

SLICES Data Management Infrastructure for Reproducible Experimental Research on Digital Technologies

Yuri Demchenko, University of Amsterdam

On behalf of

Yuri Demchenko, Paola Grosso, Shashank Shrestha

Acknowledgement to SLICES partners: Sebastian Gallenmuller, Serge Fdida,
Panayiotis Andreau, Damien Sauzes, Thijs Rausch

Track: CompSys Research for a Responsibly Digitalised Society

ICT.OPEN2024 11 April 2024, Utrecht

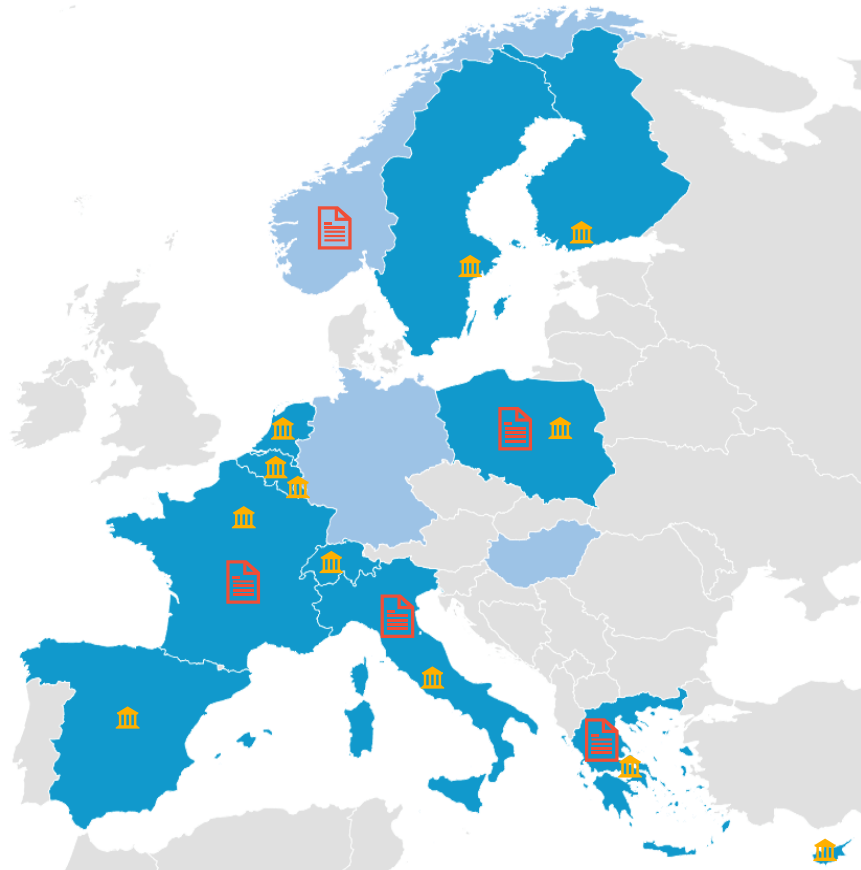
Outline

- SLICES-RI: European Scientific Large-Scale Infrastructure for Computing/Communication Experimental Studies
- SLICES Data Management Infrastructure for Experimental Research
 - Experimental data lifecycle
- Experimental Research Reproducibility as a Service
 - Experiment organization and workflow
 - Metadata definition
- Tools to support data management and FAIR data principles
 - EOSC Data Management and Metadata Tools: MSCR, FDO, PID, others
- Discussion




SLICES-RI for Research on Digital Infrastructures

Includes extensive experimentation with new technologies



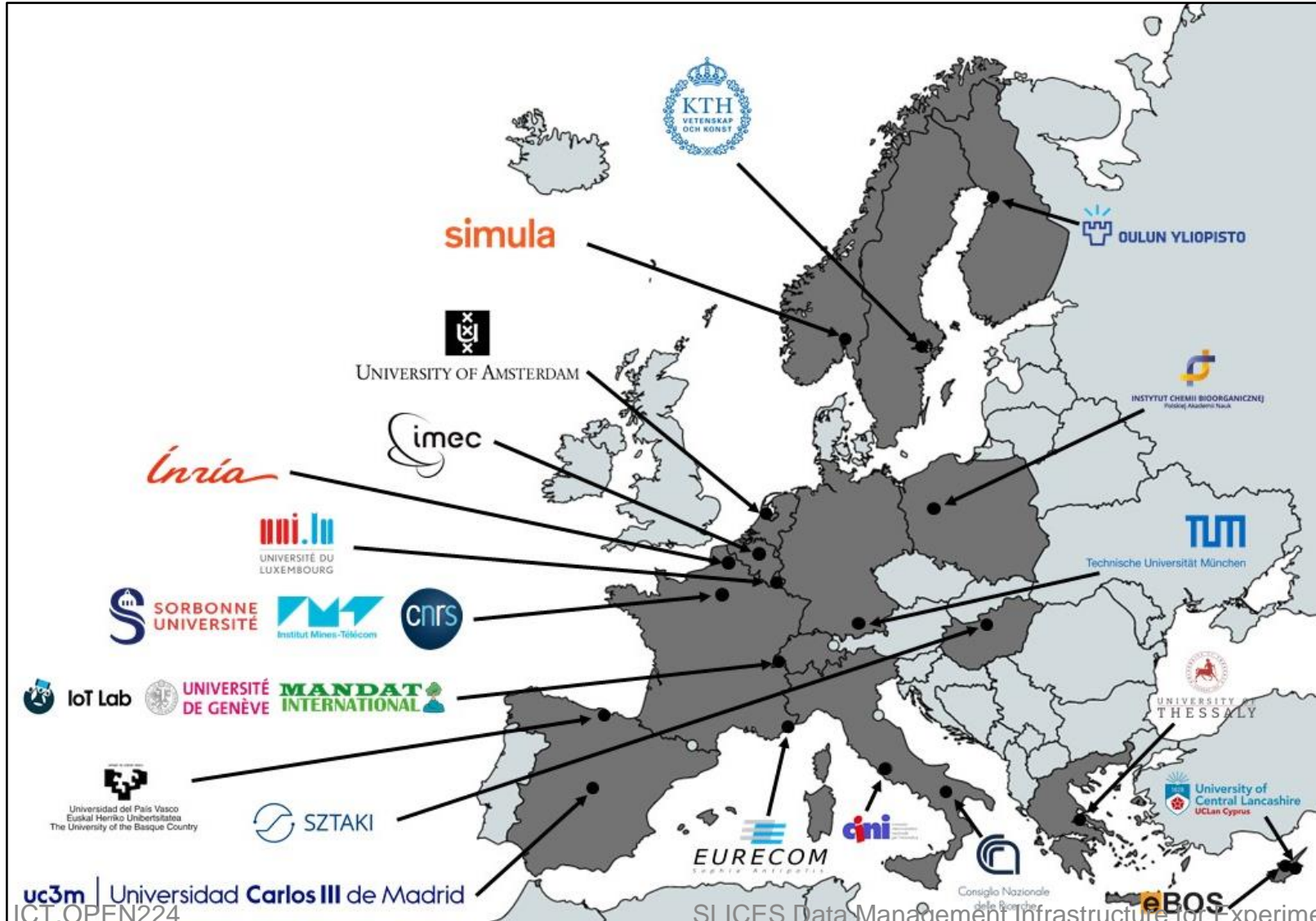
Initiated in 2017, **25 partners** from 15 countries:

- **12 political supports** from National Ministries 
- included in **5 national roadmaps** 

SLICES will enable **scientific excellence and breakthrough** and will **foster innovation in the ICT domain**, strengthening the **impact of European research**, while contributing to European agenda to address **societal challenges**, and in particular, the twin transition to a sustainable and digital economy.



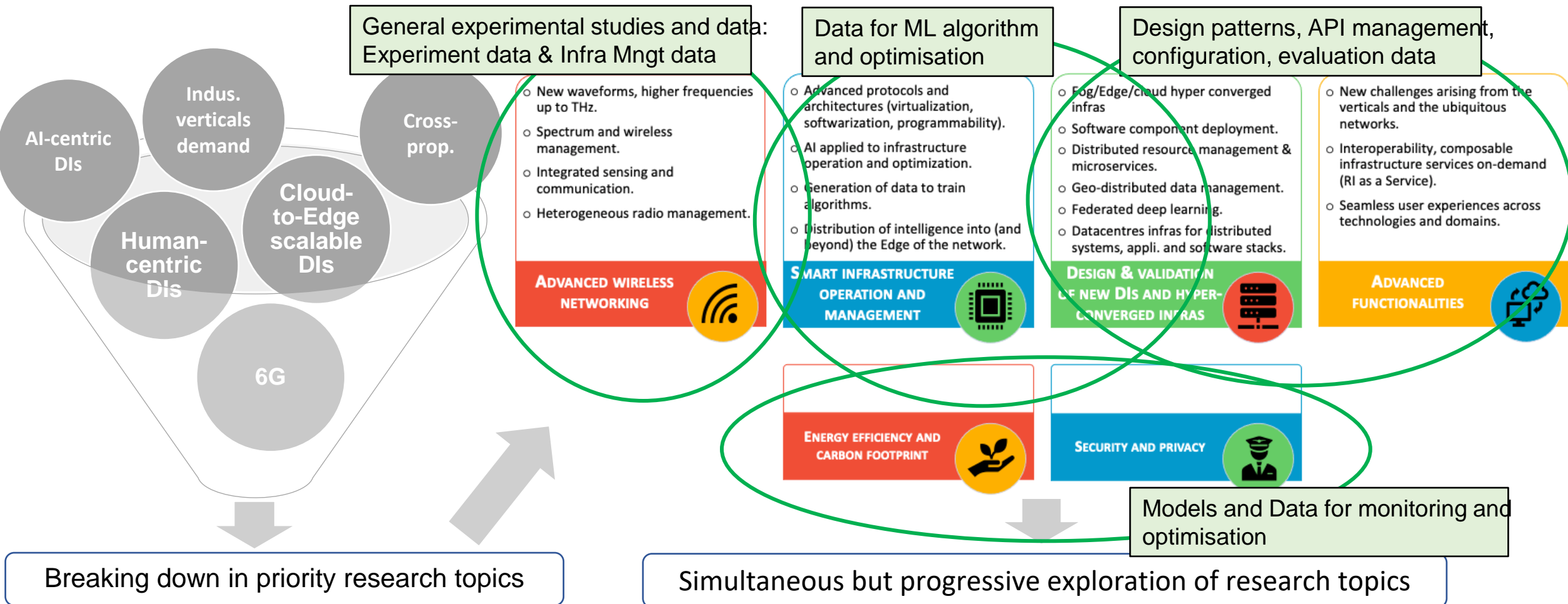
SLICES-PP (2022-2025): Consortium members



25 Partners from 15 countries

- INRIA, FR
- Sorbonne University (SU), FR
- **Univ of Amsterdam (uvA), NL**
- Univ of Thessaly (UTH), GR
- CNR, IT
- PSNC, PL
- Mandat International (MI), CH
- IoTLAB, FR
- UC3M, ES
- IMEC, BE
- UCLan, CY
- EURECOM, FR
- SZTAKI, HU
- CINI, IT
- CNIT, IT
- Univ Luxemburg, LU
- TUM, DE
- EHU, ES
- KTH, SE
- Univ Oulun, FI
- EBOS, CY
- SIMULA, NO
- IMT, FR
- Univ Geneve, CH

Different Types of Data for Different Experimental Studies



Experimental Research Reproducibility as a Service (ERRaaS)

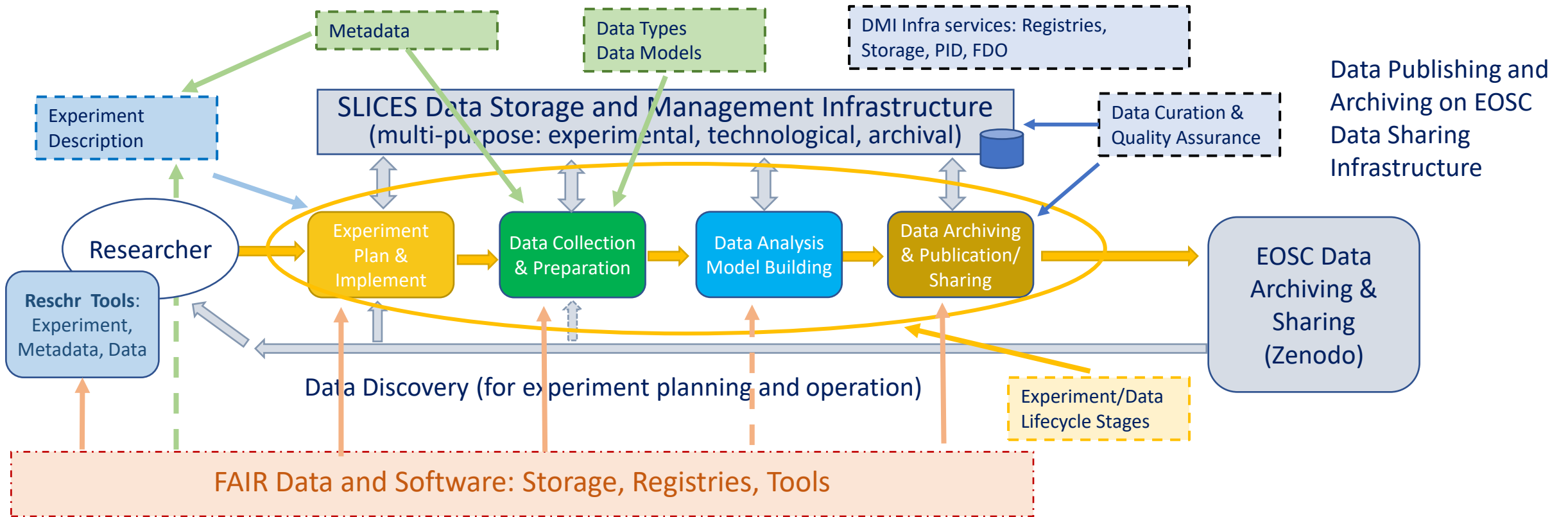
- **SLICES to provide the Robust Data Infrastructure for Experiment/Data Driven Research**
 - Interoperability and integration with EOSC as Federated data infrastructure
- **Experiment as a Research Object (RO)**
 - Identified with unique ID and containing smart metadata (for discovery and FAIR compliance)
 - Complying with the FDO/SFDO metadata schema
 - RO Registry and LOCrates bundles: Local and integrated with EOSC
- **Containing full experiment (infrastructure) setup**
 - Components/nodes, parametrized infrastructure description and deployment sequence
 - Automation of deployment with tools: Ansible, Terraform, shell script, others
- **Experiment description and orchestration/workflow**
 - Jupyter Notebook, CWL/Galaxy, Github
 - Interactive Experiment configuration and management (web console and CLI)
- **Input/test data**
- **Data storage and preprocessing**
 - Data ingest link and API
 - Data model and interoperable/standard data format
 - **FAIR by design: primarily metadata management**
- **Measurement points and monitoring**

ACM Recommendations

1. **Repeatability:** *Same* team executes experiment using *same* setup
2. **Reproducibility:** *Different* team executes experiment using *same* setup
3. **Replicability:** *Different* team executes experiment using *different* setup

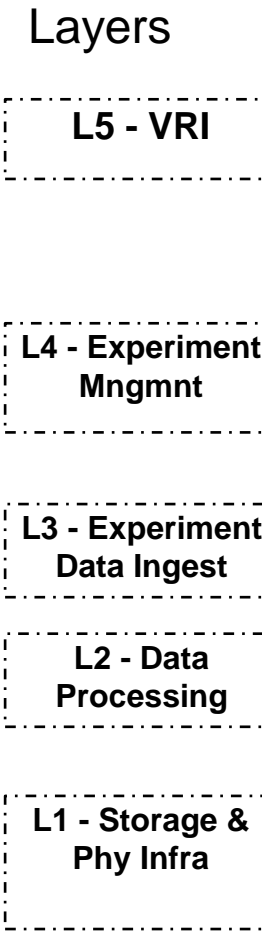
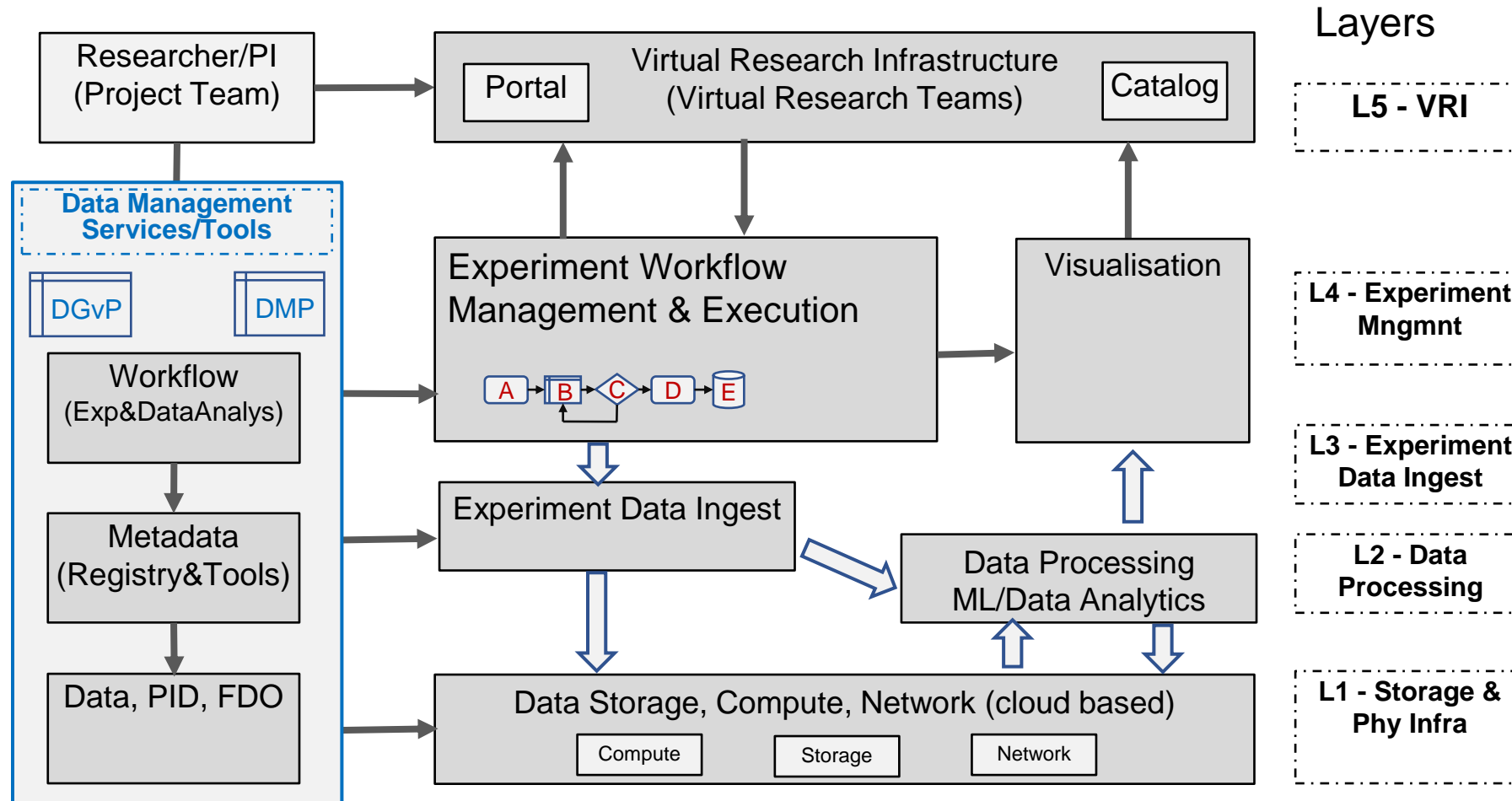


SLICES Experimental Data Lifecycle Model and Dataflow



- **Each Data Lifecycle stage** – experiment, data collection, data analysis, and finally data archiving, works with own **data set**, which must be **linked**.
 - All data sets need to be stored and possibly re-used in later processes.
- Many experiments and research require already existing datasets that will be available in SLICES data repositories or can be obtained/discovered in EOSC data repositories

Experimental Research Data Management Infrastructure



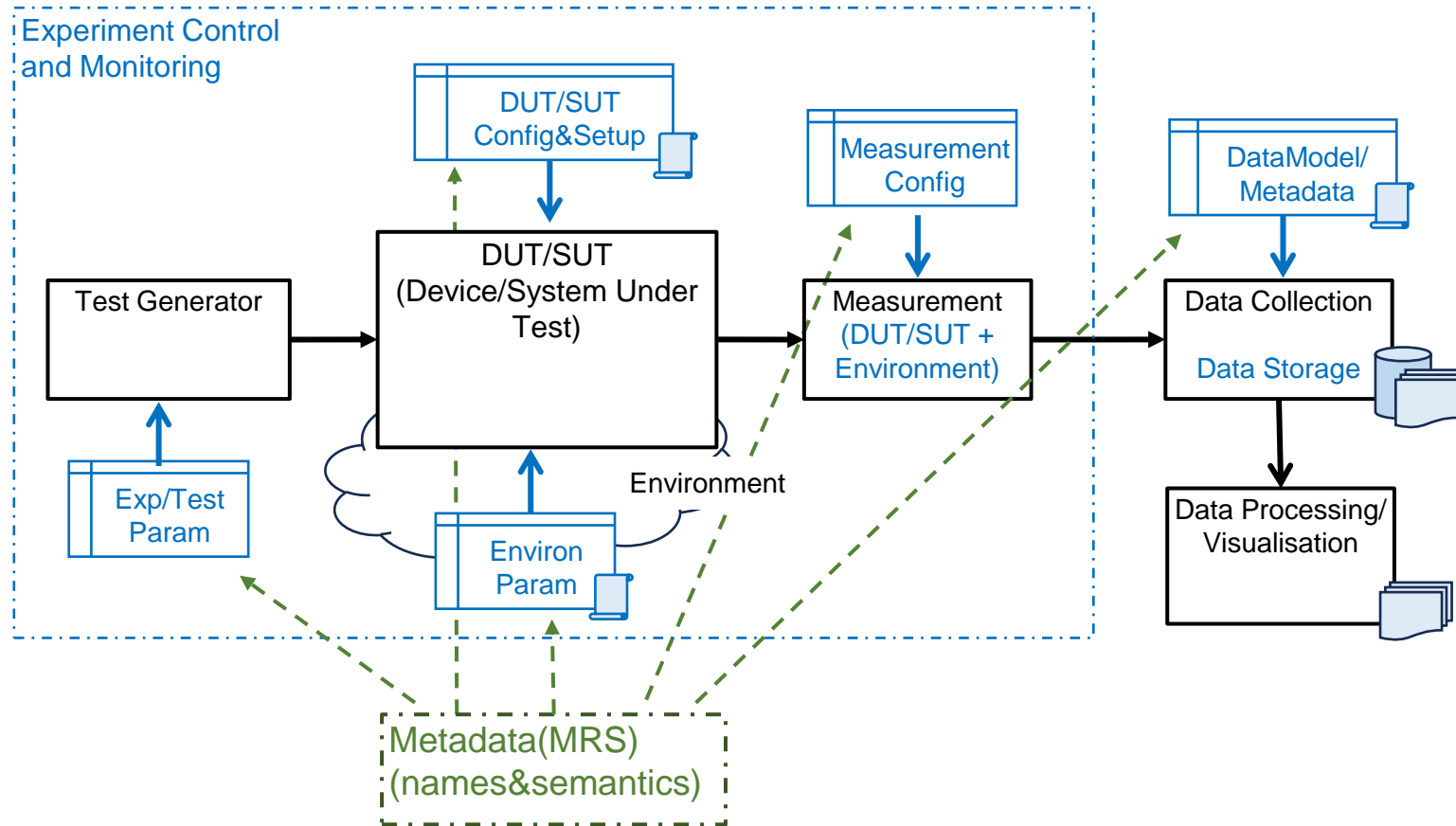
DMI layers

- Virtual RI and Researcher Portal
- Experiment Workflow Management
- Data Ingest
- Data Processing
- Storage & Physical Infrastructure

DGvP – Data Governance Policy DMP – Data Management Plan



Generic Experiment Model for Reproducibility



Questions to be answered before starting Experiment

- Device/System under Test model (variables, parameters, environment)
- DUT/SUT Configuration&Setup
- **DUT/SUT data model**
 - Relational model with multiple tables
- Test/Stimulus Variables& Parameters
- Measurement (instruments) configuration
- **Metadata defined and applied for all experiment components and stage**

Experiment Description: Metadata Requirements

- SLICES Data Management Infrastructure (DMI) Requirements groups - Part of SLICES DMI Blueprint
 - (1) Architecture and services
 - (2) General Metadata definition and management
 - (3) Experiment description and metadata
 - (4) Domain specific (e.g. SLICES Blueprint Architecture)
 - (5) Metadata Management tools

Existing practices

- Jupyter Notebook (Python based) – Popular but limited portability
- GitHub and GitHub Actions (CI/CD tools)
- Shell script
- Common Workflow Language (CWL)

What metadata should describe

- **Data models:** storage, databases, metadata
- **Experiment**
 - Orchestration; configuration; equipment: DUT, test generators, measurement; data storage; data models/metadata
- **Dataflow:** Stages, transformations, lineage/provenance, data models
- **Workflow:** Stages, Operations/conditions, workstations

New/emerging technologies and tools for data and metadata management

- EOSC Core Metadata Tools
- **Research Object (RO) and ROCrate** – Packaged information about and data from experiment – [Experimental Research Profile by SLICES](#)
 - [DVC for Experimental data versioning and lineage in complex data processing](#)
- EOSC Catalog – Data(set) and services registration
- FAIR Data Object (FDO) and PID for data publication and discovery
- Machine Actionable DMP (maDMP)
- [Metadata Tools for researchers:](#)
 - [Metadata Registry Service \(MRS\)](#) developed by SLICES/UCLan Cyprus
 - Metadata/data annotation, mapping and search
 - Namespace/semantics definition (SLICES namespace)
 - [Metadata extraction for experiment description \(legacy/unstructured format\)](#)

Discussion and Questions

- Open to cooperation with research and industry
- Pilot services planned for summer 2024

- Where to learn background information
 - EOSC (European Open Science Cloud) developments, services and products
 - RDA (Research Data Alliance) recommendations and best practices
 - FAIR expertise centers
 - Data services: datasets repositories and scientific data archives