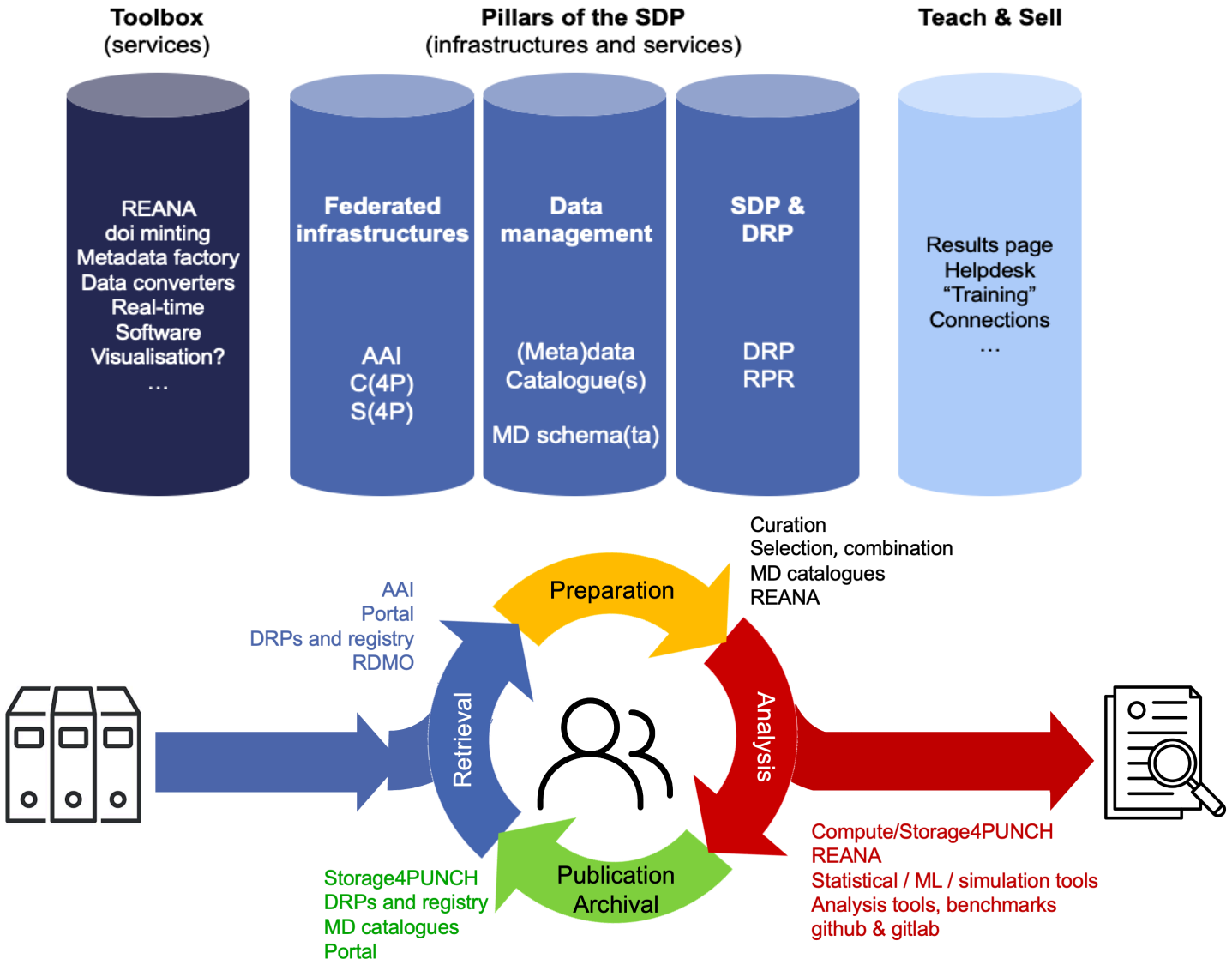
Preparation PUNCH 2.0

The Unnamed Pillar (formerly known as Data Management)

Hamburg 4 March 2025

Frank Wagner; Hubert Simma; Uwe Hernandez Acosta; Jörn Künsemöller; Andrew Mistry; Stephan Hachinger; Victoria Tokareva; Tim Wetzel; Alexander Wellmann; Tilo Wettig; Gunnar Bali; Sagar Borra & PUNCH-EB

Idea on Structure (before Kassel)



Definition of Data Management

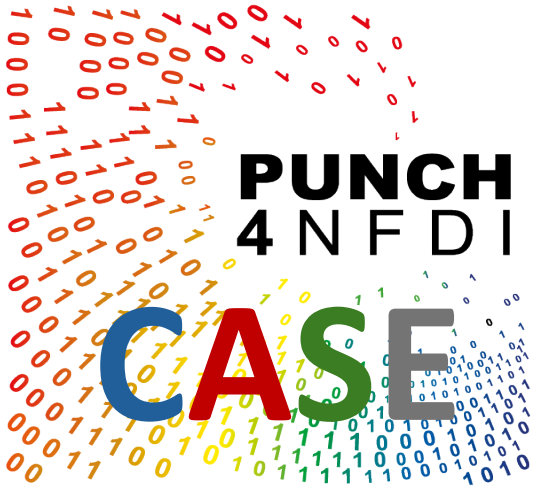
Data Management refers to the processes, policies, and tools used to collect, store, organize, preserve, and share data efficiently and securely throughout its lifecycle. In **scientific research**, data management ensures that data remains **accessible, reproducible, and interoperable** while adhering to best practices such as the **FAIR Principles (Findable, Accessible, Interoperable, Reusable)**.

(chatGPT)

Conceptional discussions in and after Kassel

What does need PUNCH to provide? What a user needs?

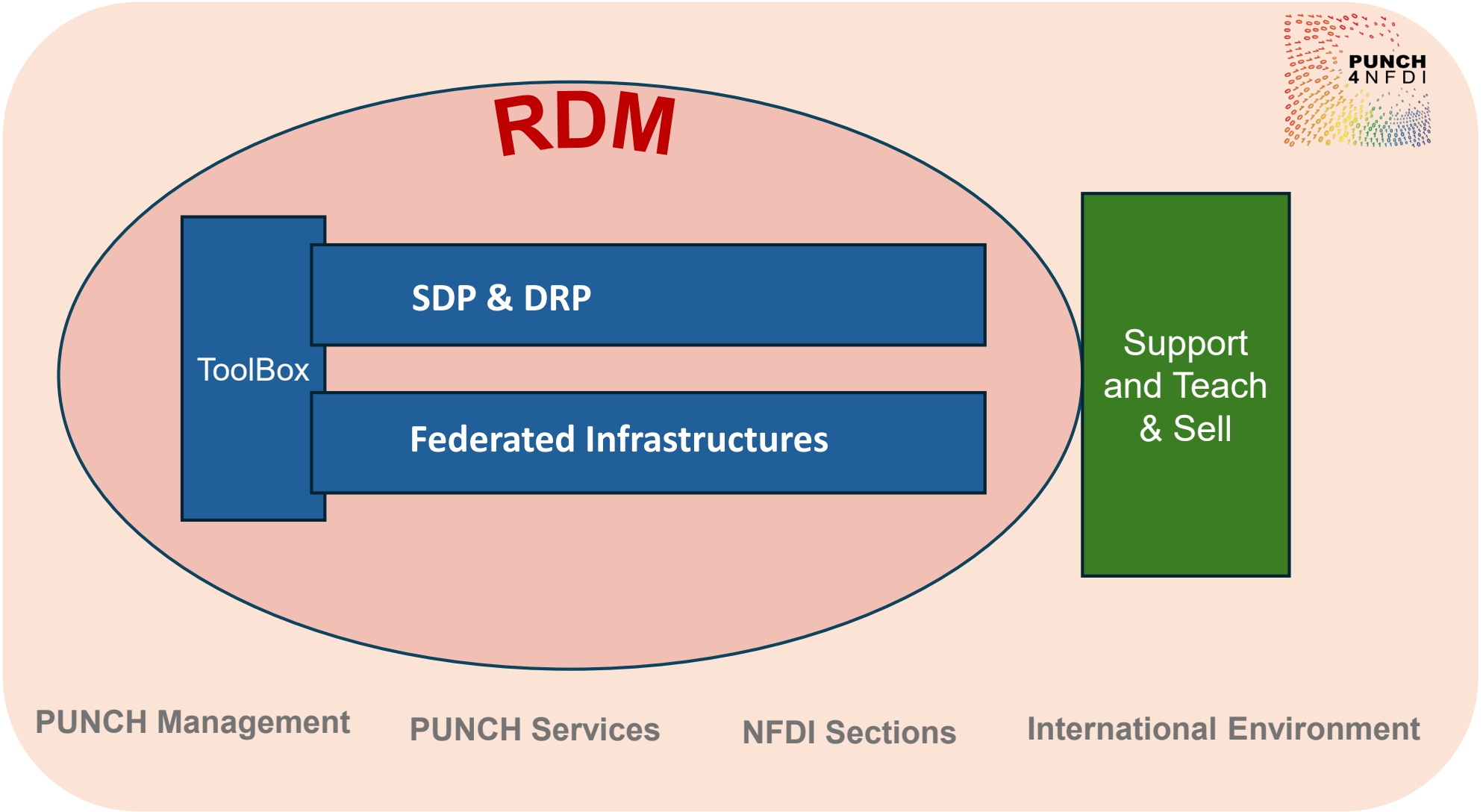
- **C**omponents (Technologies) for transforming data
  [data, computing, storage,LLM, AAI, …]
- **A**ssembling of use case specific components
  [catalogues, interfaces, workflows, ….]
- **S**upport
  [training, services, help, ….]
- **E**nvironment
  [management, NFDI, EOSC, Collaborations, ….]



➔ **The task of the „assembling team“ is to glue the components together**

Adapted from Hubert Simma / Uwe Hernandez Acosta

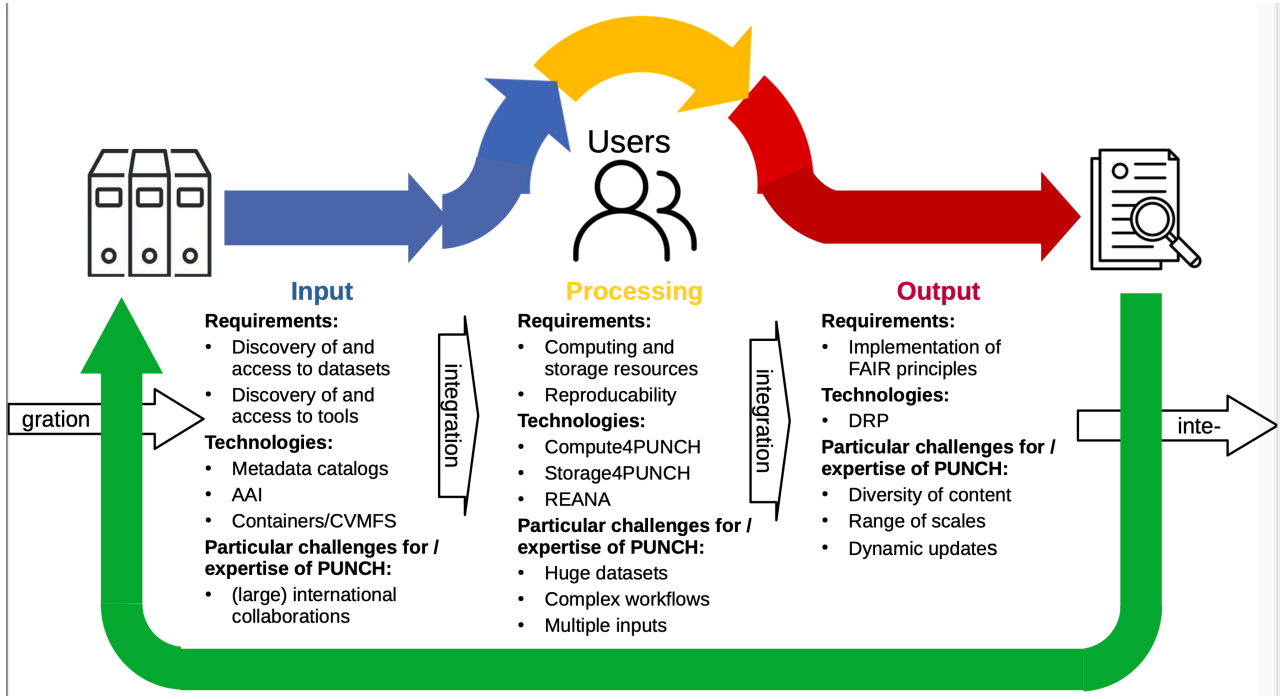
RDM = Assembling components (technologies)



Components (can exist or to be futher developed)

- OAI-PMH
- Metadata catalogues
- Metadata transformation (crosswalk tools/services)
- File catalogue (identifiers)
- Storage4PUNCH
- File transfer service
- File management/orchestration (RUCIO)
- Search portals (SDP) (here LLM??)
- Metadata ingestion/generation
- Compute4PUNCH (including monitorand accounting systems)
- AAI (needs evaluation) (may be use IAM4NFDI)
- DOI minting, publishing (via Zenodo, e.g.), PID4NFDI(?), InHPC Fair Data Portal
- ….

Idea on Structure (Thomas Kuhr)



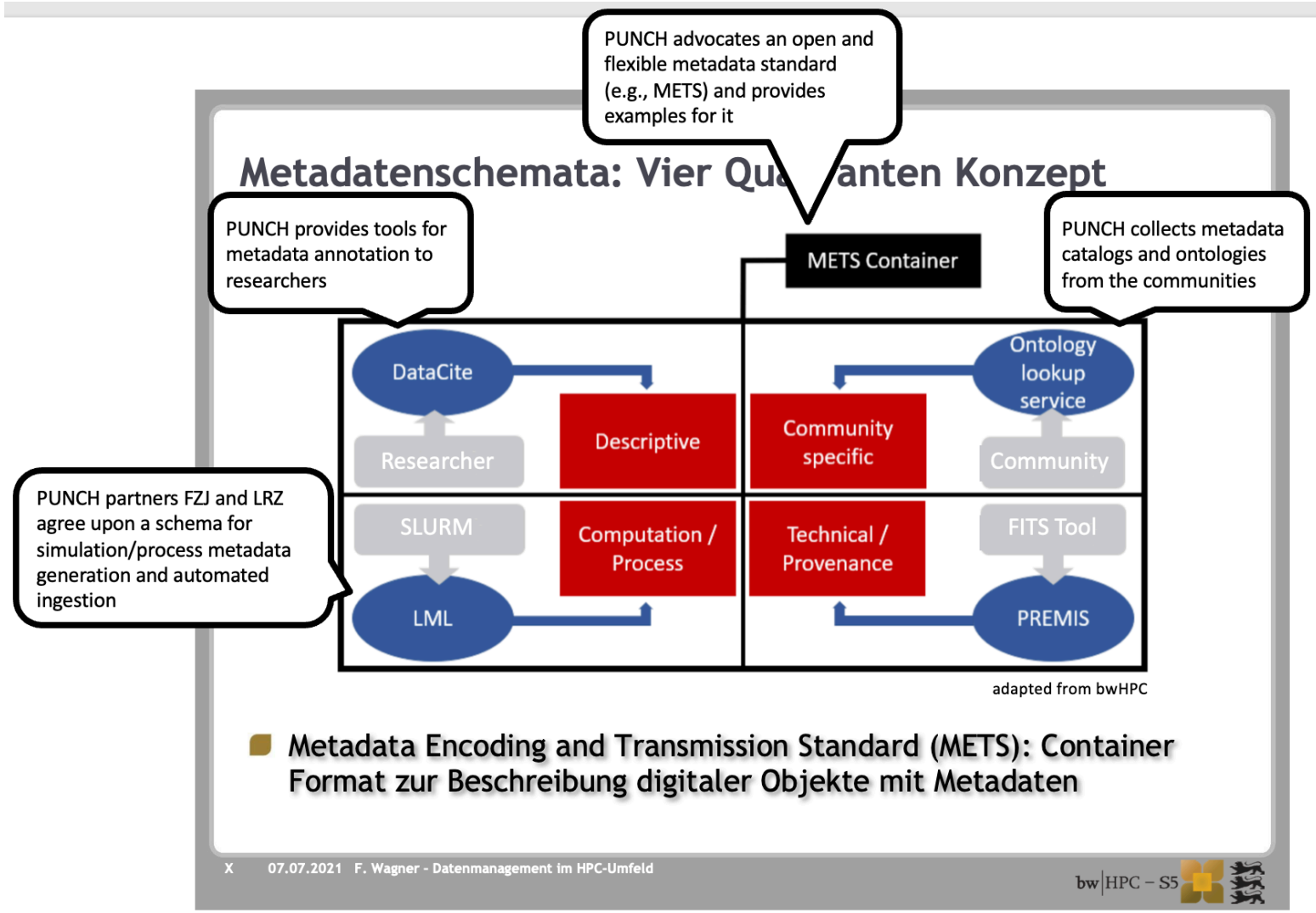
Assembling: (Meta)Data Management in PUNCH

Ideas from input of applicants in overleaf:

- Metadata schemata ➔
- Metadata conversion ➔
- Metadata and DoI minting/finding ➔
- Metadata for simulation / software ➔
- Metadata for Instruments ➔
- Metadata provenance / archive ➔
- Energy-efficient data handling ➔
- RDM with/at HPC ➔
- (Knowledge Graph / Large Language Model) ➔

- **Needs to be elaborated!**
- **What is important for PUNCH?**
- **What we define belonging to the RDM?**
- **Which tools we need and/or have to develop?**
- **What is requested for the "typical" use cases**
- **What results in sustainable services from PUNCH?**

One idea for possible overarching Metadata schemata



By Frank Wagner

The unnamed pillar - input

-- DESY will contribute to **metadata schemata development and metadata catalogue solutions**—e.g. SciCat and / or ILDG for all DESY activities and beyond (nuclear, small accelerators, ...).
– DESY will develop a defined procedure for a scientific community to arrive at a metadata schema (best practices, tools, ...)
– DESY will develop metadata schemata & cross-walk registry for **conversion between different metadata schemata** for use with various catalogues → would allow for higher degree of interoperability between different schemata
– DESY will contribute to **DOI minting and publication processes**.

– FZJ and DESY will contribute the **minting of persistent identifiers** (“DOI minting”) and data publishing workflow as part of a modular distributed (meta-)data management.

– FZJ and LRZ will take a role in the next step towards **FAIR simulation data** by implementing the strategies developed in PUNCH 1.0 into practice. This includes developing metadata schemes, easy storage tools, data accessibility tools, connective DOIs, training, etc.)
– FZJ and LRZ will develop workflows and tools for **publishing data automatically with metadata**, but with the possibility to augment this data by the user, if required. Also, the integration of these published datasets in DRPs is envisaged.
– FZJ will take a role in developing strategies, implementing and teaching **energy-efficient research data handling**.
– FZJ and LRZ will contribute to connect PUNCH to **HPC data handing strategies**. JGU is also willing to contribute to this.

The unnamed pillar - input

– AIP will contribute to **DOI minting and publication processes** using DRP
– AIP will contribute to improve **findability** of data across PUNCH communities and accessibility using community specific protocols (astropy/TAP, etc)

– KIT will contribute to **metadata** schemata development, metadata catalogue solutions and metadata extraction, in particular taking into account the demands and needs for Astroparticle Physics (…also simulations and experiments).

– FAU will contribute to **metadata for software** (link to CodeMeta)

– GSI will work on enhancing **metadata schema** for Nuclear physics communities, with improved end-user interfaces and catalogues.
– GSI will contribute to API developments to enable transfer of metadata across interfaces: development of **knowledge graphs** to link schema with bibliographic systems, data repositories, data management planning tools, code repositories, etc.
– GSI will contribute to establishing metadata cross-links with broader communities like NAPMIX to ensure interoperability.

The unnamed pillar - input

– UBI will continue to work on **metadata catalogue** improvement both in lattice and astro
– DESY, FZJ, UBi and UR will contribute the development of **modular distributed (meta-)data manage**ment services for the hep-lat (lattice) and further communities.

– DZA will contribute to R&D on **metadata in the petabyte range** (relevant for SKAO) and provide access to this data.

– UR will contribute the development and extension of the **data provenance** information for use in the modular distributed data and metadata management services.

Common Tools and Services for Research Data Management

1. **Data Storage & Infrastructure**
   - **HPC & Cloud Services**: SDP C4P S4P
   - **Federated Computing Systems**:
   - **Databases**:
2. **Data Organization & Curation**
   - **Metadata Standards**: METS, DataCite, DublinCore?
   - **Metadata Management**: RUCIO, DIRAC, iRODS?
   - **Data Catalogs & Repositories**: Zenodo, HEPData, OPUS
   - ~~**Electronic Lab Notebooks (ELNs)**:~~
3. **Data Sharing & Accessibility**
   - **FAIR Data Repositories**: CERN Open Data Portal, KCDC, GAVO, …
   - **Persistent Identifiers (PIDs)**: DOI (Digital Object Identifier), ORCID
4. **Data Analysis & Processing**
   - **Big Data Processing**:
   - **Machine Learning for Data Management**:
   - **Workflow Management**:
5. **Data Preservation & Long-Term Archiving**
   - **Version Control & Reproducibility**: Git, GitHub, GitLab, Zenodo Integration
   - **Backup & Archival Services**:
6. **Compliance & Policy Support**
   - **Regulatory Compliance Tools**:
   - **Data Management Planning Tools**: ?