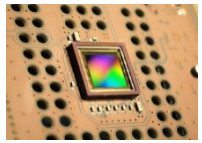
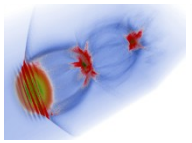
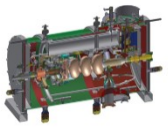
Matter and Technologies

MT-DMA ST1 status
The Matter Information Fabric

11th Annual MT Meeting

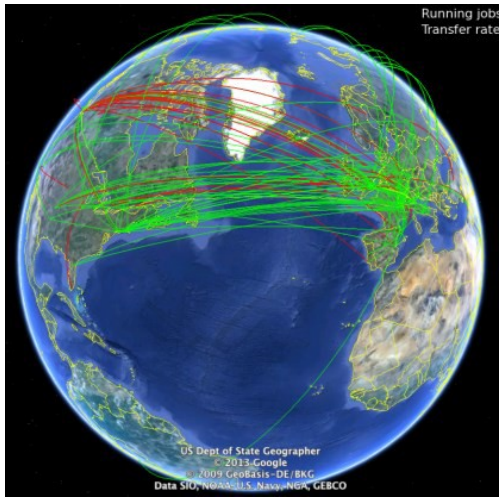
November 04, 2025, GSI

Kilian Schwarz

HELMHOLTZ ASSOCIATION

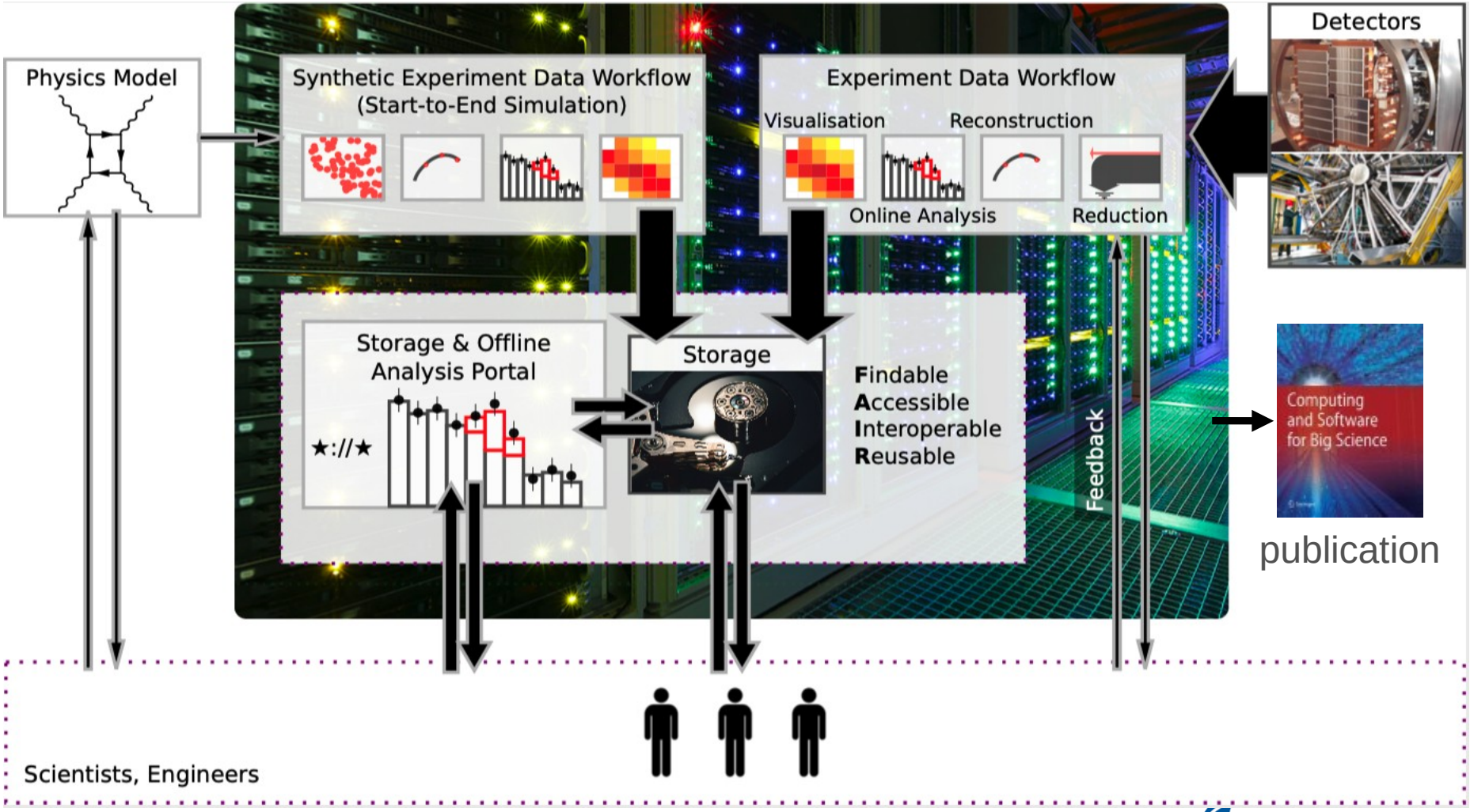
ST1 – the Matter Information Fabric
High Level Goals

- data analysis at large scale facilities
- solutions for the complete data lifecycle
  - modular, interoperable, reusable
  - scalable with increasing data rates and volumes
- data analysis platform
  - providing access to data at all levels
  - data transformation following FAIR principles



The WLCG connects analysis and simulation sites worldwide

full data lifecycle



DMA ST1 – current centres involved

- Current centers involved (in alphabetical order)
  - DESY (Photon, HEP, WIMP, accelerator, Lattice/theory)
  - GSI & HIJ (HEP, HI, Laser, accelerator)
  - FZJ (Photon, neutron, accelerator)
  - HZDR (Laser, photon, accelerator, ions)
- Plus HZB as observer (accelerator, photon, protons)

ST1 – bridges to other subtopics

- ST1 will address both technical and organisational aspects of integrated solutions
- technical interfaces to other STs have to be considered
  - on-site data reduction (ST2/3)
  - real-time online analysis (ST3)
  - new and efficient algorithms (ST2)
  - data flow models (ST3)

ST1 – how to achieve this ?

- analyse requirements from DMA communities
- identify possible synergies
- review of existing tools and tools to be developed
- gap analysis

➔ through workshops and direct engagement

- define a prototype of a data life cycle management system in a distributed computing environment
  - preferably fitting for most communities and centres in DMA
  - prototype should be modular, generic, open, portable, adaptable
  - Concentrating on interfaces

HELMHOLTZ ASSOCIATION

Milestones

- From the proposal … working towards achieving these

Milestone		Subtopic	Year
DMA-1	Definition of the structure and content of the S4M portal	ST1–3	2023
DMA-2	Launch of the S4M portal	ST1–3	2024
DMA-3	Online availability of all solutions provided by MT-DMA via S4M	ST1–3	2027
DMA-4	Organization of a workshop that defines and strengthens synergies in data lifecycle management among the participating facilities and communities	ST1	2022
DMA-5	Review and gap analysis of existing common tools for implementing a data lifecycle management system in a distributed computing environment that respects FAIR principles	ST1	2024
DMA-6	Review of and documentation of “lessons learned” from the implementation of a generic prototype of a data lifecycle management system in a distributed computing environment that respects FAIR principles	ST1	2027

Autumn 2023

Seminar Series
White Paper

HELMHOLTZ ASSOCIATION

DMA ST1 seminar series

- Review of existing tools for implementing a data life cycle system

ST1 – Whitepaper and infrastructure

- Describing status of centres and communities, requirements, gap analysis as well as a blueprint towards a common federated infrastructure including a concrete workplan towards a pilot system
- Milestone 5: review and gap analysis will be covered via white paper
- First integrated services for DMA ST1 infrastructure (HIFIS FTS/Rucio, SciCat Catalogue, ...) have been identified and tests will start

ST1 – Whitepaper and infrastructure

- Collaborative writing in collabtex
  - https://collabtex.helmholtz.cloud/project/65671b8f4a3cb804aec8b369

DMA-ST1 Whitepaper

Kilian Schwarz, Yves Kemp, Gerrit Günther, Oliver Knodel, Hans-Peter Schlenvoigt,
Christian Felder, Marina Ganeva, Martin Gasthuber, Thomas Gruber,
Radoslaw Karabowicz, Sonal Ramesh Patel, and Hubert Simmer
for the DMA-ST1 team

1 Introduction

The Helmholtz program *Matter and Technologies* (MT) was developed to meet technological challenges and current as well as future demands of the Helmholtz research field *Matter*. The MT program is built on the three pillars of *Accelerator Research and Development* (ARD), *Detector Technologies and Systems* (DTS) and *Data Management and Analysis* (DMA). The *DMA* topic facilitates challenge-driven and discovery science across all user communities on the unique data produced at Helmholtz research facilities, creates knowledge and fosters innovation. Within DMA, three focus areas are implemented, dedicated to sustainable data-driven science, algorithms for frontier technologies, and autonomous, intelligent systems for science. The objectives of DMA are reflected by a three-fold subtopic structure according to *Data*, *Algorithms*, and *Systems*.

Covering the subtopic *Data*, DMA-ST1 addresses the digitization of large-scale infrastructures at all levels, as seen in Figure 1.

Chapter 2&3 – centres and communities

2 Participating centers

3 Participating communities

Communities stretch across facilities sharing similar scientific interest and needs. The applied granularity for distinguishing communities is rather arbitrary and remain on a basic level here. The White Paper ErUM-Data TG Federated Infrastructures [KS23] has compiled a similar list of community requirements w.r.t. federated IT infrastructures for the ErUM communities: astro-particle physics, high-energy physics, accelerator physics, neutron science, photon science, ion

- Current centers and communities involved + MT-IDAF
  - DESY (Photon, HEP, WIMP, accelerator, Lattice/theory)
  - GSI & HIJ (HEP, HI, Laser, accelerator)
  - FZJ (Photon, neutron, accelerator)
  - HZDR (Laser, photon, accelerator, ions)
- Plus HZB as observer (accelerator, photon, protons)

As well as cross community activitities: DAPHNE4NFDI, PUNCH4NFDI, PaNOSC, LEAPS, ESCAPE

Chapter 4 – analysis of requirements

4 Analysis of requirements

The digitization of large-scale research facilities affects the entire data life cycle starting from experiment planning over data taking and analysis [MBB+19] until publication of scientific results. From a technical point of view, the data life cycle relies on an infrastructure of hard- and software on different levels ranging from local instrument control systems to facility-wide services. Extending this infrastructure across Helmholtz facilities fosters the data life cycle, enables synergies in data processing and improves F.A.I.R. data management in the Helmholtz Association, significantly. For this, DMA-ST1 conducted a survey on participating centers to shed some light on the maturity of data life cycle, current infrastructure, and technologies in use.


Questionnaire describing the tools used in the different steps of the data life cycle.

please fill in one questionnaire per site and per community.

Center & Community:			
Step	Tools	Competences	
Datataking / detector			
Start-To-End-Simulation			
Online processing and online data reduction			
Data storage			
Offline data analysis			
FAIR data handling, publication and archiving			
Step-overarching: Metadata handling & ELN			
Step-overarching: Used data formats, Data flow & automatisation			

HELMHOLTZ ASSOCIATION

Chapter 5 – KPI & user stories

5 Evaluation criteria

In order to be able to evaluate the federated computing infrastructure prototype, key performance indicators need to be identified. These predefined evaluation criteria shall serve as a guideline towards the milestone DMA-6 (review and documentation of "lessons learned" from the implementation of a generic prototype of a data life cycle management system in a distributed computing environment that respects F.A.I.R. principles). Typical key performance indicators which can be applied are the number of users, the numbers of participating centers and communities, the number of collaboratively used services, tools, and infrastructures and also feedback from users and communities. One of the main evaluation criteria, though, will be community provided user stories and to what extent these users stories will be supported by the provided infrastructure. The more user stories will be able to run successfully on this federated infrastructure the better the evaluation will be.

5.1 User Stories

5.1.1 User Data Processing

The unique instrumentation at Helmholtz facilities often produces complex data sets and relies on unique data processing workflows, ranging from adaptations over customized scripts to sophisticated, self-written computer programs. During user operation, a guest from the same or another facility (even from outside Helmholtz) get used to such an instrument-specific workflow which allows to create reliable and technique-relevant results. Once leaving the instrument, the user looses access to the data as well as the data processing workflow which can be an obstacle in publishing scientific results, often leaving the scientist to deal with infrastructure problems.

The federated infrastructure should provide remote access to instrument data as well as processing workflows. Since the access to the instrument is often limited due to security reasons, the entire data processing workflow could be moved to a separate unit which is used by the current user of the instrument (during measurement) as well as past users processing former experiment data. It would be advantageous to restrict on a single workflow for all users to avoid additional maintenance, realized by a central service or container technology.

HELMHOLTZ ASSOCIATION

Chapter 6 – Blueprint of the infrastructure

See also presentation of H-P Schlenvoigt today in this session
and presentation of C. Schneide yesterday in plenary

A first idea describing the blueprint of a pilot infrastructure can be seen in Figure 9. In the initial setup phase a prototype of the DMA-ST1 federated compute infrastructure, running on the participating DMA-ST1 centres, will be provided along with a detailed documentation of the involved software and tools. The documentation will contain information on how the different parts can be assembled into a common federated infrastructure based on multiple community resources providing a common DMA-wide Science Cloud. In this stage, resources of other communities and centres can be used in an opportunistic way to, eventually, allow more efficient use of provided resources.

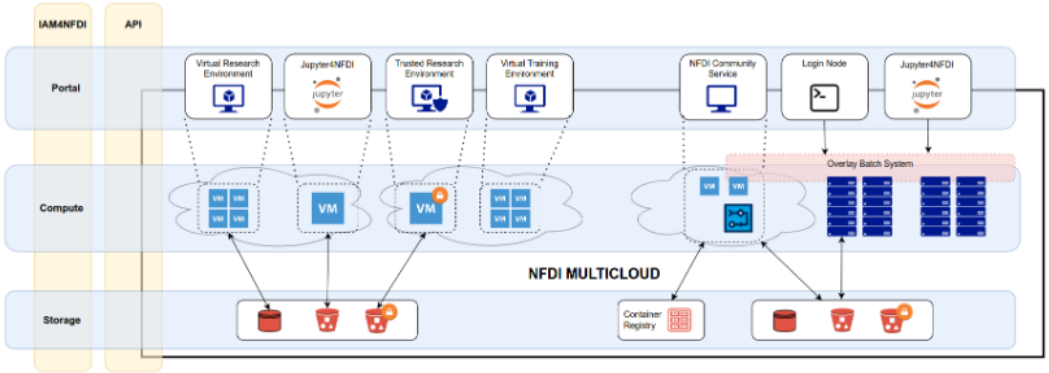


Figure 9: Federated infrastructure consisting of federated portal, federated compute and federated storage layer.

PUNCH4NFDI provides a prototype environment for federated computing and storage which can be used as a starting point. The distributed resources are combined via an overlay batch system (COBalD/TARDIS [15]) which provides interfaces to cloud, HPC, and HTC infrastructures. Access to these compute resources is provided using established technologies like JupyterHub, or traditional login nodes, and also data portals as VISA or the PUNCH Science Data Platform are

Services and testbed are done in close collaboration with PUNCH4NFDI.
The idea is to set up an infrastructure based to a large extent on existing services.
DMA can contribute to the sustainability of the PUNCH4NFDI infrastructure.

Chapter 7 – Gap Analysis

Based on this gap analysis open tasks for POF V are identified

7 Gap analysis

In the Helmholtz Association, an improvement of the data life cycle requires basic measures on an administrative, technical, and semantic level. Some software solutions are still missing or in development which could fill gaps, such as a data management plan (DMP) establishing a connection between proposal, experiment, data, and publications.

In general the idea is to provide a prototype infrastructure based on existing tools and services, which are already in use or in development by the participating centres and communities. This shall be compared the the requirements analysis (section 4) and existing gaps must then be identified. The identified gaps should be filled by required additional development work in order to be able to provide a complete infrastructure which is able to fulfill the requirements of the centres and communities. A draft of the intended pilot infrastructure including the identified tools, serrvices and (sub-) systems is shown in Figure 10.

Open Gaps which still require research: common set of metadata and schemata, federated metadata catalogue infrastructure, web portals as user interface, workflow languages and interactivity, REANA and Kubernetes, resource usage monitor, software policies, federated proposal system, collaborative Jupyter Nbs, Search APIs, ...

Chapter 8 – Work Plan towards pilot system

See also presentation of H-P Schlenvoigt today

- Identification and setup of initial services
- Location of pilot services (MT-IDAF and participating centres)
- Pilot resources
- Initial test cases

Pilot resources:

- FZJ:
  - compute/storage resources
  - local SciCat instance
  - Visa portal ?
- HZDR:
  - compute/storage resources
  - metadata portal with search API
  - Knowledge graph
- GSI:
  - compute/storage resources
  - container/CVMFS services
- DESY:
  - compute/storage resources
  - central SciCat instance
  - central Rucio instance (HIFIS)
- HZB:
  - metadata portal with search API
  - ICAT endpoint
  - HELPME
- ILDG:
  - metadata catalogue
  - alternative workflows w/o Rucio

This list is based on initial discussion. It is an expression of will and no commitment. The list is not complete.

Chapter 9 – Sustainability aspects

9 Sustainability aspects

Although not anchored in the initial topics of DMA, concern about sustainability in all areas of computing is rising and is becoming a critical topic that has become necessary to address. While this topic is important in all areas of research, the environmental impact of IT alone is expected to triple over the next few years. A lot of this will be due to the increasing of use of data-centres for AI/ML training and deployment [And20, IEA25, Sop25]. Consequently sustainability will also play a more central role in the upcoming funding period.

Analysing the aspects of sustainability pertaining to the design and operation of data-centres will be crucial in both minimising a research centre's impact on the environment, but also helpful for determining policy to make data-centres in general run more sustainably. Some of the main concerns include but are not limited to: the amount of power the technology running in the data-centre draws while in operation, and the attributable cooling requirements required for said operation; the effectiveness of software run on servers; the computing workflows of experiments that the data-centre services; the data access model of said experiments; the efficiency of the infrastructure housing the data-centre; the ability to be flexible to variable energy-market conditions; and the embedded carbon coming into the data-centre via new hardware.

Also here tasks for POF V have been identified:
In general: minimising CO2 equivalent output in scientific computing; fostering cultural change
- procurement guidelines including end of life policy
- green energy depending load management
- blue print for sustainable data centres
- white paper describing change of culture

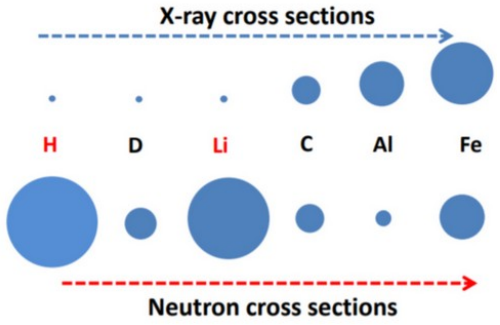
HELMHOLTZ ASSOCIATION

Chapter 10 – Summary and Outview

Things to do:
- Summary and Outview
- bring chapter 8 into plain text
- how different communities deal with open data
- cost estimates
- add community overarching use cases:
  example use case:

neutrons and x-ray photons, due to their different ways of interaction to matter, deliver a complimentary information about properties of the materials.
Thus, as a natural consequence, a lot of scientific groups employ both techniques to study their samples. As an example use case can serve neutron and x-ray diffraction in batteries (see figure below): neutrons provide information about light elements, while x-ray photons have higher scattering cross sections for heavier elements and only both techniques together provide comprehensive structural information about battery under investigation.

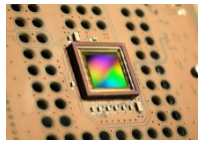
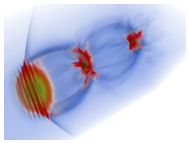
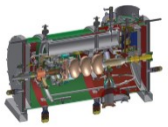
Follow the blue print and build and test the described architecture !!!



HELMHOLTZ ASSOCIATION

ST1 – outlook onto POF-V

- Timeline: 2028-2035
- ST1 is glue between LK1 and LK2 (IDAF)
- Sustainability of solutions
- Concentrating on interfaces between existing approaches at communities/centres

MT-DMA ST1 status
The Matter Information Fabric

11th Annual MT Meeting

November 04, 2025, GSI

Kilian Schwarz

questions/comments ?