Summary of centre questionaires and seminar series

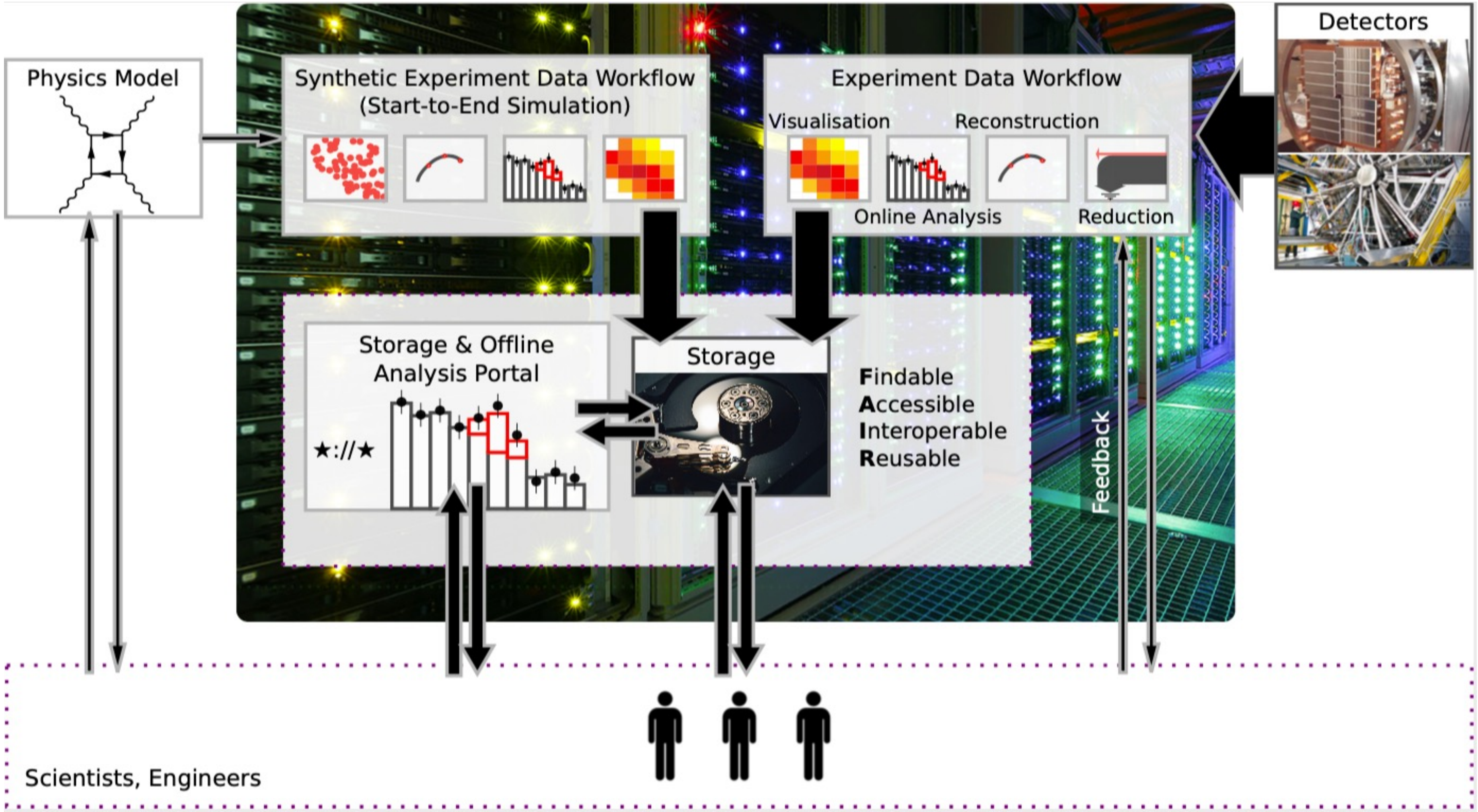
Kilian Schwarz, Yves Kemp

DMA ST1 synergy workshop
DESY 8.11.2023

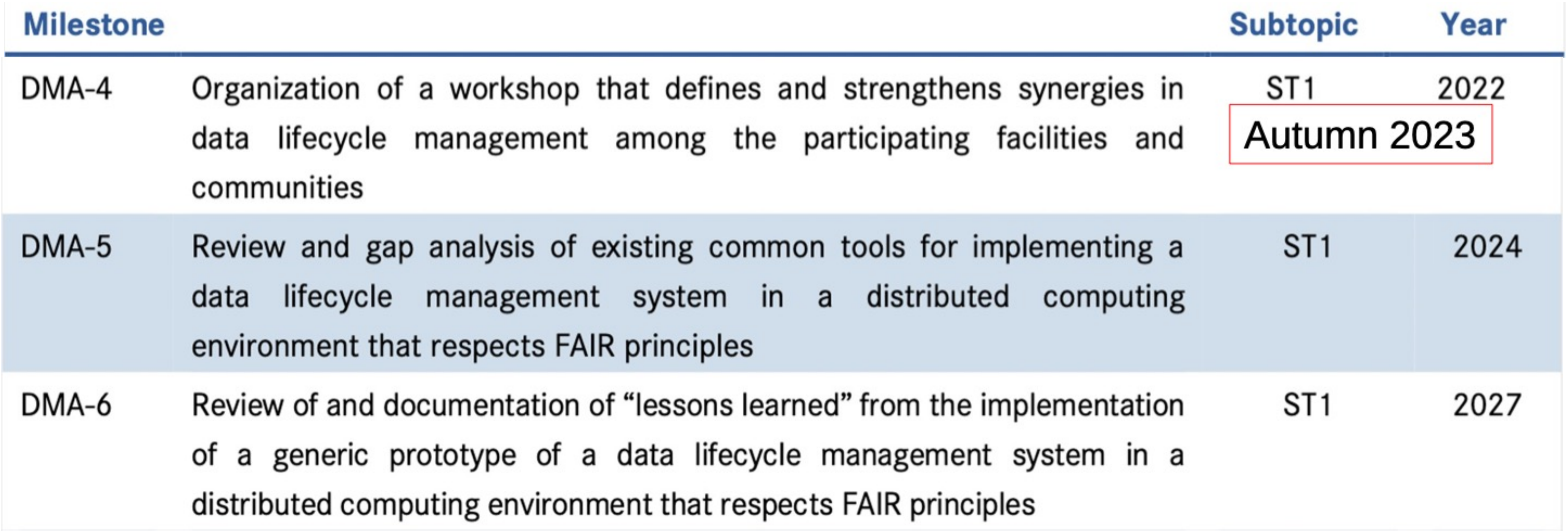
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

What is ST1 about?
… the full data lifecycle: workflows, infrastructure, tools



DMA ST1 milestones



Milestone		Subtopic	Year
DMA-4	Organization of a workshop that defines and strengthens synergies in data lifecycle management among the participating facilities and communities	ST1	2022 Autumn 2023
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

Analysis of full data lifecycle

- Identified several steps:
  - Actions on raw data (reduction, reconstruction, visualization, transfer and storage, …)
  - Actions on simulated data (reconstruction, transfer and storage, …)
  - Offline analysis (data discovery, accessibility, analysis, reusability, …)
- Identified several actors:
  - Centers, facilities
  - Individual scientists, experiments, collaborations, communities
- Identified possible tools

Preparation for the first milestone: Two-fold

- Learn what tools are available, what these tools can offer, …

  → Seminar series

- Learn what tools communities use, want, are missing, …

  → Questionaire

Seminar series: In general:

- Out of the previous items, have created seminar series. So far:
- Storage and data management solutions
- Meta-Data handling
- Log-Books
- Data acquisition
- Call for volunteers for further seminars!

- Opened seminar beyond DMA/ST1
- → Announced to all of DMA, well attended

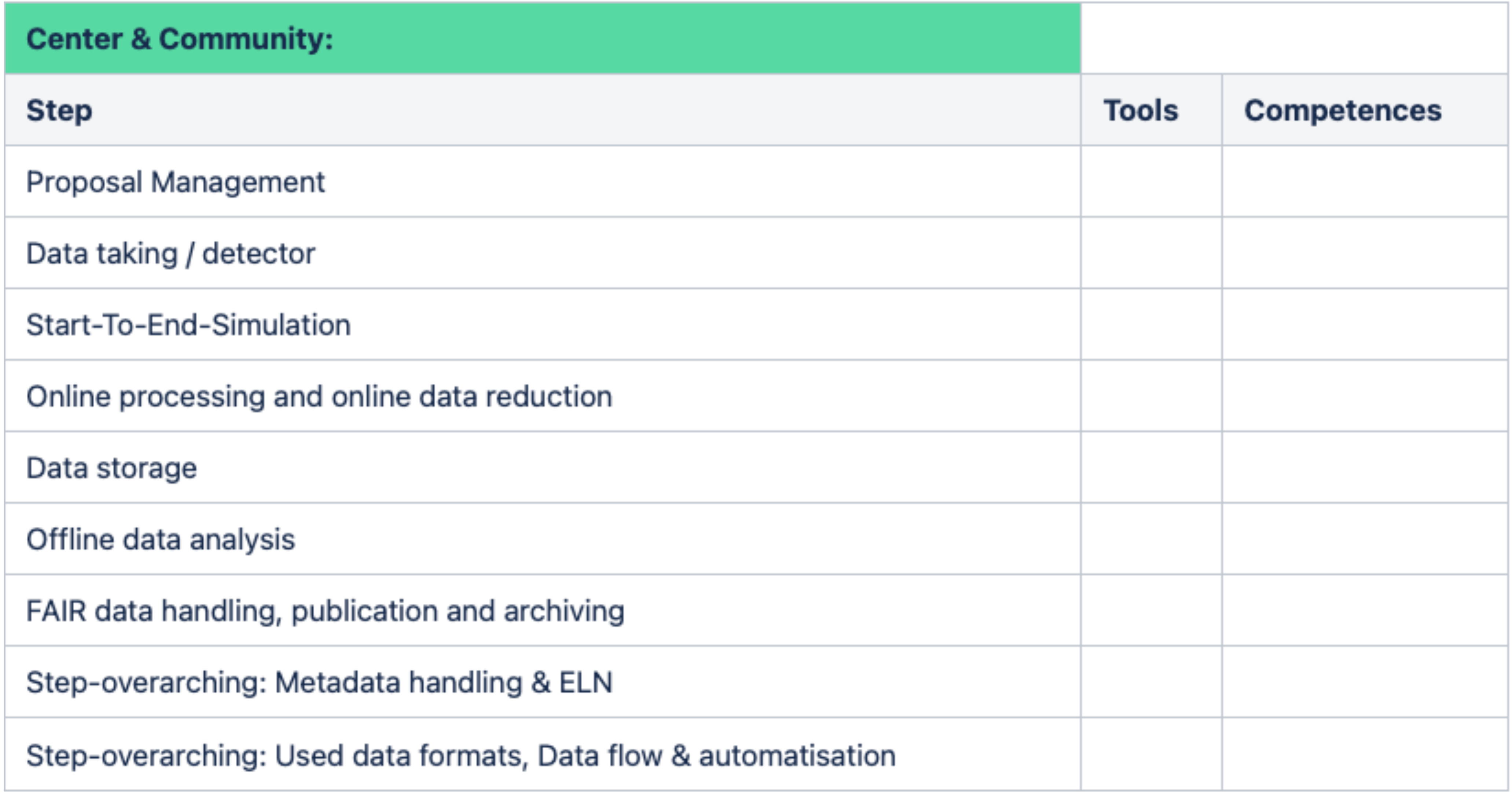
(some) topics discussed for seminar

Need to chose topics well:

- Not too basic, fabric topics
  - e.g. Linux distros, cluster file systems
- Not ”too much physics” or community specific
  - e.g. monte carlo simulators
- Related to full data lifecycle
  - Online analysis, reconstruction, data reduction, data movement
  - Offline analysis, portals
  - Metadata and logbook,

We want your feedback!

Questionaire: Idea: Get input from ST1 sites about their communities



Center & Community:		
Step	**Tools**	**Competences**
Proposal Management		
Data taking / 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		

Answers:

- DESY
  - HEP
  - PhotonScience
- HI Jena
  - HI Laser
- HZB
  - Photon Science

- HZDR
  - Photon Science
  - THz Science
- FZJ
  - Neutron
  - Photon Science
- GSI:
  - HEP / CBM / Panda

Thanks everyone who participated in filling out the questionaires!

- We know:
  - Not all relevant centers
  - Not all relevant communities
  - … but a start

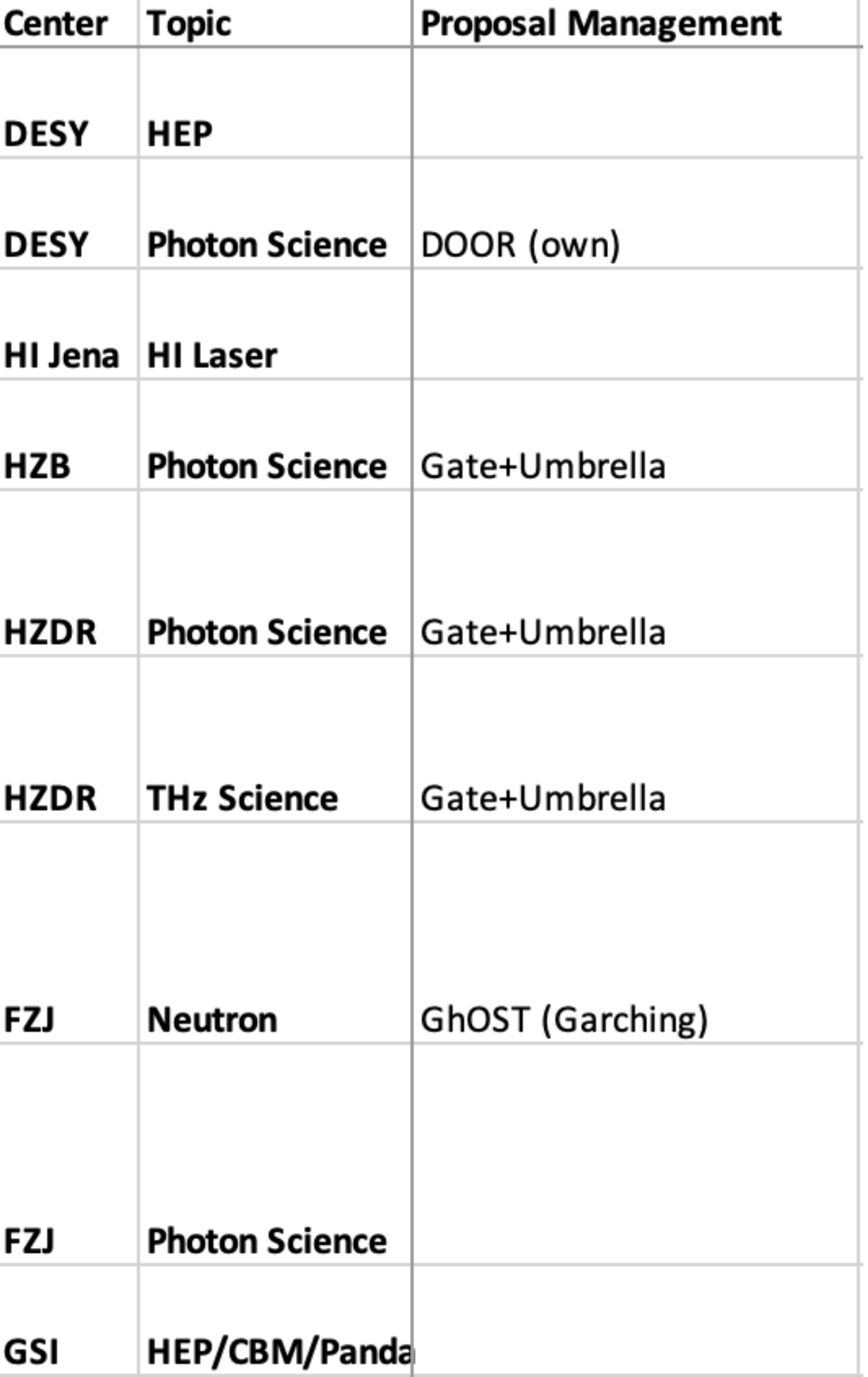
Preparatory workshop June 2023 @ HZDR

- Finalizing & discussing questionaires
- summarizing them

→ The following slides will show a summary,

→ question by question

→ base for discussion

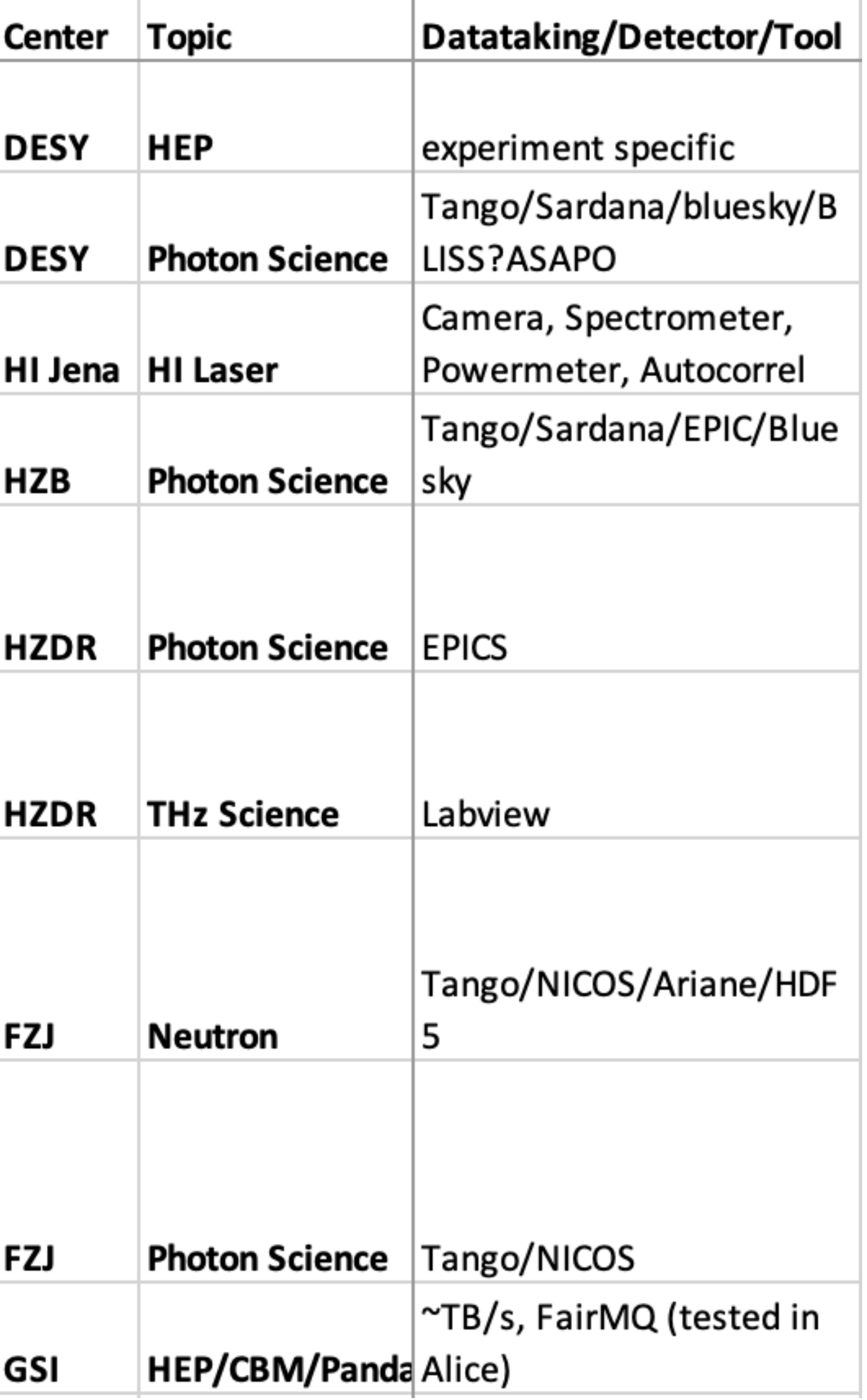


Center	Topic	Proposal Management
DESY	HEP	
DESY	Photon Science	DOOR (own)
HI Jena	HI Laser	
HZB	Photon Science	Gate+Umbrella
HZDR	Photon Science	Gate+Umbrella
HZDR	THz Science	Gate+Umbrella
FZJ	Neutron	GhOST (Garching)
FZJ	Photon Science	
GSI	HEP/CBM/Panda	

Proposal management

- Gate users already connected
- Where ST1 can help:
  - Offer ST1 communication platform on proposal management systems
- Essential for an ST1 federated infrastructure
  - proposal management metadata accessible for integration with other steps of lifecycle

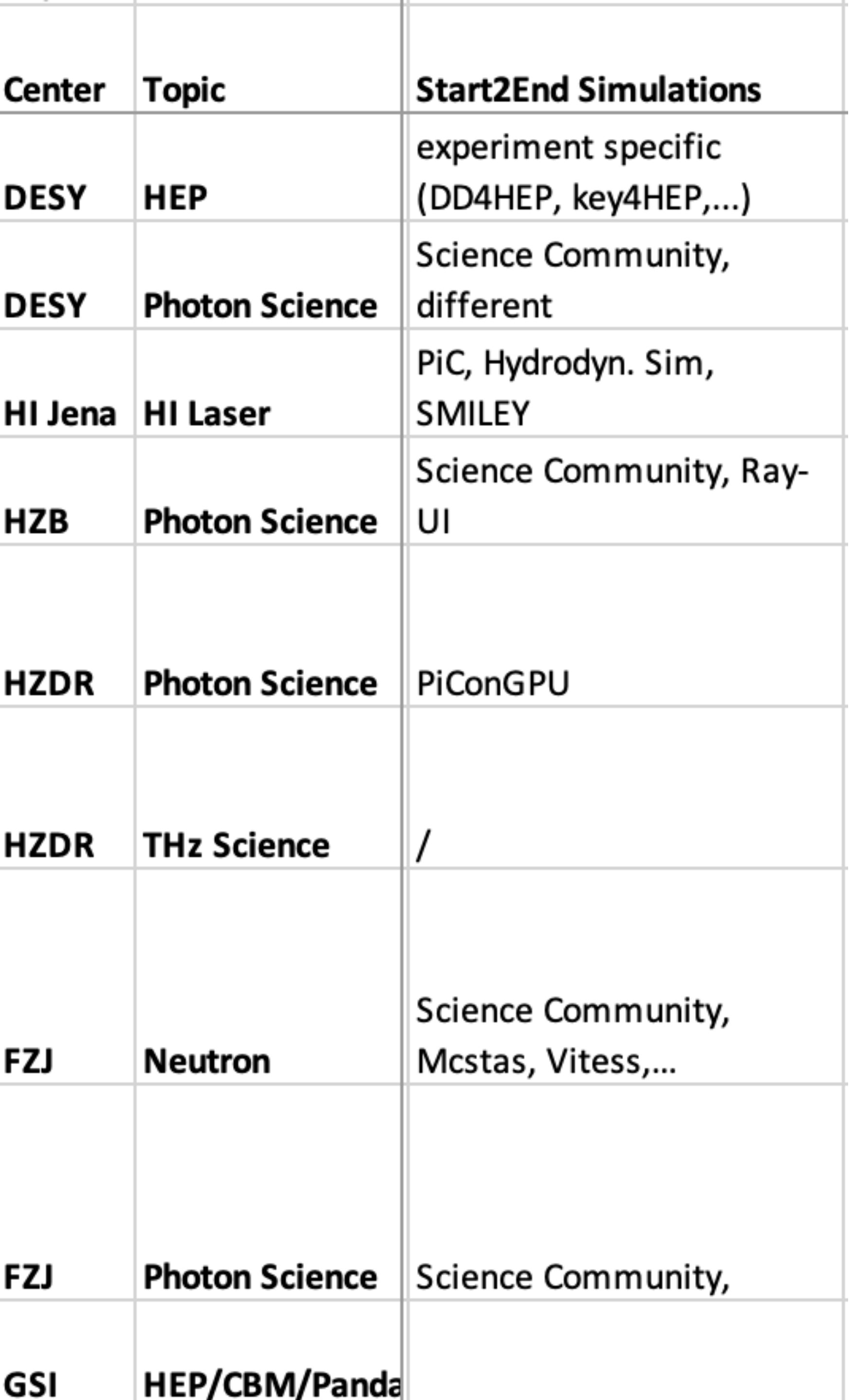
Datataking, detectors



Center	Topic	Datataking/Detector/Tool
DESY	HEP	experiment specific
DESY	Photon Science	Tango/Sardana/bluesky/BLISS?ASAPO
HI Jena	HI Laser	Camera, Spectrometer, Powermeter, Autocorrel
HZB	Photon Science	Tango/Sardana/EPIC/Bluesky
HZDR	Photon Science	EPICS
HZDR	THz Science	Labview
FZJ	Neutron	Tango/NICOS/Ariane/HDF5
FZJ	Photon Science	Tango/NICOS
GSI	HEP/CBM/Panda	~TB/s, FairMQ (tested in Alice)

- Common points:
  - Data in-memory-handling
  - HDF5 / Nexus
  - bluesky / tango /Epics / NICOS
- Where ST1 can help:
  - Offer ST1 communication platform
  - especially in-memory handling // future of labview // control systems and cross-control-systems-communities
- Essential for an ST1 federated infrastructure
  - Common file & metadata formats

start-to-end simlations

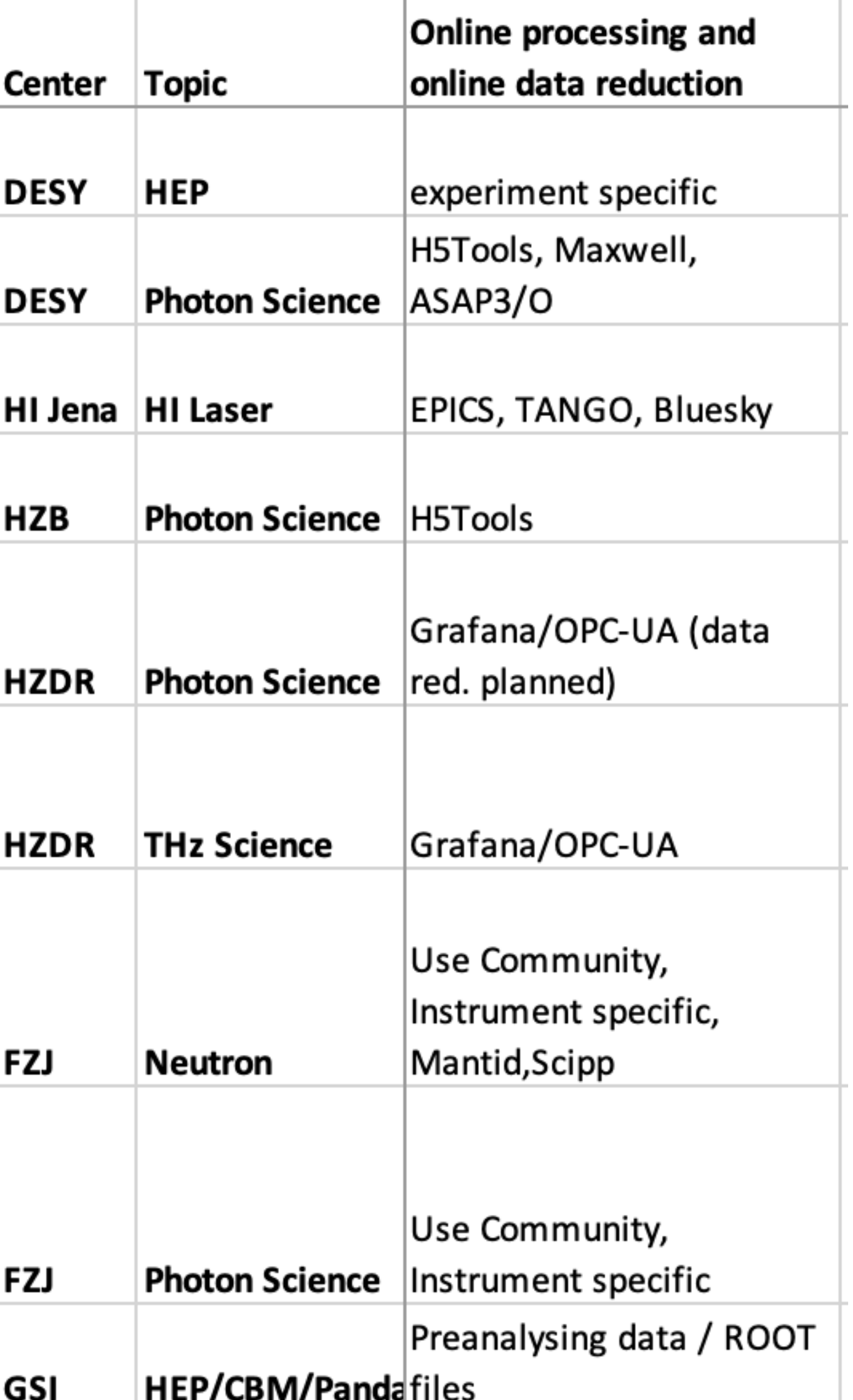


Center	Topic	Start2End Simulations
DESY	HEP	experiment specific (DD4HEP, key4HEP,...)
DESY	Photon Science	Science Community, different
HI Jena	HI Laser	PiC, Hydrodyn. Sim, SMILEY
HZB	Photon Science	Science Community, Ray-UI
HZDR	Photon Science	PiConGPU
HZDR	THz Science	/
FZJ	Neutron	Science Community, Mcstas, Vitess,...
FZJ	Photon Science	Science Community,
GSI	HEP/CBM/Panda	

- no obvious gap
- more an application than infrastructure itself
- surrogate modeling (as an alternative to expensive simulations)
  - → Interface for ST2 & ST3 + Helmholtz AI
- Essential for an ST1 federated infrastructure
  - common exchange formats, metadata formats
  - Metadata input/output standardized for Chaining applications together

→ data processing pipeline / workflow

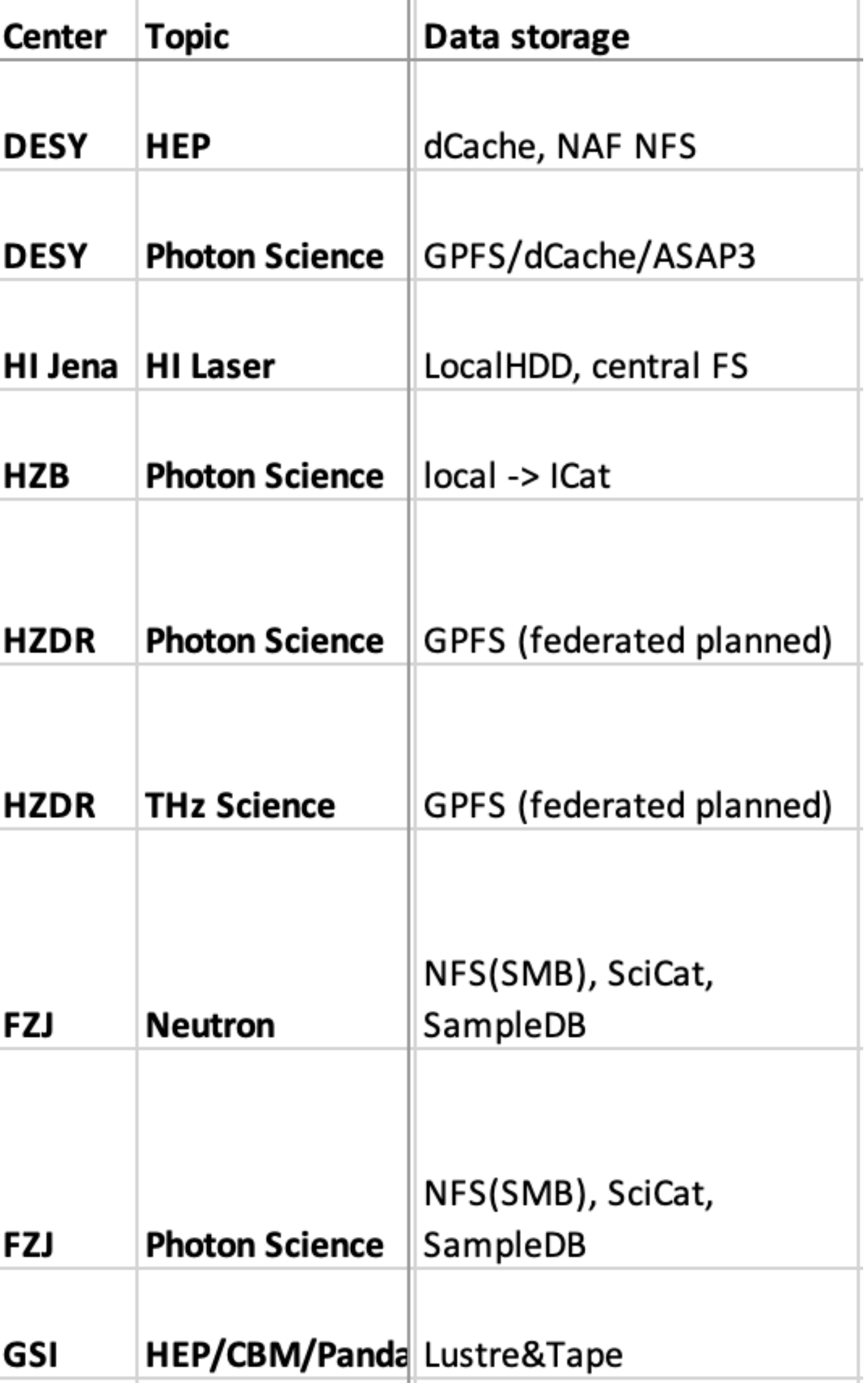
Online processing and data reduction



Center	Topic	Online processing and online data reduction
DESY	HEP	experiment specific
DESY	Photon Science	H5Tools, Maxwell, ASAP3/O
HI Jena	HI Laser	EPICS, TANGO, Bluesky
HZB	Photon Science	H5Tools
HZDR	Photon Science	Grafana/OPC-UA (data red. planned)
HZDR	THz Science	Grafana/OPC-UA
FZJ	Neutron	Use Community, Instrument specific, Mantid,Scipp
FZJ	Photon Science	Use Community, Instrument specific
GSI	HEP/CBM/Panda	Preanalysing data / ROOT files

- Discussion on significance of online data analysis
- Common points:
  - data reduction
- Communication:
  - ”how fast should fast be” (experiment specific) → ST2
  - data reduction (experiment specific) → ST2
  - (Offer ST1 communication platform)
- Essential for an ST1 federated infrastructure
  - Monitoring
  - data & metadata formats
  - dynamic archiving

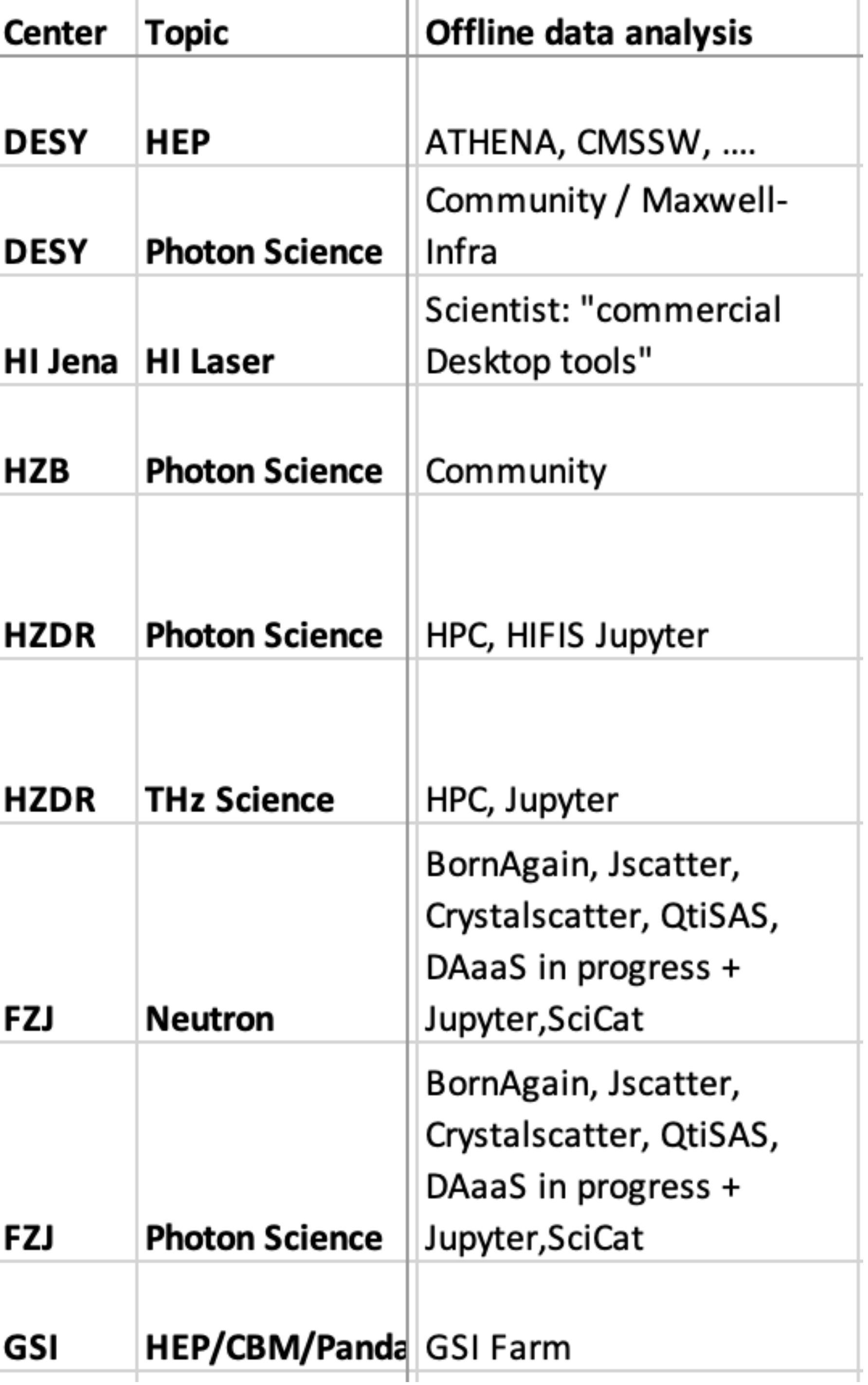
Data storage & data exchange



Center	Topic	Data storage
DESY	HEP	dCache, NAF NFS
DESY	Photon Science	GPFS/dCache/ASAP3
HI Jena	HI Laser	LocalHDD, central FS
HZB	Photon Science	local -> ICat
HZDR	Photon Science	GPFS (federated planned)
HZDR	THz Science	GPFS (federated planned)
FZJ	Neutron	NFS(SMB), SciCat, SampleDB
FZJ	Photon Science	NFS(SMB), SciCat, SampleDB
GSI	HEP/CBM/Panda	Lustre&Tape

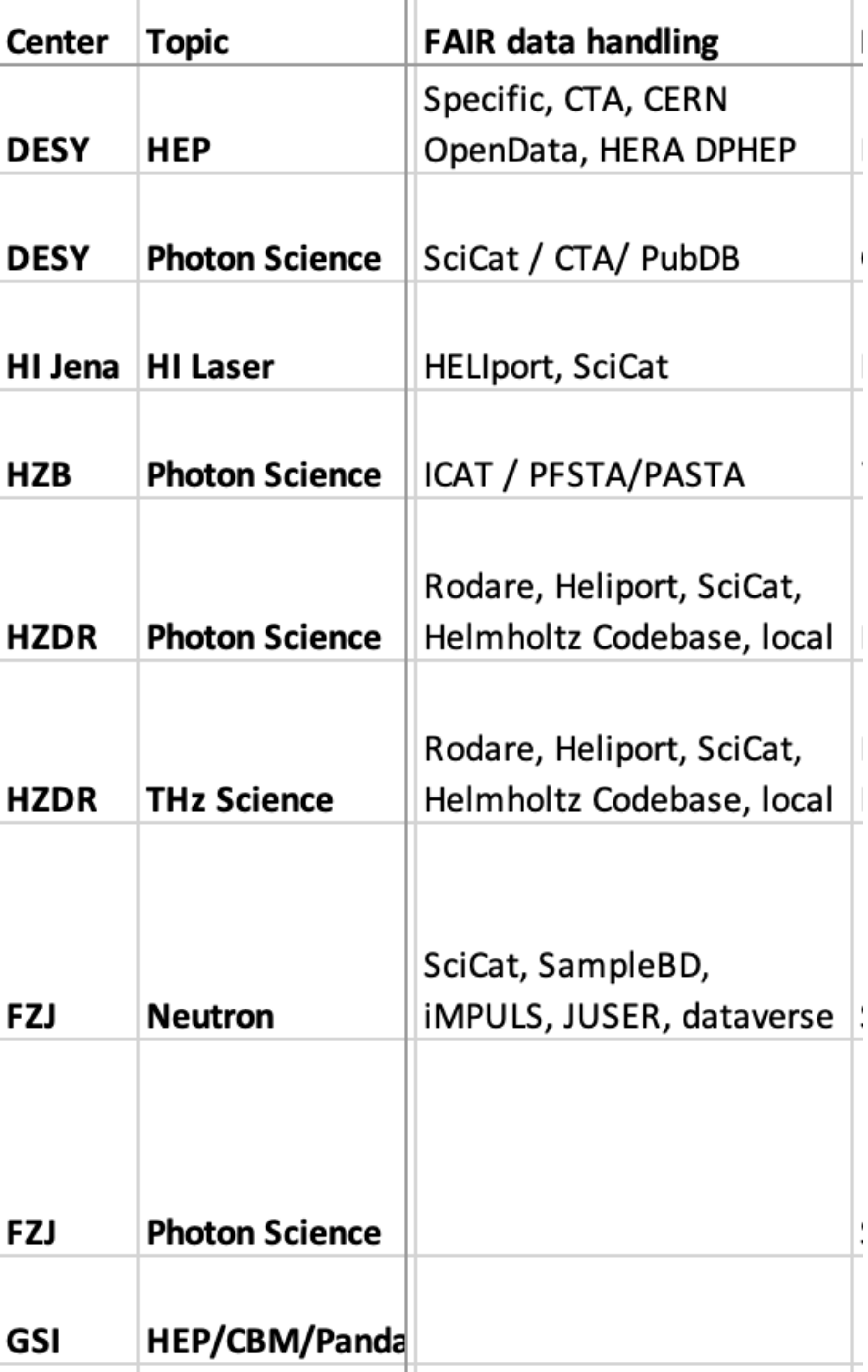
- Common points:
  - All (locally) rely on POSIX (network) filesystems – with their strengths and limitations
- Essential for an ST1 federated infrastructure
  - data exchange: technical & (meta)-data formats
  - define minimal content of data policies → HMC
- Work-item:
  - Setup data lake, e.g. join PUNCH and/or DAPHNE ?

Offline data analysis



Center	Topic	Offline data analysis
DESY	HEP	ATHENA, CMSSW, ….
DESY	Photon Science	Community / Maxwell-Infra
HI Jena	HI Laser	Scientist: "commercial Desktop tools"
HZB	Photon Science	Community
HZDR	Photon Science	HPC, HIFIS Jupyter
HZDR	THz Science	HPC, Jupyter
FZJ	Neutron	BornAgain, Jscatter, Crystalscatter, QtiSAS, DAaaS in progress + Jupyter,SciCat
FZJ	Photon Science	BornAgain, Jscatter, Crystalscatter, QtiSAS, DAaaS in progress + Jupyter,SciCat
GSI	HEP/CBM/Panda	GSI Farm

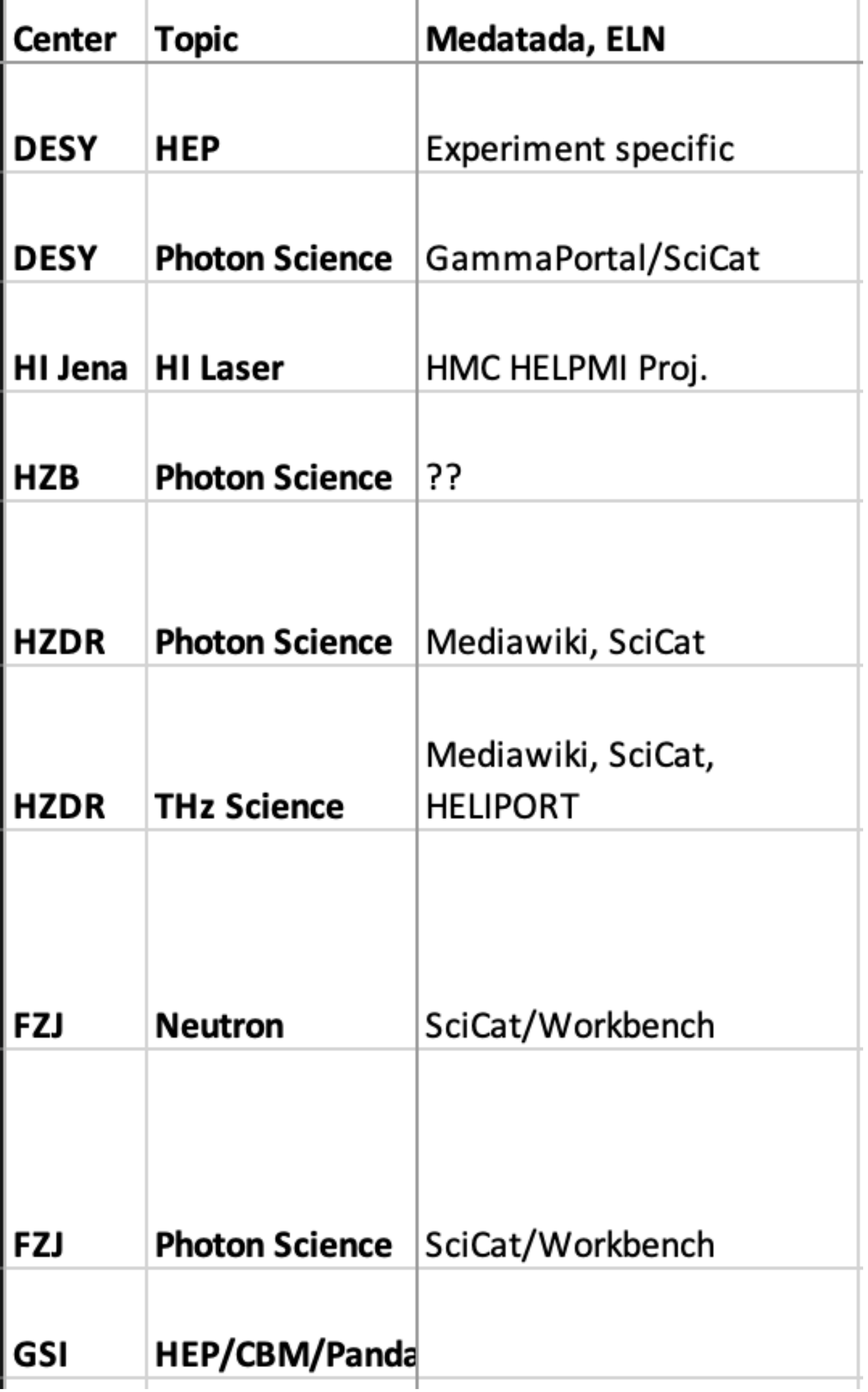
- Common points:
  - Jupyter
  - Wish list: Sharing interactive access to Jupyter notebook … maybe via screen-sharing?
- Essential for an ST1 federated infrastructure
  - Jupyter + tools (e.g. combine with Base4NFDI Jupyter service / FZJ lead)
  - Notebooks in Gitlab (up to users/facilities to enforce)
- Work-item:
  - Distributed computing (e.g. integration, see previous slide)



Center	Topic	FAIR data handling
DESY	HEP	Specific, CTA, CERN OpenData, HERA DPHEP
DESY	Photon Science	SciCat / CTA/ PubDB
HI Jena	HI Laser	HELIport, SciCat
HZB	Photon Science	ICAT / PFSTA/PASTA
HZDR	Photon Science	Rodare, Heliport, SciCat, Helmholtz Codebase, local
HZDR	THz Science	Rodare, Heliport, SciCat, Helmholtz Codebase, local
FZJ	Neutron	SciCat, SampleBD, iMPULS, JUSER, dataverse
FZJ	Photon Science	
GSI	HEP/CBM/Panda	

FAIR data handling

- Common points:
  - SciCat
  - Heliport
- Communication platform to Zenodo community
- Work-item:
  - Interconnecting SciCat installations, e.g. connect to b2find (maybe small development, HZDR)
  - Provide DMA test SciCat instance
  - Potential MongoDB licensing issues, alternatives?
  - Investigation: Model ST1 Lifecycle in Heliport … Heliport test instance will not make much sense
- Essential for an ST1 federated infrastructure
  - data repositories should be able to report Metadata to b2find
  - Discipline specific repos can be found using re3data

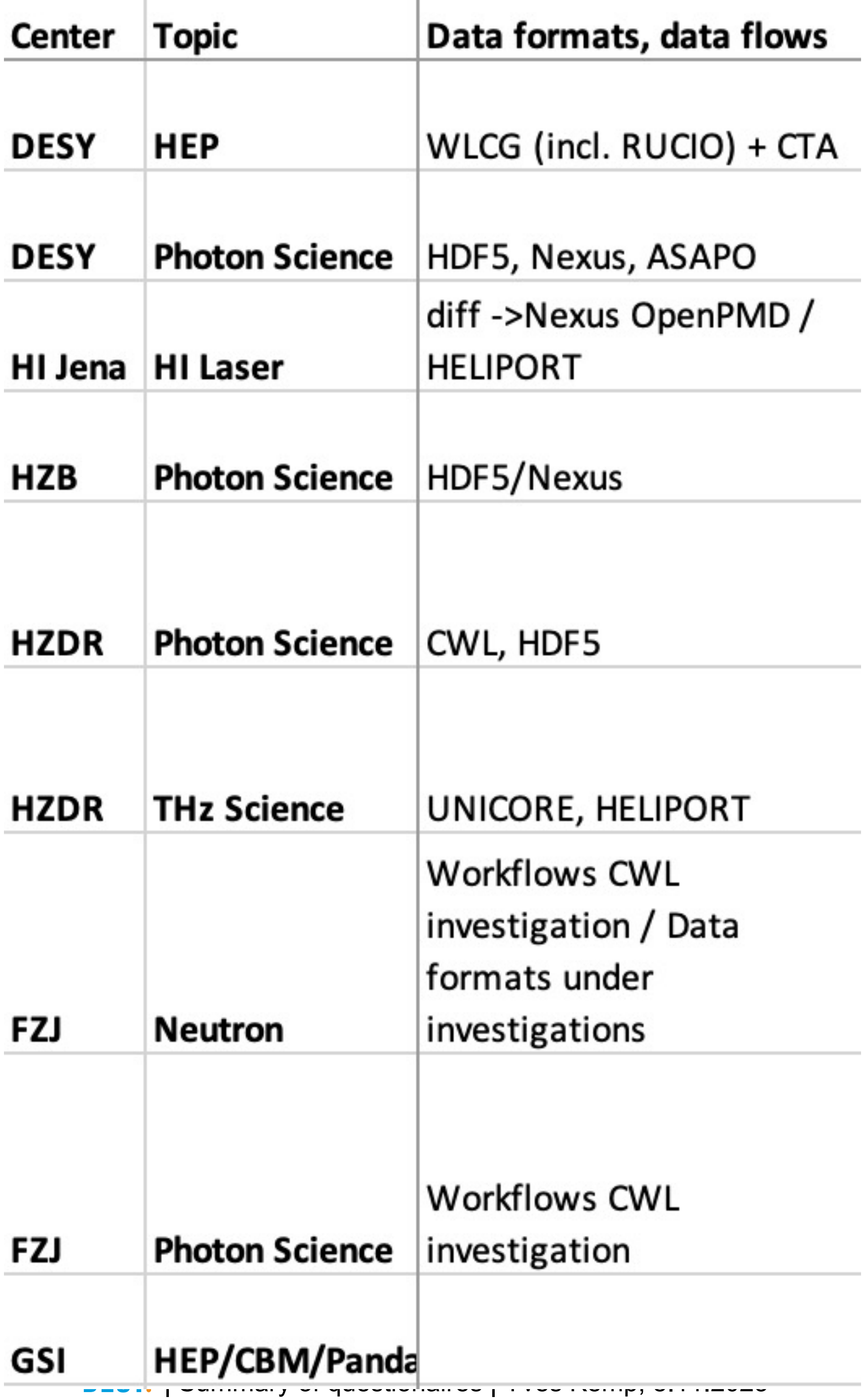


Center	Topic	Medatada, ELN
DESY	HEP	Experiment specific
DESY	Photon Science	GammaPortal/SciCat
HI Jena	HI Laser	HMC HELPMI Proj.
HZB	Photon Science	??
HZDR	Photon Science	Mediawiki, SciCat
HZDR	THz Science	Mediawiki, SciCat, HELIPORT
FZJ	Neutron	SciCat/Workbench
FZJ	Photon Science	SciCat/Workbench
GSI	HEP/CBM/Panda	

Metadata, ELN

- Common points:
  - SciCat
- Choice of ELN up to sites
  - offer ST1 communication channels
- Essential for an ST1 federated infrastructure
  - ELN standarized exchange formats (e.g. XML, .eln, …) and API
  - observe development (e.g. DAPHNE, ROCK-IT)

Data formats, data flow, automatisation



Center	Topic	Data formats, data flows
DESY	HEP	WLCG (incl. RUCIO) + CTA
DESY	Photon Science	HDF5, Nexus, ASAPO
HI Jena	HI Laser	diff ->Nexus OpenPMD / HELIPORT
HZB	Photon Science	HDF5/Nexus
HZDR	Photon Science	CWL, HDF5
HZDR	THz Science	UNICORE, HELIPORT
FZJ	Neutron	Workflows CWL investigation / Data formats under investigations
FZJ	Photon Science	Workflows CWL investigation
GSI	HEP/CBM/Panda	

- HDF5
- RUCIO → Dev. team presented in seminar

- ST1 Communiction platform on file formats
- Essential for an ST1 federated infrastructure
  - File formats should be standardized and documented
- Work-Item: Setup RUCIO test installation (DESY,GSI,HZDR usage interest)

Other observations _ 1

Items that came up and did not fit any category

- User support is mandatory! → Best-effort & self-support via Mattermost
- Globus endpoints / RUCIO / FTS …how to organize large data transfers?

  → RUCIO see other slide

  → large data transfers: try out HIFIS FTS Service (later stage: Heliport integration)

  (general comment: avoid data transfers, e.g. data aware meta-scheduling)
- DESY: connect to WIMP community (ALPS, ...)
- Community: Control of experimental components → Mattermost community / ST1 communication
- User management, Authentication, Authorisation, … ownership of data …

  → AAI

  → Mattermost community for config & conceptual exchange

  → e.g. DAPHNE&PUNCH policy defining

Other observations _ 2

Items that came up and did not fit any category

- Quota management, user data migration to tape → Mattermost community
- Strengthen PUNCH-DAPHNE-DMA interplay … have interoperable view

  → Invite DAPHNE/PUNCH/ROCK-IT to large workshop → Welcome ☺
- Portals: Investigate on VISA and other portals → Seminars
- Points of interest: DAaaS, remote access, cloud infrastructure, data transf. between facilities
- One Goal: enable multi-modal experiments

What is the DMA ST1 communication infrastructure?

- So far:
  - DESY based mailing list → OK
  - DESY based confluence → needed DESY accounts, deprecated at DESY
  - DESY based Indico for conferences → OK, make more use of it
  - DESY based ZOOM for meetings / seminars → OK
- New channels:
  - Confluence → (HZB/nubes) → DESY SyncAndShare (+Helmholtz AAI!) (for minutes & internal documents)
  - Rapid communciation: Mattermose channels: HIFIS instance
    - Goal: Buidling up community, offering communication platform beyond the scope of the „Matter information fabric“ infrastructure → ignite & moderate discussions on special topics
  - A face for non-ST1 people: Website
    - DESY hosted website
- Status: Working on this, need some underlying work on AAI … hope still in 2023

… and two new topics of relevance after DMA setup

- Sustainability
  - different meanings: this time, the envirnmental definition:
  - Infrastructures and workflows should be sustainable
- Security
  - Services and centers are being attacked
  - Find suitable new services and infrastructures that are inherently secure and user friendly