Federated Data Management Lifecycle

We suppose a small-to-medium size international community, that ...

- ❒ has sizable data volumes (e.g. from simulation, detector, telescope)
- ❒ knows the metadata to describe it
- ❒ uses or plans to use distributed storage infrastructures
- ❒ wants to adopt FAIR data management.

Example communities

- ❒ Lattice QCD (ILDG)
- ❒ Radio Astronomy (GLOW)
- ❒ Small and medium-size particle and nuclear physics experiments (ALPS, IAXO, ...)

- ✗ **Not** WLCG- or SKA-scale communities

Use case

❒ Dataflow
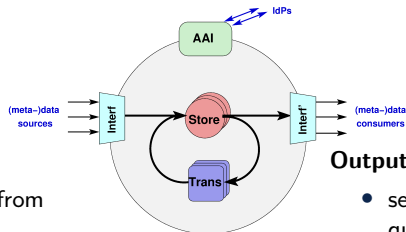
Input / write access

- ingest data and metadata from source(s)
- mint unique ID
- validate, store (possibly redundant), and register the (meta-)data
- manage hierarchical access permissions

Output / read access

- search data by (possibly complex) queries of metadata
- retrieve data and metadata by ID
- expose and/or convert metadata
- optional data publishing process
- optional data transformations

❒ Rough resource estimate: 3 FTE at DESY (at least 1 SW developer)
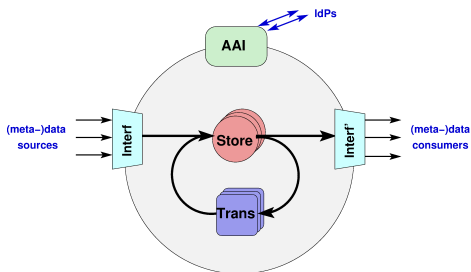See "Components" slide for details.

❒ Involved / interested partners: FZJ, UBi, UR, ... (+ synergies with e.g. CTAO)

Backup

Components

class	component	activity
Store	Edge service (teapot)	C**AS**
	Metadata Catalogue	C**AS**
	File Catalogue	C**AS**
Trans	File Transfer Service	**A**S
	Rucio	**AS**
	Crosswalks	C**AS**E
Interf	Ingestion from detector	**CA**S
	Ingestion from control system	**CA**S
	Markup GUI	CAS
Interf'	Metadata Harvesting (OAI-PMH)	C**A**SE
	Search GUI	**CAS**E
	Data publishing	ASE
AAI	Access Control Service	**CAS**
	IAM extensions	C**AS**
	Onboarding iam4nfdi	**AS**

C=component development, A=assembly/integration, S=support, E=scientific exploitation/results

Architecture

- **AAI:** Authentication and Authorization Infrastructure
  Examples: Unity AAI, Indigo IAM, . . .

- **Interf, Interf':** Input and output Interface(s)
  Components: Ingestion or packaging tools, DAQ systems, web portals, landing pages, search or harvesting interfaces, APIs, . . .
  Examples: LLM assisted markup, HDF5 packager

- **Store:** (Meta-)Data Storage
  Components: Storage4PUNCH, metadata catalogue, file catalogue, code- or software repository, container registry, . . .
  Examples: GitLab, CVMFS, SciCat, Rucio, . . .

- **Trans:** (Meta-)Data Transformations
  Components: Compute4PUNCH, file transfer service, metadata cross-walk, DOI minting, . . .
  Examples: CERN FTS, Globus online . . .

Use-case: detailed characterization

Users	international, possibly world-wide distributed community
Storage	possibly geographically distributed, different types (tape, disk) and interfaces (dCache, S3, ...)
Transformations	trivial (only replication and/or transfer)
FAIR compliance	desirable (incl. persistent ID, rich metadata, registration, access control, standard protocols and markup, provenance, license, ...)
Metadata	well-defined and known
Metadata schema	community-specific, possibly complex, possibly including DC
Data growth	$\leq$ 10 PB / y
	$\leq$ 100'000 data objects / y
Read access	restricted by embargo, or VO-wide, or public
Write access	fine-grained control delegated to collaborations/projects
Interfaces (optional)	ingestion, markup, search, harvesting, data publication
GDPR compliance	possibly required