

European XFEL Users' Meeting 2024

Satellite Workshop: Data management, analysis, and reduction at European XFEL

The concept of data management plans for European XFEL proposals

Fabio Dall'Antonia

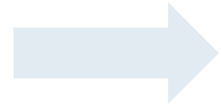
on behalf of the Data Policy and Reduction working group

26 January 2024



High-level purpose and motivation of DMPs

Challenges for
data management and analysis



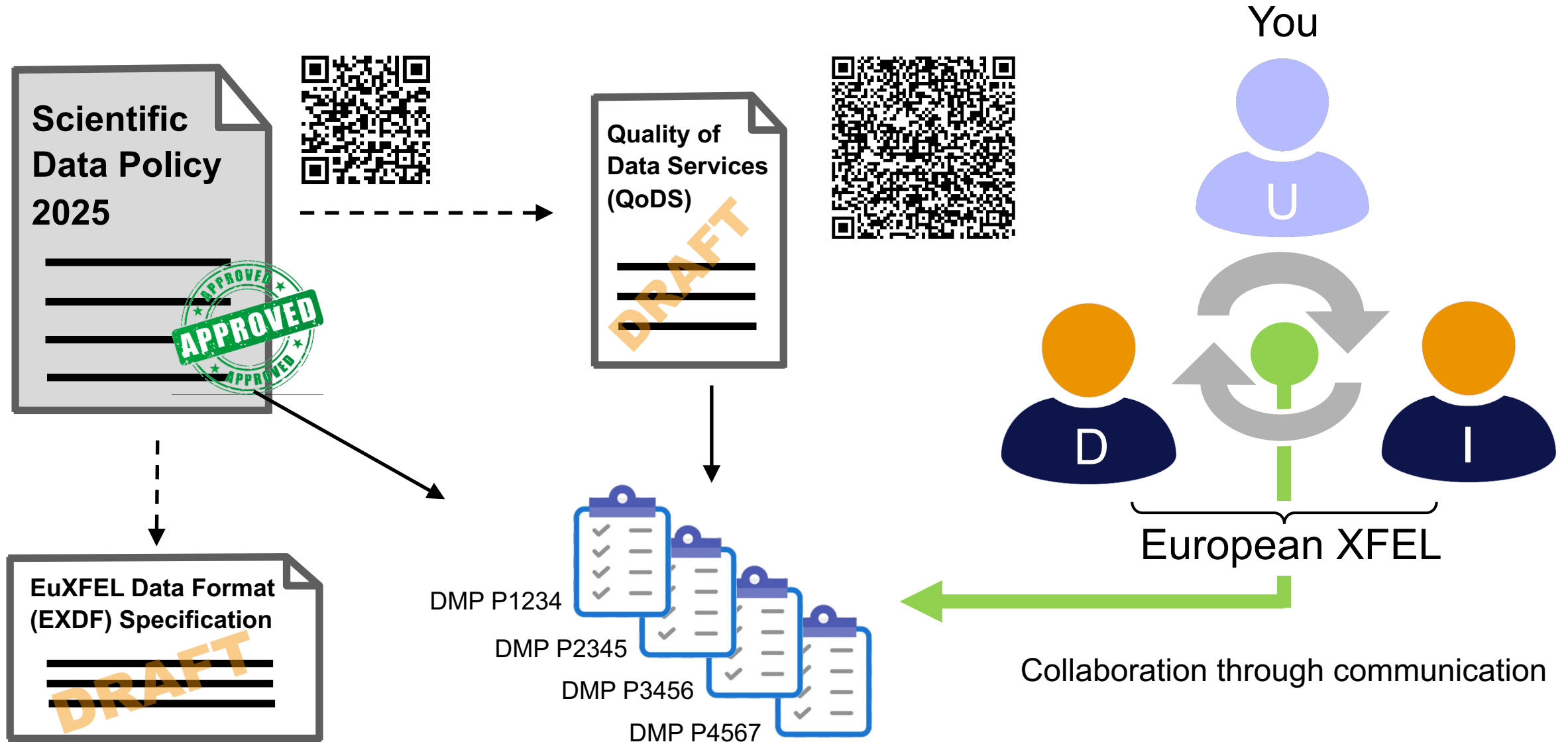
Recommendations from
Open Science projects



Purpose as motivation driver: DMPs are intended to

- optimise your experience before, during and after the beamtime regarding data aspects
- serve as documentation and reference, including future proposal submissions, experiment planning and (open) data-reuse
- fulfill formal requirements by funders with respect to FAIR/Open data, both for experiment groups and facilities

DMPs in the SDP framework



DMP characteristics

- The DMP is conceptually a **service**, not a document
- DMP instances are **specific** for every proposal
- DMPs aim at **collaborative** planning and agreements on all data-related requirements between stakeholders, addressed in a systematic way.
- DMPs represent **dynamic** information pools, that is, they are incrementally enriched and adapted at each stage of a proposal lifecycle
- Gathering DMP information is **streamlined** by integrating relevant communication processes to existing ITDM infrastructure

5 Data management plan

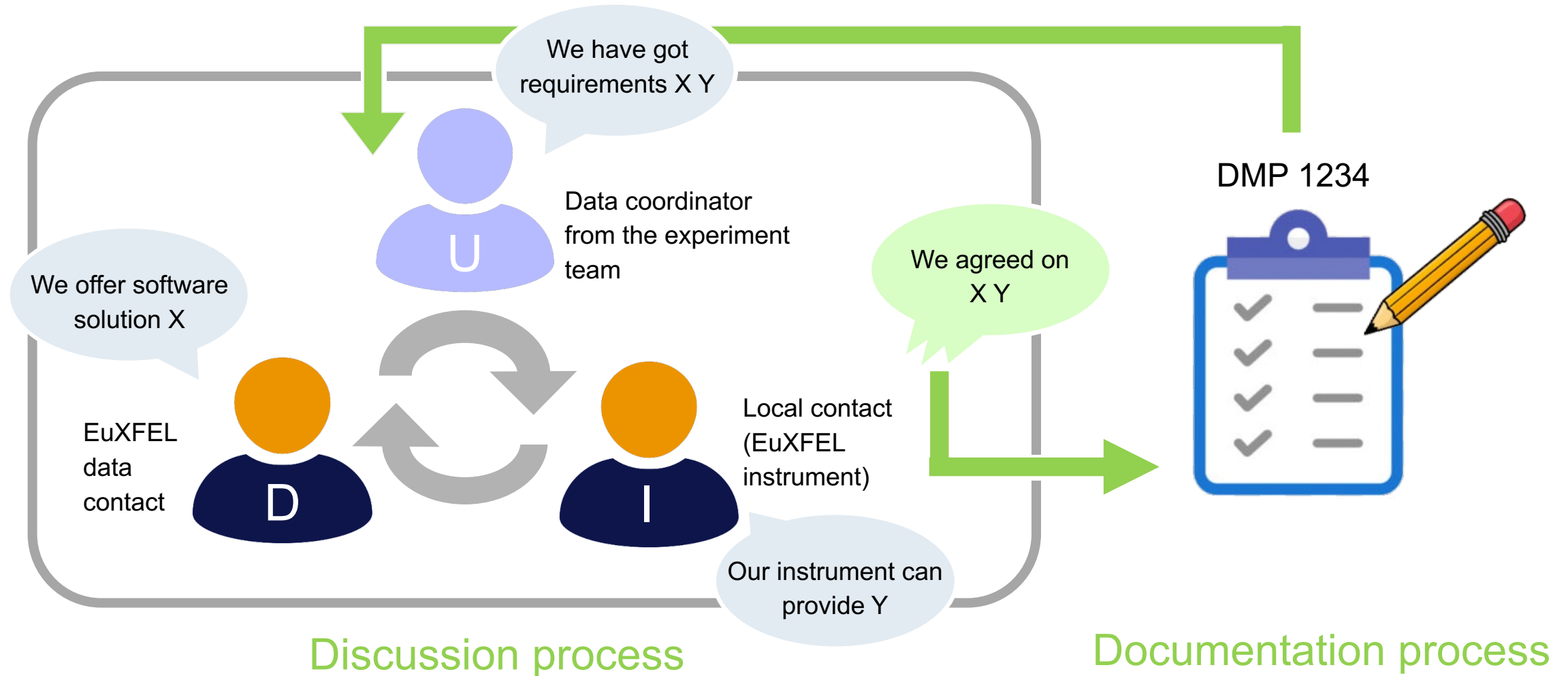
- 5.1. A data management plan (DMP) will enhance the communication between the European XFEL GmbH and the proposal and experiment teams with respect to data management.
- 5.2. A DMP must be specifically defined for each proposal.
- 5.3. A DMP will be created at the proposal submission stage and the European XFEL GmbH will ask the users to confirm or update the DMP at different stages until the end of the embargo period. The European XFEL GmbH reserves the right to reject a DMP and any of its updates on the grounds that the DMP is neither comprehensible nor accurate.
- 5.4. The DMP will clarify all aspects of data management and will, in particular, document agreements reached between the PI and European XFEL GmbH, with respect to the provision of data storage, transfers, processing and analysis, computing resources, data retention periods, and data disposal.
- 5.5. Users must apply their best knowledge and due diligence when completing the DMP and are required to follow the guidelines provided by the European XFEL GmbH on what constitutes good data management and guidelines for completing DMPs.

Covered in SDP Section 5



DOI: 10.22003/XFEL.EU-TR-2025-001

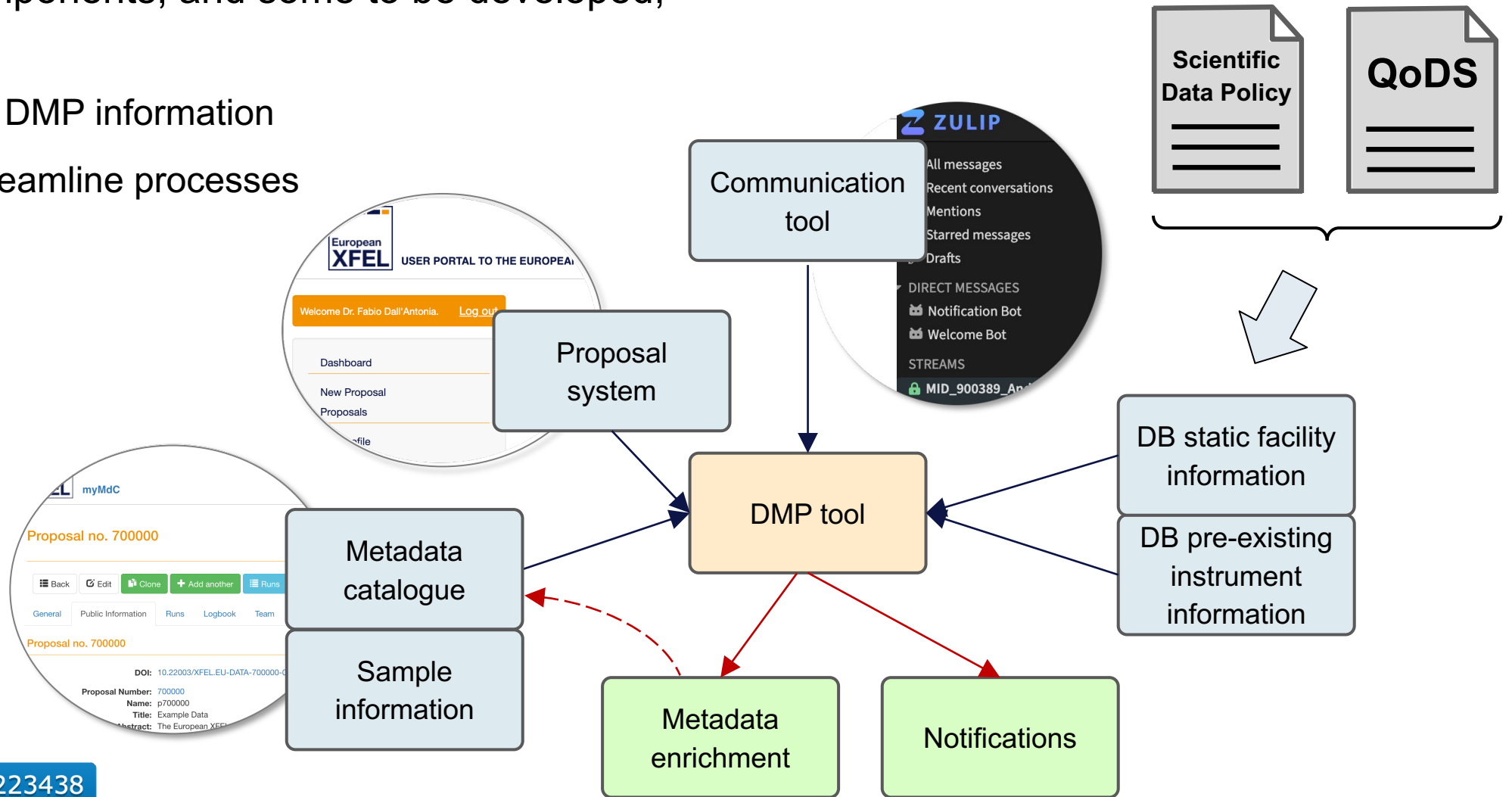
Processes and roles in the DMP concept



DMP service infrastructure

Existing components, and some to be developed, serving as:

- ▶ Sources of DMP information
- ▶ Tools to streamline processes

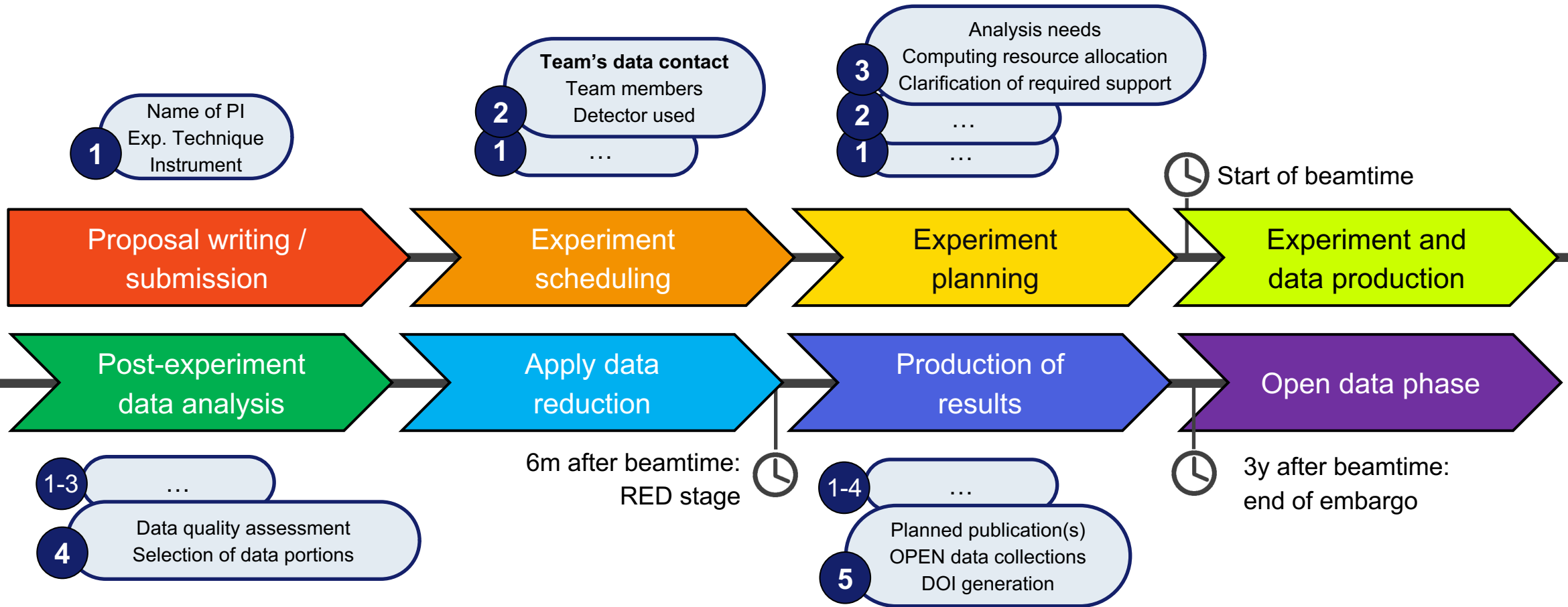


Scheme based on



Proposal lifecycle and DMP checkpoints

What is known or needs to be discussed when?

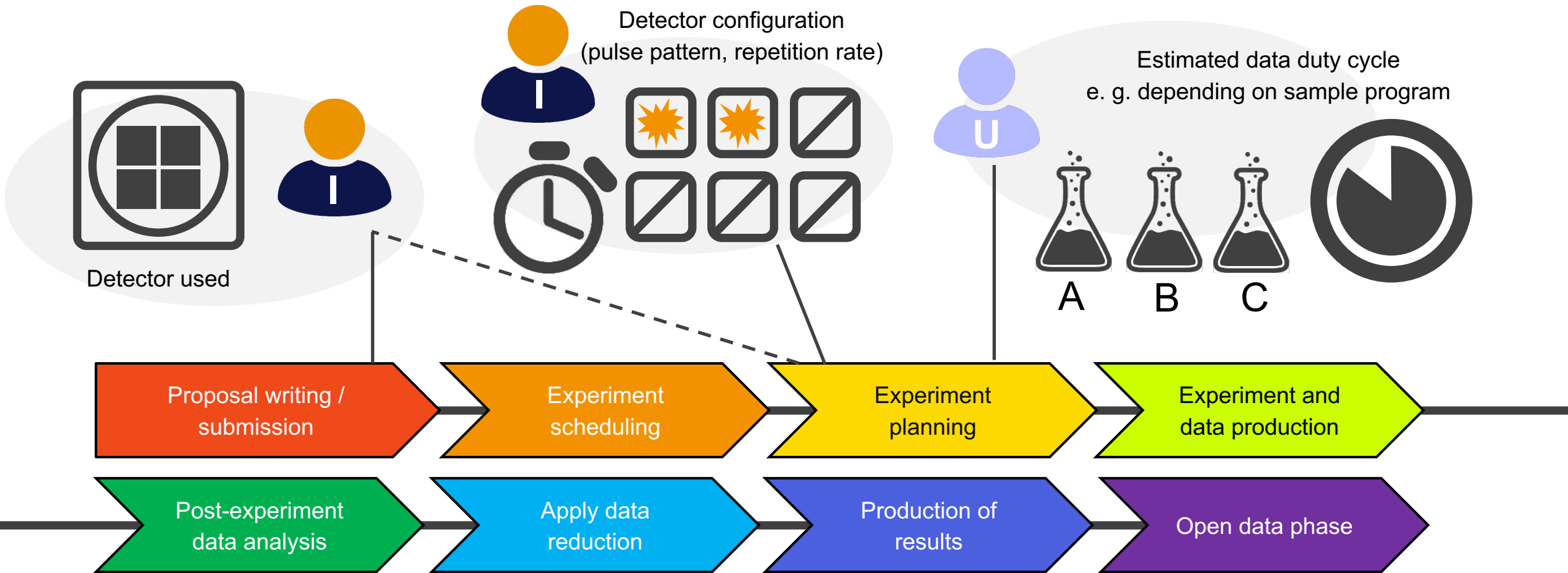


Information targets (from facility perspective)

■ What we need to know in order to support you best

We need to know ...	Parent information to be retrieved via DMP
... the storage space to keep free for an expected volume of experiment data that will be generated	What detector will be used? What is the detector configuration? What fraction of data acquisition is expected? Are corrected detector data required?
... the data analysis (support) requirements, in order to prepare SW and allocate staff resources	Is existing facility SW used? Is community SW to be deployed? Are there special SW solutions to be developed (online/offline/both)? Is in-person software support needed (online/offline/both)? Is the use of data/metadata standards like NeXus planned?
... the options for data reduction in order to prepare for the reduced-volume retention on disk (RED storage)	Which experimental technique is applied? Are suggested facility service methods for reduction to be used? If so, which one? Will a custom user method be used instead for which arrangements are needed? What is the required degree of reduction?
... expert support requirements for integration of custom hardware, in order to allocate control system staff resources	Are there new hardware components that need to be integrated to the control system? Do bridges for external control SW need to be prepared? Is personal support needed for these tasks?
... performance requirements, given the experiment characteristics, in order to allocate compute infrastructure	How many HPC compute nodes are required for offline analysis? How much latency is acceptable for quasi-real-time monitoring and feedback?

Example: how to derive the expected data volume

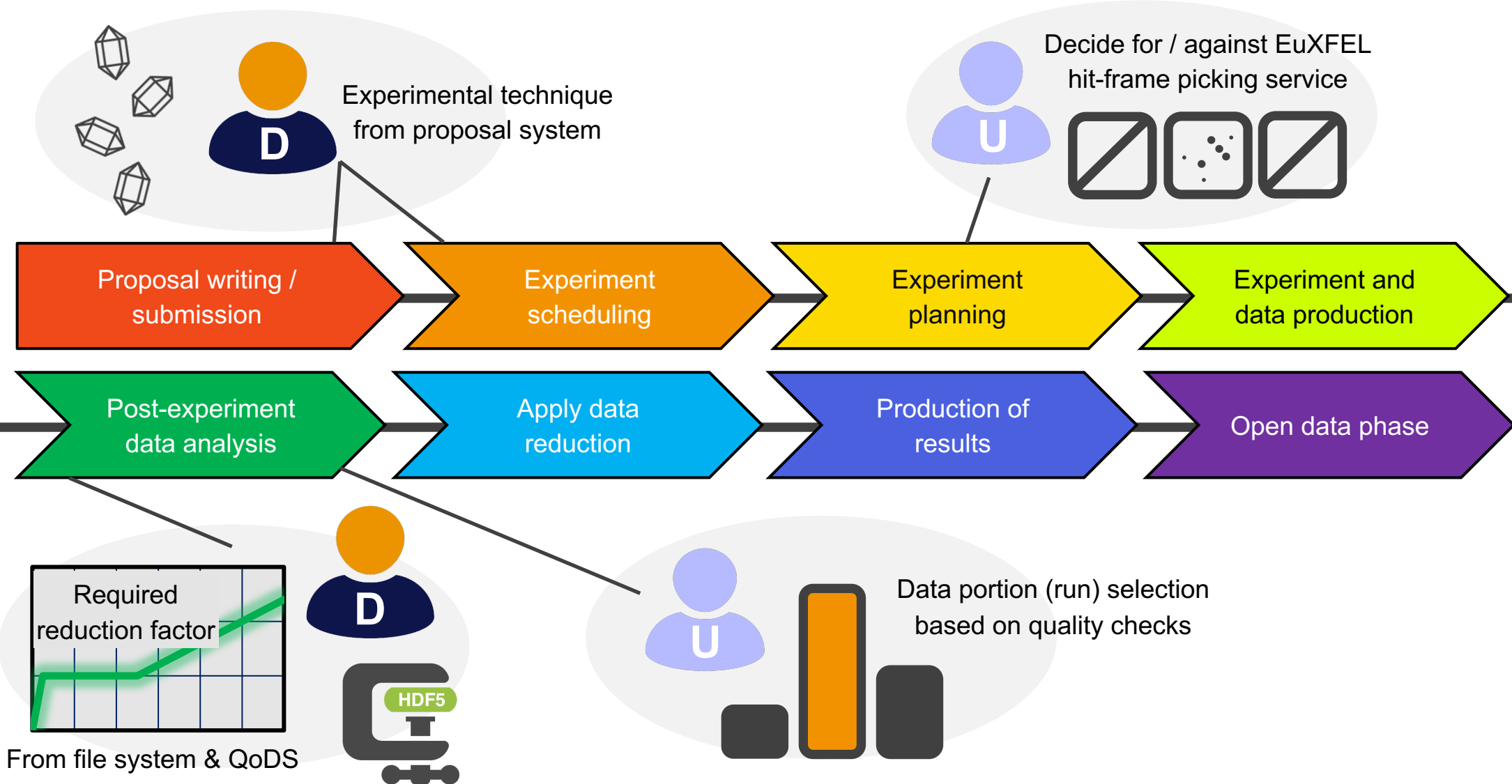


Information targets (from facility perspective)

What we need to know in order to support you best

We need to know ...	Parent information to be retrieved via DMP
... the storage space to keep free for an expected volume of experiment data that will be generated	What detector will be used?
	What is the detector configuration?
	What fraction of data acquisition is expected?
	Are corrected detector data required?
... the data analysis (support) requirements, in order to prepare SW and allocate staff resources	Is existing facility SW used?
	Is community SW to be deployed?
	Are there special SW solutions to be developed (online/offline/both)?
	Is in-person software support needed (online/offline/both)?
	Is the use of data/metadata standards like NeXus planned?
... the options for data reduction in order to prepare for the reduced-volume retention on disk (RED storage)	Which experimental technique is applied?
	Are suggested facility service methods for reduction to be used? If so, which one?
	Will a custom user method be used instead for which arrangements are needed?
	What is the required degree of reduction?
... expert support requirements for integration of custom hardware, in order to allocate control system staff resources	Are there new hardware components that need to be integrated to the control system?
	Do bridges for external control SW need to be prepared?
	Is personal support needed for these tasks?
... performance requirements, given the experiment characteristics, in order to allocate compute infrastructure	How many HPC compute nodes are required for offline analysis?
	How much latency is acceptable for quasi-real-time monitoring and feedback?

Example: how to prepare for data reduction, if you need to




A mock DMP – how it could look like

dmp.xfel.eu/proposal_9999

Information / agreements Discussion

Experiment team data coordinator (e-mail):	jane.doe@university.edu
EuXFEL data contact (e-mail)	john.doe@xfel.eu
EuXFEL instrument contact (e-mail)	erika.mustermann@xfel.eu
Experimental technique:	serial femtosecond crystallography ▼
Detector used:	AGIPD 1M ▼
Pulses per train used:	352
Expected data duty cycle %:	70
Expected raw data size:	1500 TB
Data analysis software to be used:	OnDA, CrystFEL
Preferred data reduction method:	Detector frame selection ▼

Edit  Export to PDF ▼

DMPs in the timeline of the SDP update



- Legal consultations and adjustment started
- Consultation with the EuXFEL Scientific Advisory Committee

- SDP approved by the EuXFEL Council
- DMP task force formed

Start of implementation phase incl. DMP pilot (voluntary involvement of users)

SDP will come into effect *

* The SDP will apply to proposal submissions in 2025, and thus to corresponding beamtimes in 2026

Updates on implementation: subscribe to computing@xfel.eu (Log-in to UPEX and click on "Mailing lists" in the navigation)

Getting involved

We need to know ...	Parent information to be retrieved via DMP
... the storage space to keep free for an expected volume of experiment data that will be generated	What detector will be used? What is the detector configuration? What fraction of data acquisition is expected? Are corrected detector data required?
... the data requirements for the experiment, in order to prepare for the reduction (RELS) task (RELS)	Is existing facility SW used? Is community SW to be deployed? Are there special SW solutions to be developed (online/offline/both)? Are there special SW solutions to be developed (online/offline/both)? Is the use of data/metadata standards like NeXus planned?
... the operational requirements for the experiment, in order to prepare for the reduction (RELS) task (RELS)	Which experimental technique is applied? Are suggested facility service methods for reduction to be used? If so, which one? Which experimental technique is applied? Are suggested facility service methods for reduction to be used? If so, which one?
... expert support requirements for integration of custom hardware, in order to allocate control system staff resources	What is the required degree of reduction? Are there new hardware components that need to be integrated to the control system? Do bridges for external control SW need to be prepared? Is personal support needed for these tasks?
... performance requirements, given the experiment characteristics, in order to allocate compute infrastructure	How many HPC compute nodes are required for offline analysis? How much latency is acceptable for quasi-real-time monitoring and feedback?



We highly appreciate your suggestions in order to complement this list with aspects that are relevant for you.

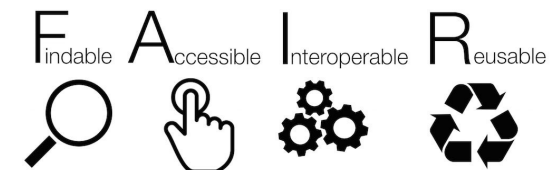
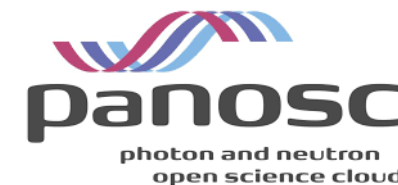
Contact us at data-policy@xfel.eu

Conclusions

- The main motivation for DMPs is to ***improve your experience*** during the entire proposal lifecycle, while also meeting FAIR data requirements
- The DMP implementation concept originates from existing processes involving instrument scientists and data staff – based on that, we aim to ***enhance communication***, in particular with you, on data management matters
- DMP services shall as much as possible be based on existing infrastructure, some dedicated technical solutions will be developed on top
- The conceptual ideas presented here are still in a drafting stage and need further shaping
- The implementation pilot/test phase will start in 2024: your feedback and voluntary participation will help to achieve an implementation rolled out officially in 2025



Thank you for your attention!



If you are interested in being part of the implementation process, let us know.
Contact us at: data-policy@xfel.eu

