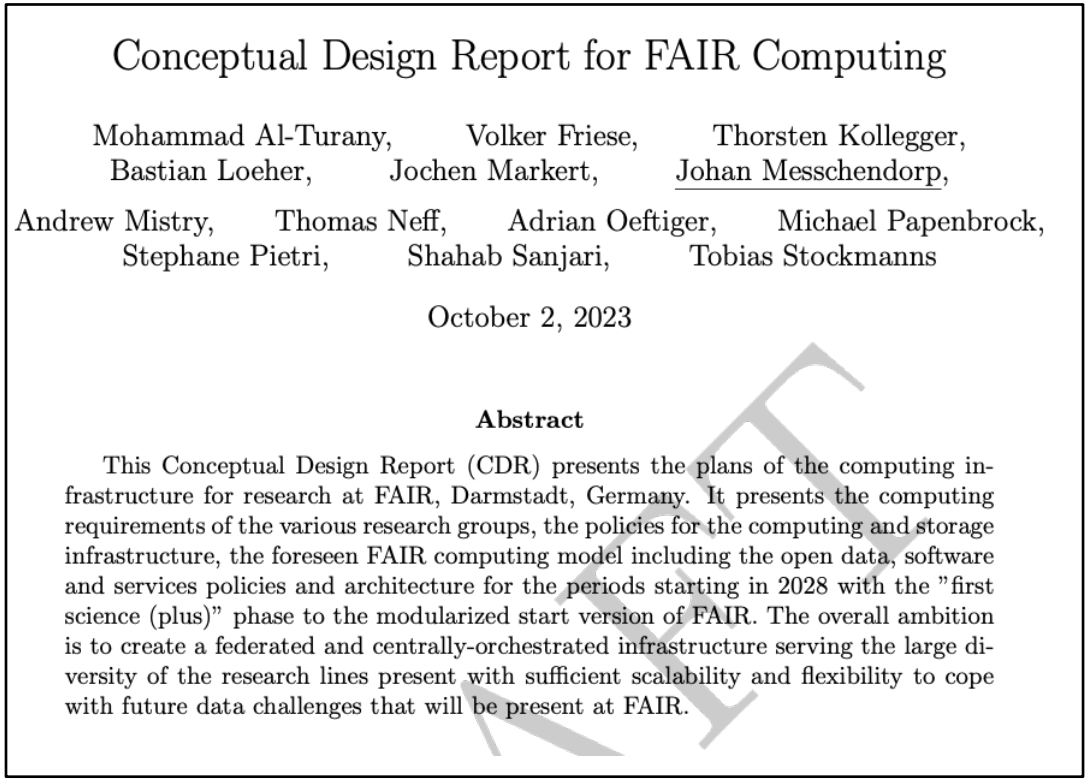
- GSI research-IT, FAIR research lines, accelerator group

What?

- Conceptual design for research IT
- Research requirements, research responsibilities versus IT support, FAIR computing model, F.A.I.R principles, R&D, …
- FS(+) and MSVc timelines

When?

- A.S.A.P.
- Draft (~70 pages) submitted beginning October to ECE/ECSG, in review



Conceptual Design Report for FAIR Computing

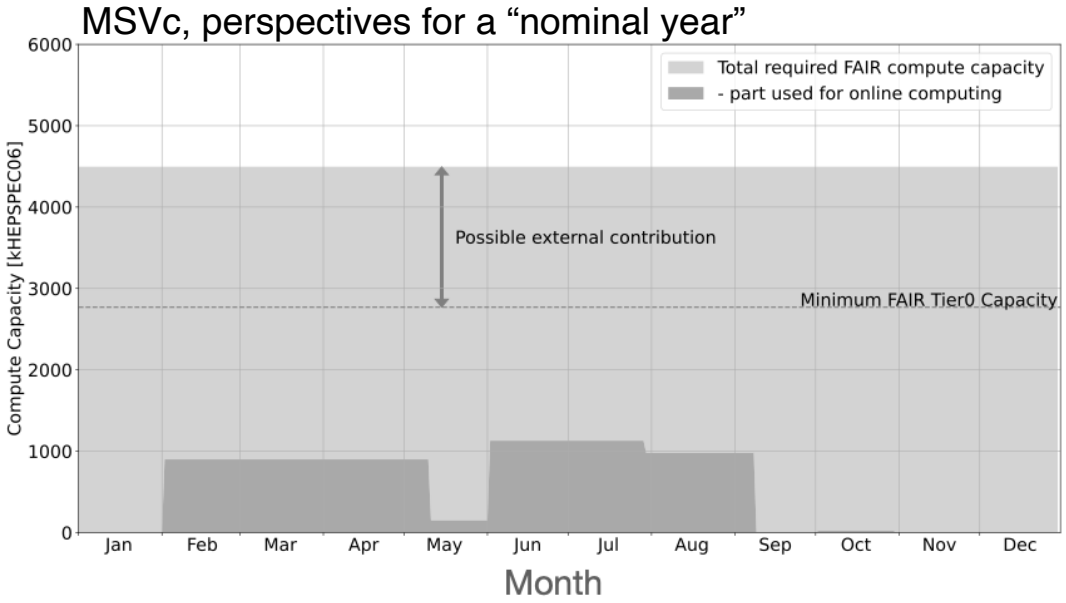
Mohammad Al-Turany, Volker Friese, Thorsten Kollegger, Bastian Loeher, Jochen Markert, Johan Messchendorp, Andrew Mistry, Thomas Neff, Adrian Oeftiger, Michael Papenbrock, Stephane Pietri, Shahab Sanjari, Tobias Stockmanns

October 2, 2023

Abstract

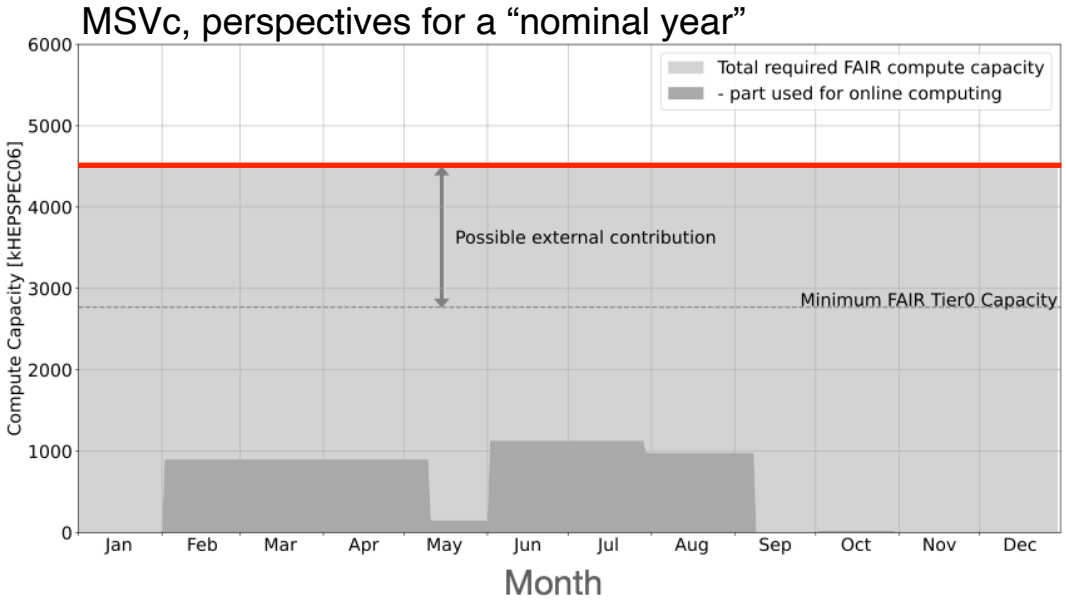
This Conceptual Design Report (CDR) presents the plans of the computing infrastructure for research at FAIR, Darmstadt, Germany. It presents the computing requirements of the various research groups, the policies for the computing and storage infrastructure, the foreseen FAIR computing model including the open data, software and services policies and architecture for the periods starting in 2028 with the "first science (plus)" phase to the modularized start version of FAIR. The overall ambition is to create a federated and centrally-orchestrated infrastructure serving the large diversity of the research lines present with sufficient scalability and flexibility to cope with future data challenges that will be present at FAIR.

Compute and storage requirements



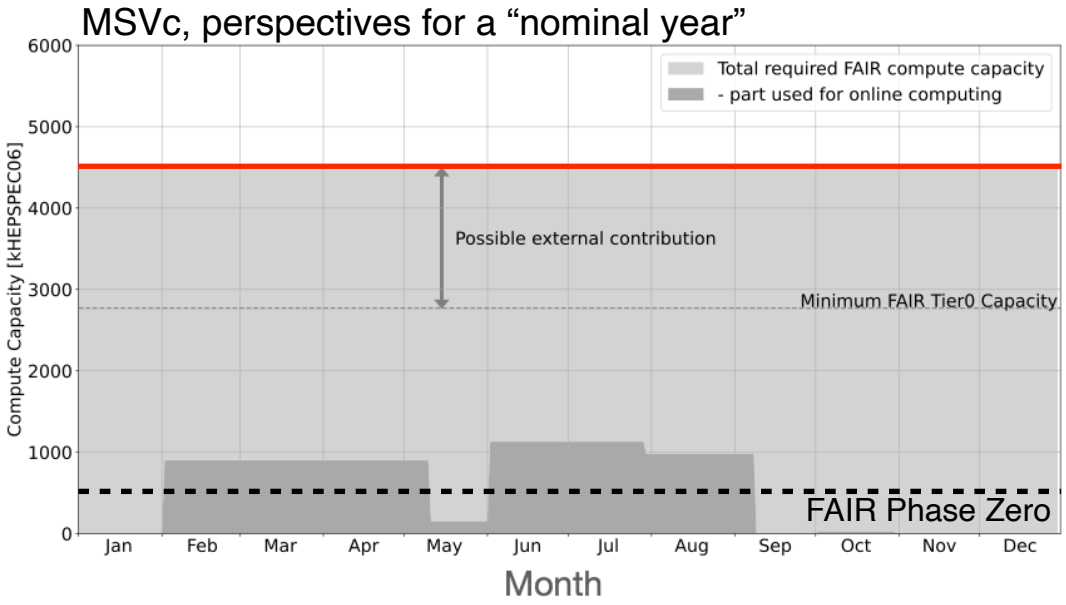
1 physical core (Intel E5-2680v4@2.4GHz) ~22 HEPSPEC06

Compute and storage requirements



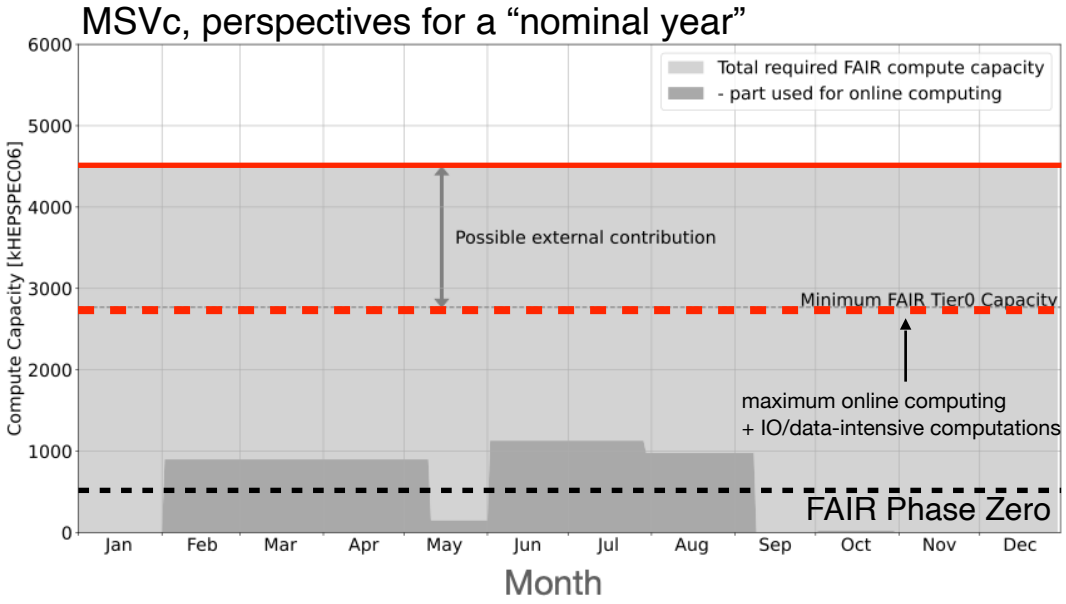
1 physical core (Intel E5-2680v4@2.4GHz) ~22 HEPSPEC06

Compute and storage requirements



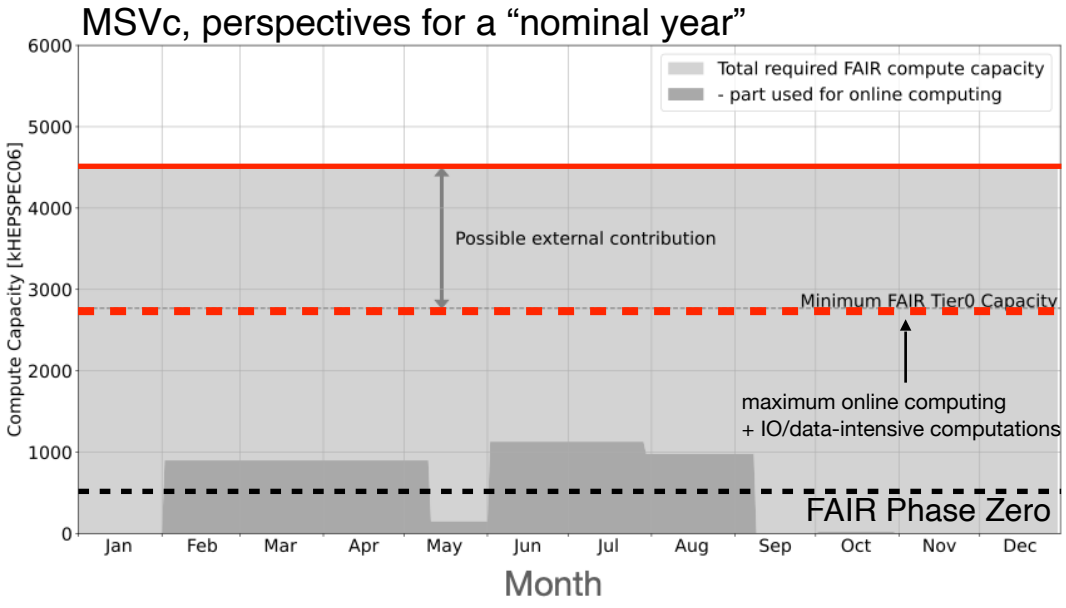
1 physical core (Intel E5-2680v4@2.4GHz) ~22 HEPSPEC06

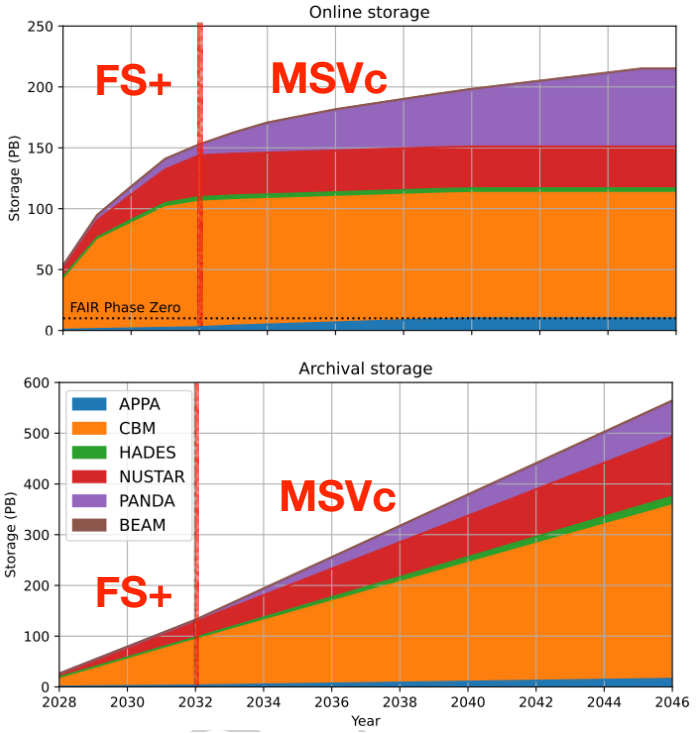
Compute and storage requirements



1 physical core (Intel E5-2680v4@2.4GHz) ~22 HEPSPEC06

Compute and storage requirements

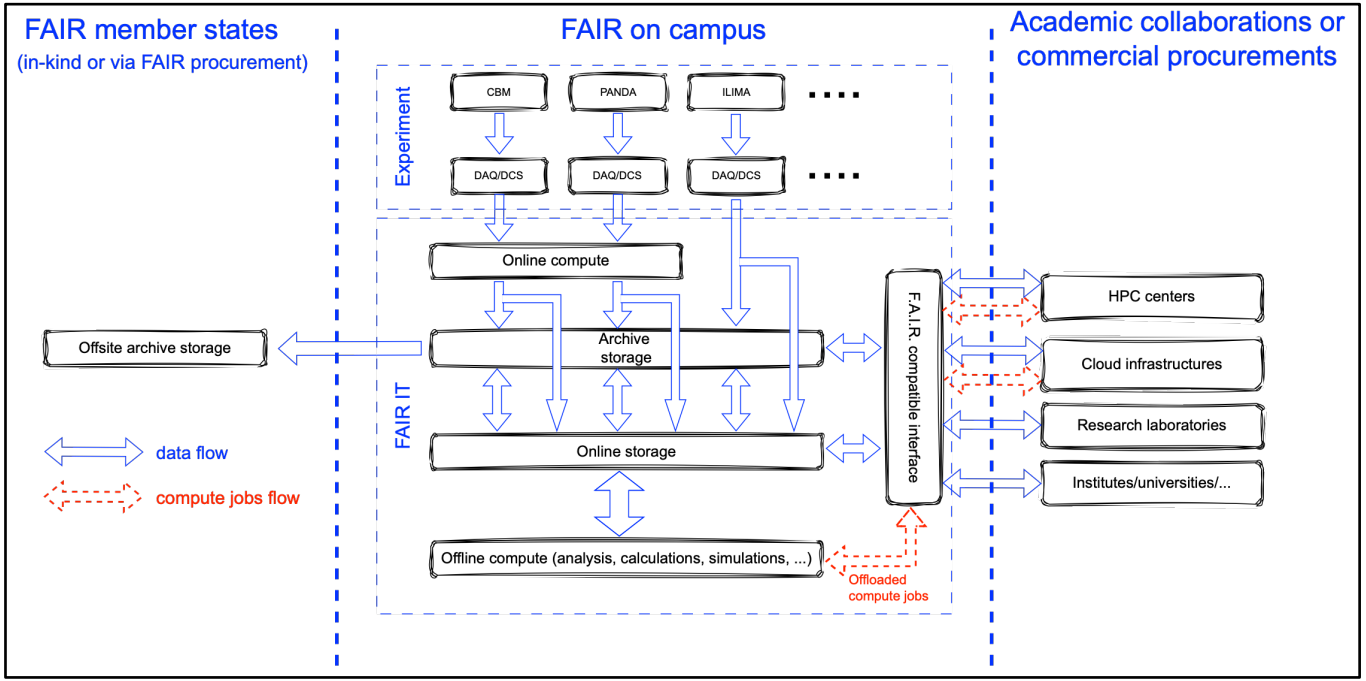




1 physical core (Intel E5-2680v4@2.4GHz) ~22 HEPSPEC06

FAIR Computing Model

- TIER-0 model supporting **online** & **offline** data **production** and **processing**
- TIER-0 infrastructure based on existing **GREEN-IT Cube**
- **Centralised storage** with external copies of archived data
- Distributed computing based on **federated model**
- FAIR-IT responsible of implementing **F.A.I.R interface**
- Research lines responsible for openness of data and services



FAIR and F.A.I.R principles

Findable

- Central orchestrated storage and access of data
- Consistent usage of PID such as DOI for data and metadata

Accessible

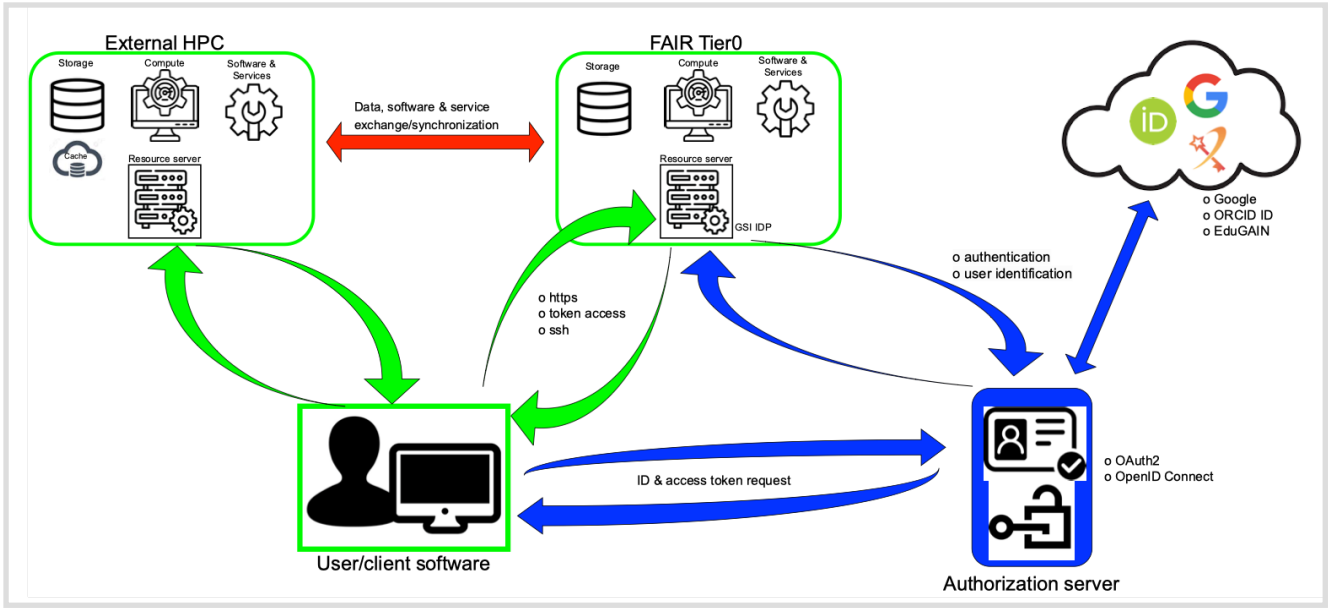
- Central storage of data and software for FAIR communities
- Data accessible via standard http protocols
- Token-based AAI, SSO authentication support, in line with domain -driven activities (ESCAPE)
- Data and software available under suitable licenses

Interoperable

- Participate in community-wide open-science initiatives on institutional, national, and European levels
- Follow-up developments in ESCAPE, e.g. “data lake” etc.
- Support OSSR service
- Agree upon controlled metadata vocabularies within research domain (GSI/Andrew Mistry)

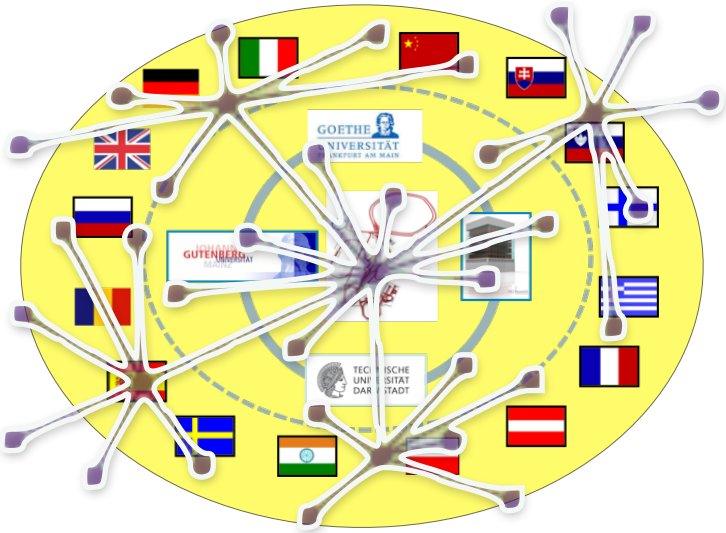
Reusable

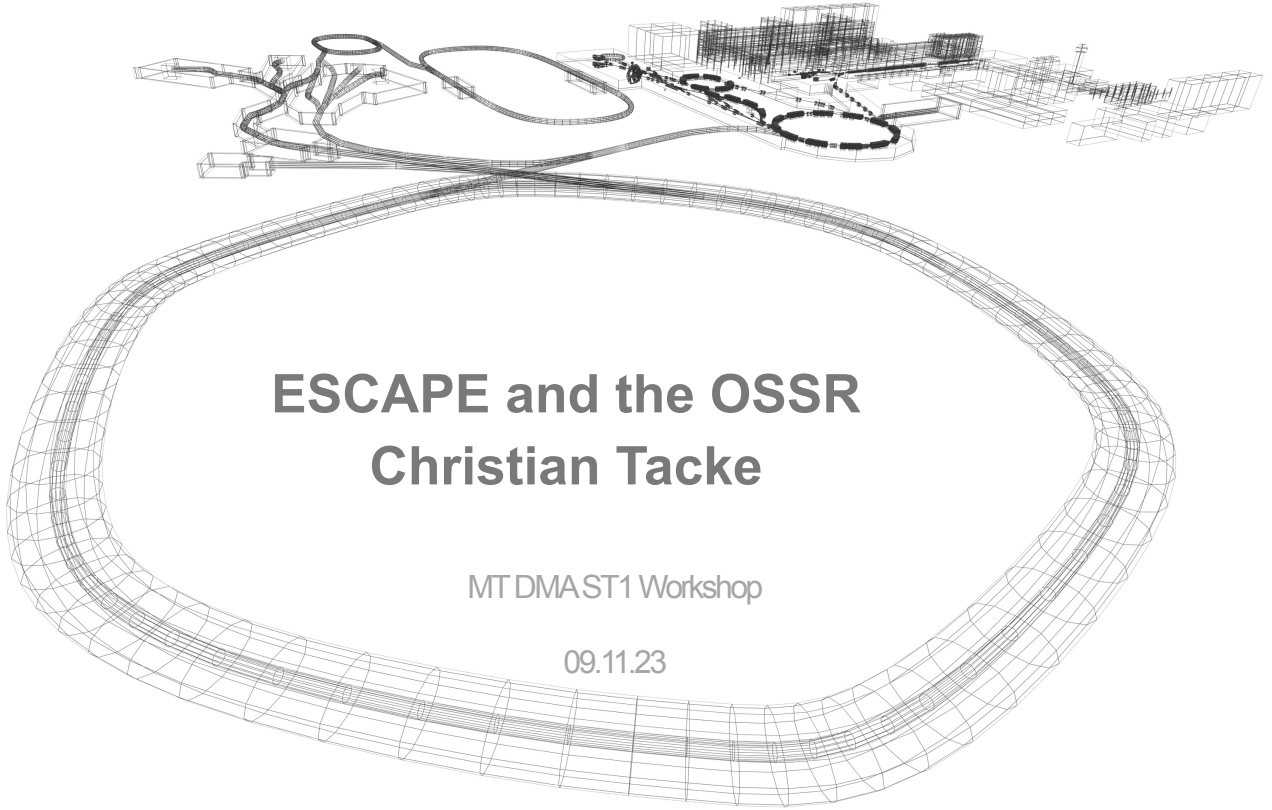
- Follows naturally from a successful implementation of “F”, “A”, and “I”



FAIR Computing in a Nutshell

- Towards **federated computing** among large centres
- Effective **resource sharing** at FAIR TIER-0 account for most of the data-driven computations
- Federated storage and computing with **local centres** using Teralink network & **commonly used standards**
- **Containerised approaches** and other **virtualisation** methods for flexible compute operations serving diverse community & optimise usage
- Data & services access using standard **token-based protocols** (http); AAI using widely accepted standards
- **Participate** in **domain-specific open-science** inspired activities





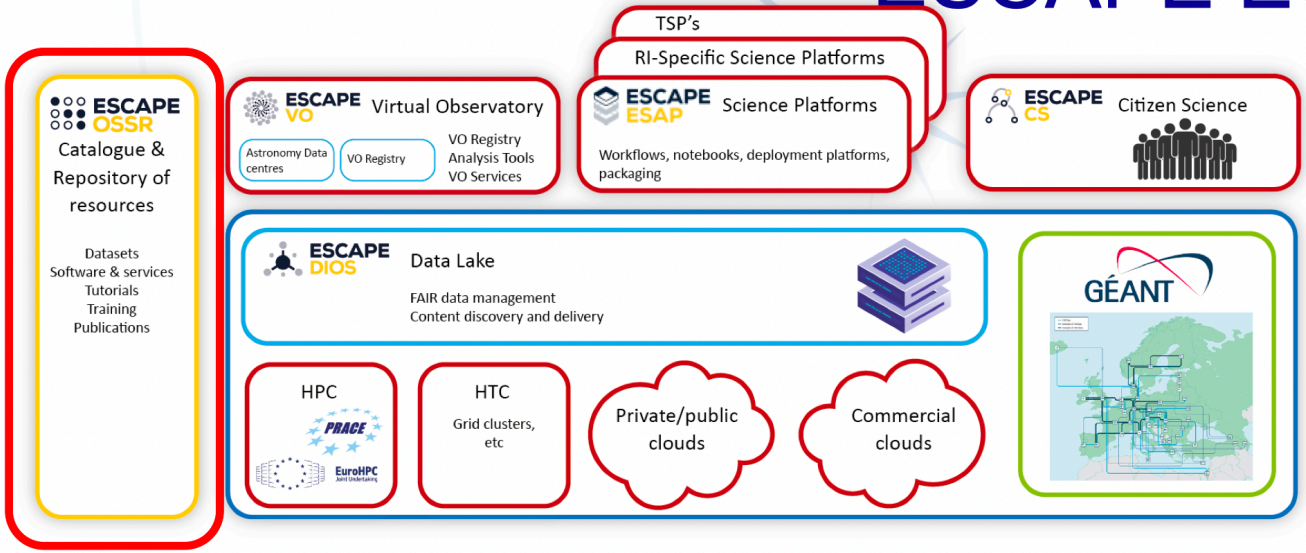
ESCAPE and the OSSR

Christian Tacke

MT DMA ST1 Workshop

09.11.23

ESCAPE EOSC cell

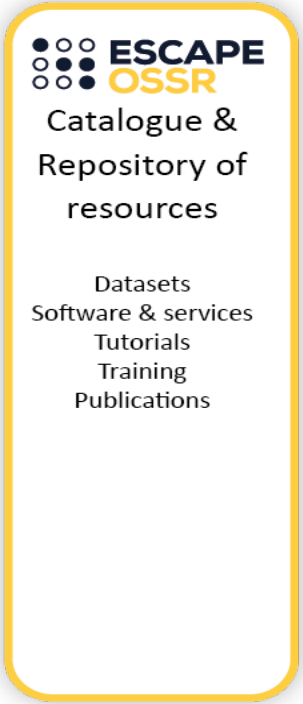


OSSR was developed as part of the ESCAPE project (Astro/particle, Particle Physics and Astronomy Research Infrastructures) in the EOSC (European Open Science Cloud)

- servicing the needs of the RIS
- Since 03/2023: ESCAPE is an Open Collaboration

Funded by the European Union's Horizon 2020 - Grant N° 824064

OSSR Overview (1)



The ESCAPE Open-source Scientific Software and Service Repository (OSSR) is a **sustainable open-access repository** to share **scientific software, services** to the **astro-particle-physics-related communities** and enable open science. It is built as a **curated Zenodo community** integrated with **dedicated tools** to enable a complete software life-cycle. The OSSR is fully onboarded into the **EOSC explorer**.

OSSR Overview (2)

ESCAPE OSSR

Catalogue & Repository of resources

Datasets
Software & services
Tutorials
Training
Publications

Development Platform

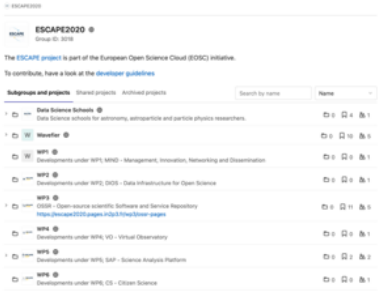
- Software Development
- Integration & Automation

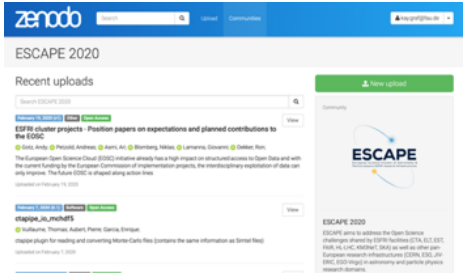
Repository

- Service Aggregation
- Preservation / Archive

Landing Page

- Entry point, Link Aggregation
- Search





Organisation

- Transition from project organisation into open collaboration under an MoU
- Three main areas of collaboration:
  - Policy, Strategy and Collaboration
  - Onboarding → Curated Repository!
  - Technical Developments
- Aims:
  - Collect software to provide additional visibility and cite-ability for RSEs; strengthen software competence with quality in focus
  - Use of OSSR as forum to foster publications/opportunities
  - Offer standards for new communities to join

Onboarding your Software

- Basic requirements
  - Mature State, Follow current software engineering practices
  - Publicly available and Open-Source licence
  - Some small things (like introductory documentation)
- Add Metadata (CodeMeta) to your software
- Give Onboarding Talk
- Publish on Zenodo
- Curation

Onboarding Talk

- Get a slot during the Onboarding Process
- 20 minute talk, 10 minutes discussion
- Template available
- Projects with recorded talk will get much higher visibility

Tooling

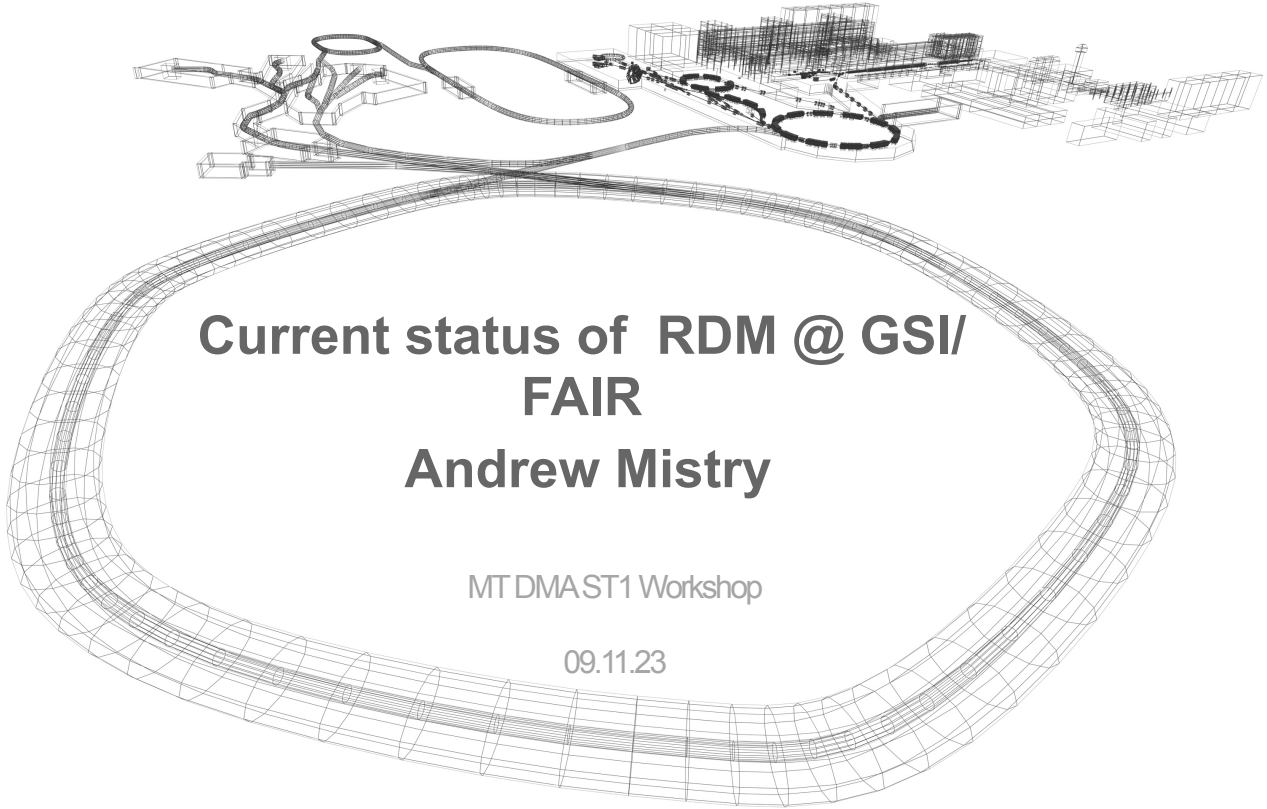
- CodeMeta (initial) Generator
- eOSSR library to interact with the OSSR
- CI / CD helpers:
  - Update your codemeta.json
  - Check your codemeta.json
  - Convert to zenodo metadata
  - Upload to Zenodo
  - Snippets / Integrations for Gitlab CI and Github Actions

Cooperations / Integrations

- Helmholtz MT-DMA has a cooperation with OSSR and will publish all its software as part of the OSSR
- Cooperation with the Netherlands eScience Center to integrate with the Research Software Directory (research-software-directory.org)
  - Currently in active development
  - Will also be used to integrate with the Helmholtz RSD (helmholtz.software)

Links

- OSSR Entry Page: https://purl.org/escape/ossr
  (nearly everything else should be linked from here)
- eOSSR Library: https://gitlab.com/escape-ossr/eossr
- Scientific Paper:
  The ESCAPE Open-source Software and Service Repository
  DOI: 10.12688/openreseurope.15692.1
- Policy Paper
  Open Source and Service Repository Policy
  DOI: 10.5281/zenodo.6757112



Current status of RDM @ GSI/ FAIR

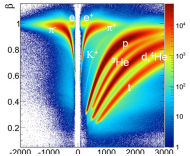
Andrew Mistry

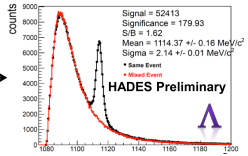
MT DMA ST1 Workshop

09.11.23

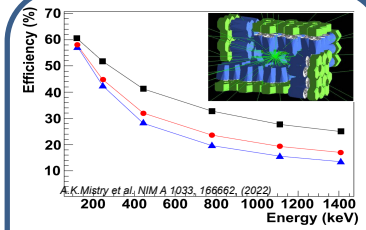
Research Data and Software at GSI: Rich and varied

Experimental Data:

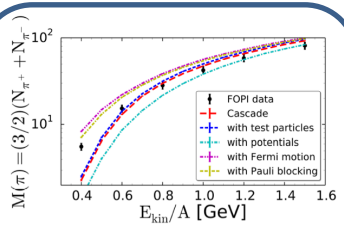




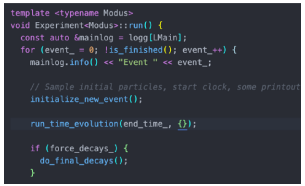
Raw *Pre-processed* *Result*



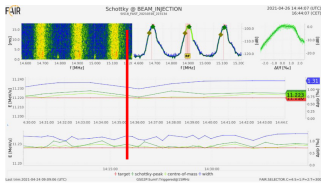
Simulation



Theory



Software



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

Who are we? Open Science working group @GSI/FAIR

Monthly meetings, started in April 2022

Endorsed by the GSI/FAIR management:Sept 2023

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. Provides the link through the hierarchy

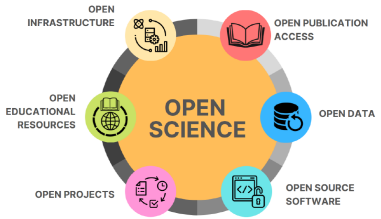
Expected Outcomes Implement OS policies and guidelines, best practices, test new tools and promotion/communication facility-wide

Email address: open-science@gsi.de

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

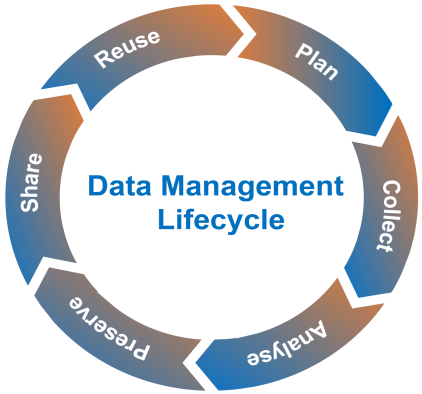
Open Science embodies the principle of making research output openly accessible and widely reusable through sustainable infrastructures, with as few barriers as possible. This can include open access publications, data, software, and hardware. Applying Open Science practices at GSI/FAIR will bring forth 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.



RDM @ GSI/FAIR: Goals and development

Goals:

- To ensure **good RDM practices** at GSI/FAIR and emphasize the benefits;
- Promote 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.



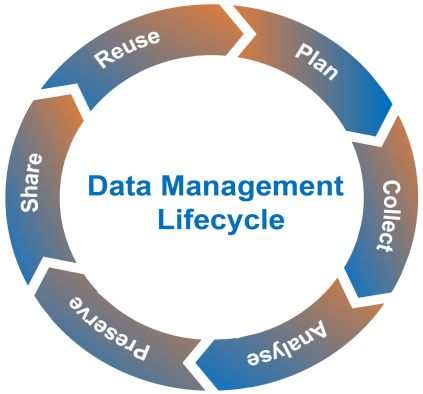
RDM @ GSI/FAIR: Goals and development

Goals:

- To ensure **good RDM practices** at GSI/FAIR and emphasize the benefits;
- Promote 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 + external projects;
- **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



RDM/Open Science workshops @ GSI: 2022+2023

RDM Slides and information: https://indico.gsi.de/event/14680/
Open Science Slides and information: https://indico.gsi.de/event/17498/

⇒ ***100+ participants*** per meeting in two hybrid events,

Variety of themes dicussed:

- **Inform and make researchers aware of RDM/Open Science, plans and progress**
- Talks from: **individual research groups, external projects, Tech Transfer, Helmholtz Open Science, Library, Grant office, IT department**
- **RDM questionnaire** distributed after workshop->identify areas to address

GSI/FAIR Research Data Management Policy/Guidelines

- **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
- ***Guidelines: Broader guide on RDM for researchers with examples***

RDM Policy: Published May 2023

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.

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

Table of Contents

Some tools for RDM currently at GSI

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

Compute Cluster Virgo

Electronic Logbooks (ELOG)

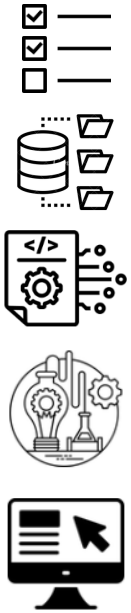
Collaboration Tools (Wiki, Indico, Mattermost, Seafile)

Code management system (GitLab)

Onboarding to the ESCAPE OSSR

GSI/FAIR publications repository (JOIN2)

Tools for RDM at GSI: Some *coming soon* examples



Data Management Planning Software (RDMO)

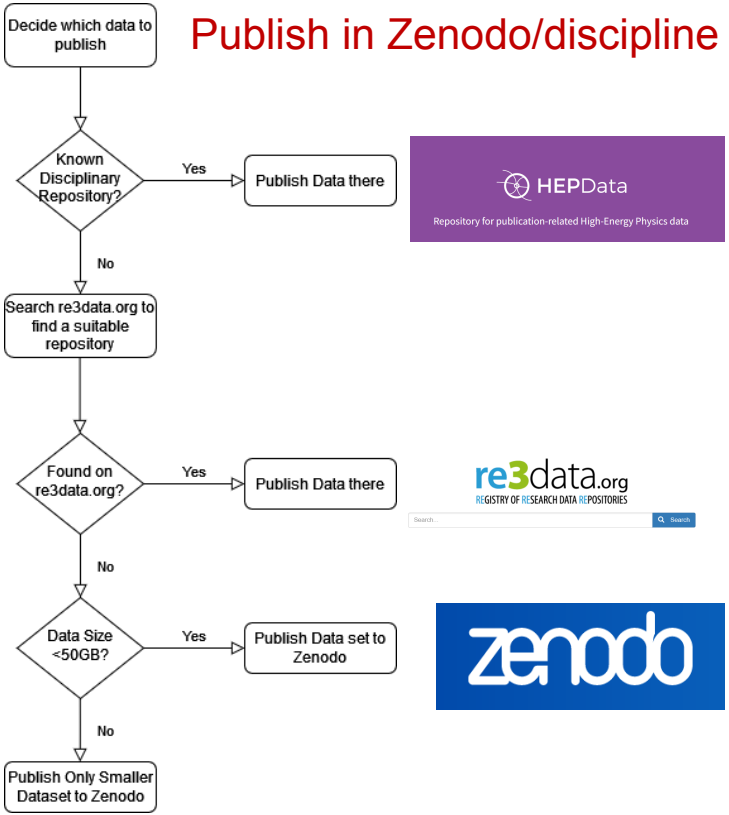
New GSI repository (MyCore-MiR)

Metadata schema (HELPMI, Nuclear Physics metadata)

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

Ecosystems + External e.g. HMC, PUNCH4NFDI, EOSC, EuroLabs…

Example: Data publication (interim solution)

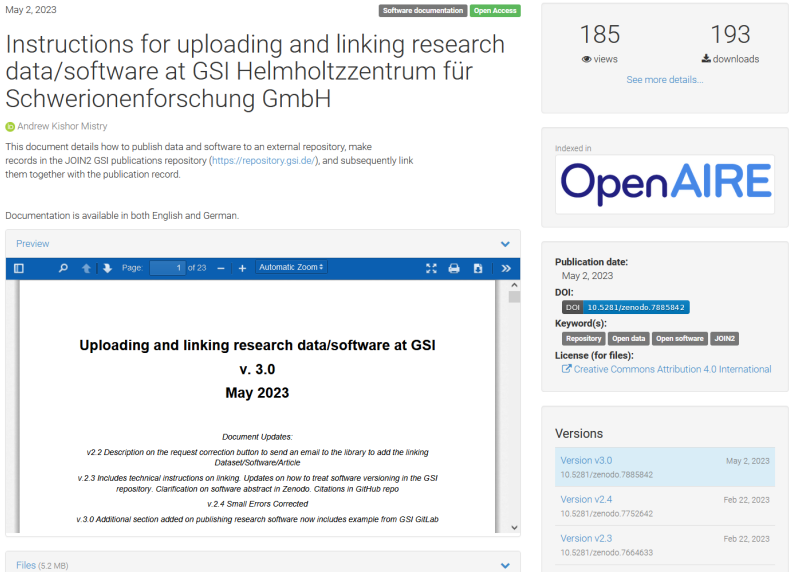


Publish in Zenodo/discipline specific repos

HEPData
Repository for publication-related High-Energy Physics data

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

zenodo



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

Instructions available on how to publish data at GSI

Summary

- Developments ongoing at GSI/FAIR to enable best RDM practices
- Use existing infrastructure plus new developments
- Data publishing: Interim solution, new repository in development
- Developing use cases within each group
- Dissemination of information: RDM team can assist!

Summary

- Developments ongoing at GSI/FAIR to enable best RDM practices
- Use existing infrastructure plus new developments
- Data publishing: Interim solution, new repository in development
- Developing use cases within each group
- Dissemination of information: RDM team can assist!

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

Summary

- Developments ongoing at GSI/FAIR to enable best RDM practices
- Use existing infrastructure plus new developments
- Data publishing: Interim solution, new repository in development
- Developing use cases within each group
- Dissemination of information: RDM team can assist!

Contact: open-science@gsi.de

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

Thanks for your attention!

Facility for Antiproton and Ion Research - "The Universe in the Laboratory"

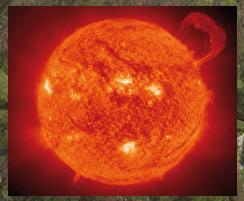
203x

Facility for Antiproton and Ion Research - "The Universe in the Laboratory"

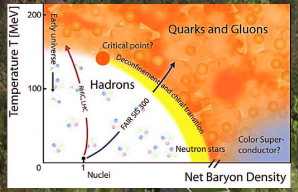
Cutting-edge science and technology

- ESFRI Landmark near Frankfurt, Germany (ESCAPE)
- Top priority for European Nuclear Physics Community
- International: 50 countries, 3000 researchers
- Diverse community from atomic to particle physics
- High intensity+precision+diversity+parallel operation
- Monolithic and modular experimental setups

APPA



CBM



panda

NUSTAR

+THEORY and BEAM physics 203x

Facility for Antiproton and Ion Research - “The Universe in the Laboratory”

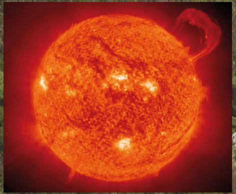
Cutting-edge science and technology

- ESFRI Landmark near Frankfurt, Germany (ESCAPE)
- Top priority for European Nuclear Physics Community
- International: 50 countries, 3000 researchers
- Diverse community from

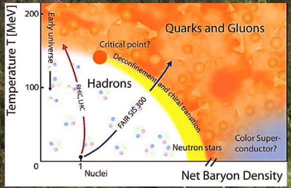
Towards the next generation “data challenge”

- High
- Intensity - precision, Velocity, Volume - Variety, Complexity!
- Multi-experiment, parallel operation
- ~TB/s data rates, online processing, ~5x105 cores
- Data stored on disk ~35 PB/year
- Distributed computing with a large user community

APPA

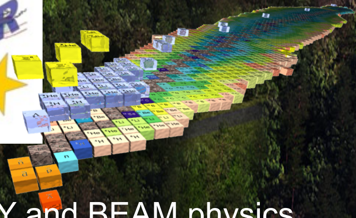


CBM



PANDA

NUSTAR



+THEORY and BEAM physics 203x