



SLICES Data Management Infrastructure for Reproducible Experimental Research on Digital Technologies

Yuri Demchenko, University of Amsterdam

On behalf of

Yuri Demchenko, Sebastian Gallenmuller, Serge Fdida, Panayiotis Andreau, Damien
Sauzes, Thijs Rausch

WS20-1: Future G Experimental Test Platforms for Advanced Systems Implementations and
Research

IEEE GLOBECOM2023 Conference, 23 November 2023, Kuala Lumpur

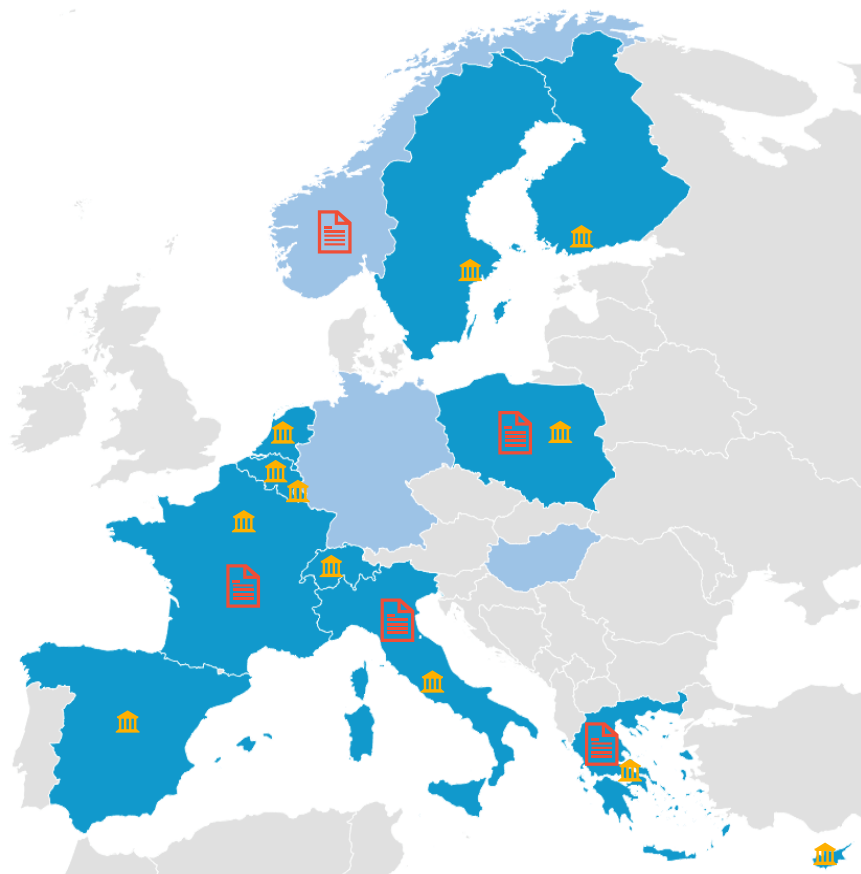
Outline

- 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
- Future developments
- Addendum: Metadata definition and Metadata Management





SLICES-RI for Research on Digital Infrastructures

Includes extensive experimentation with new technologies



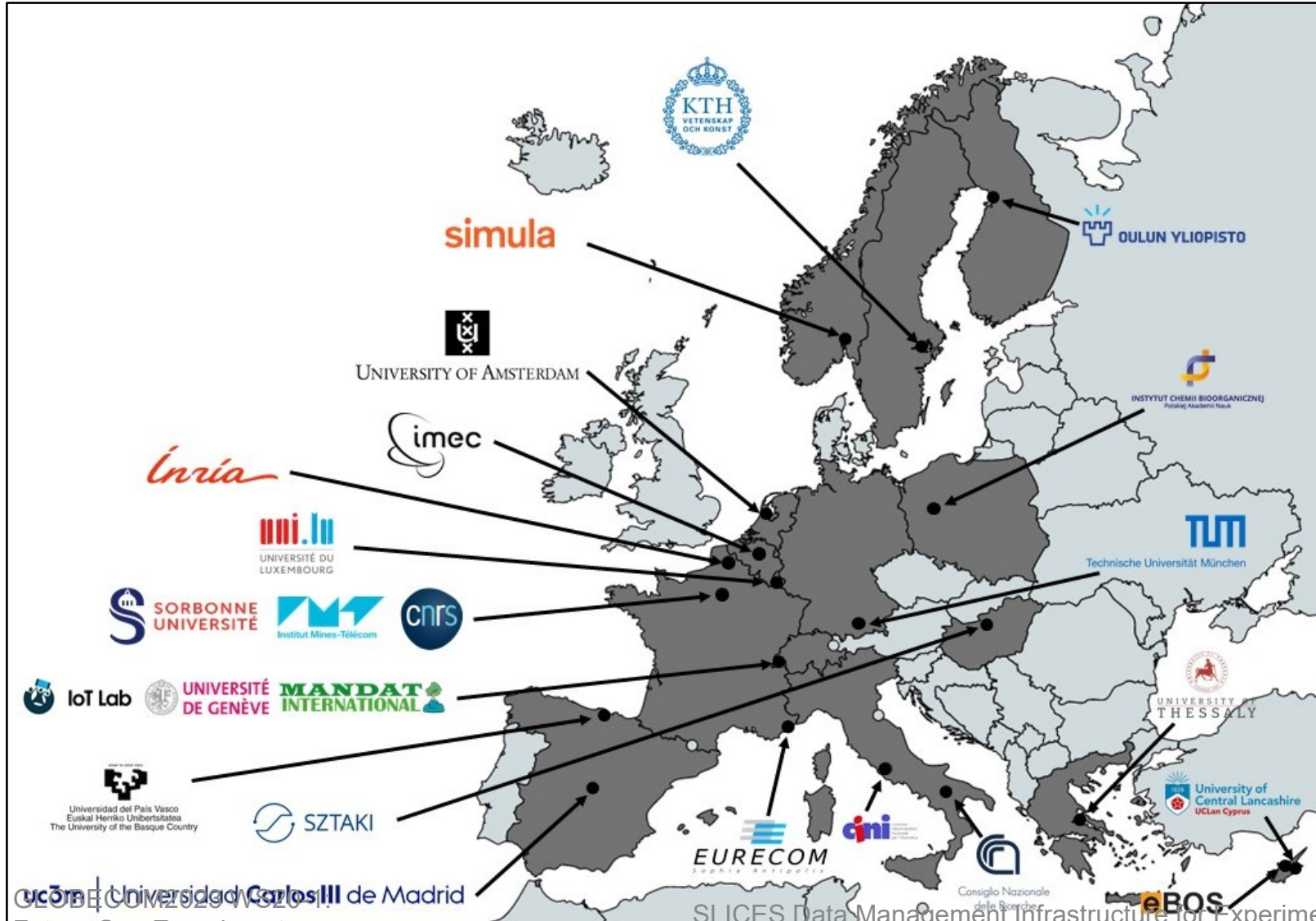
Initiated in 2017, **25 partners** from 15 countries:

- **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

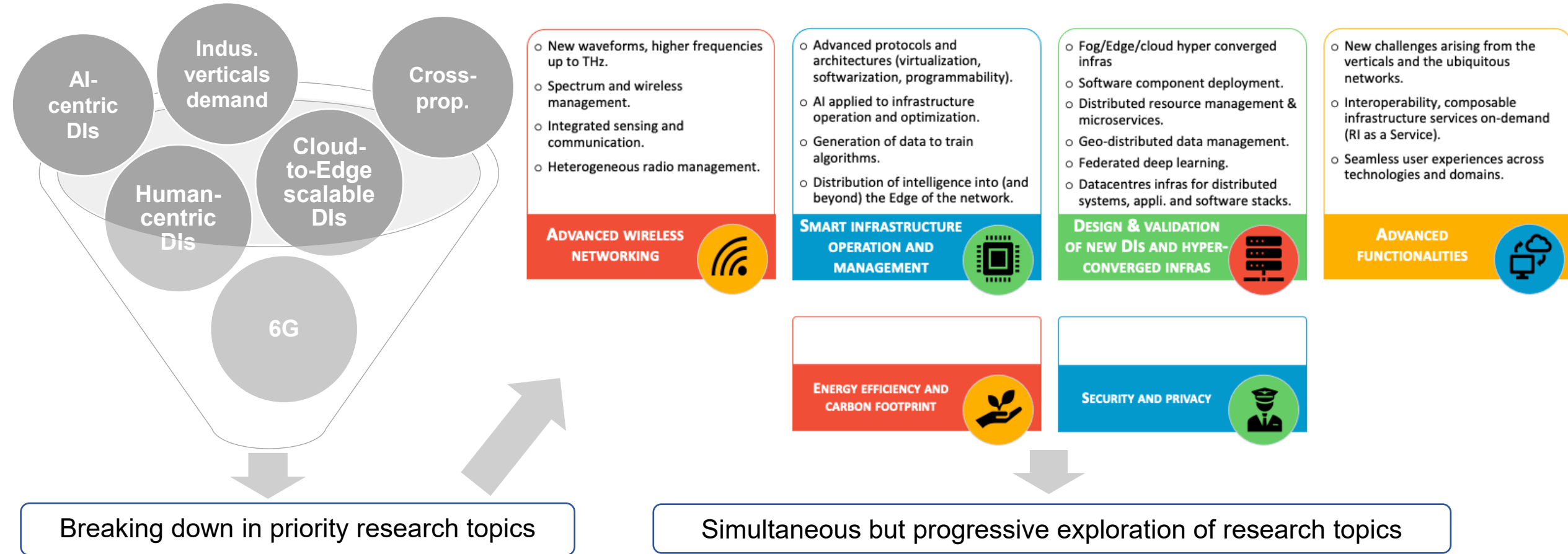


25 Partners from 15 countries

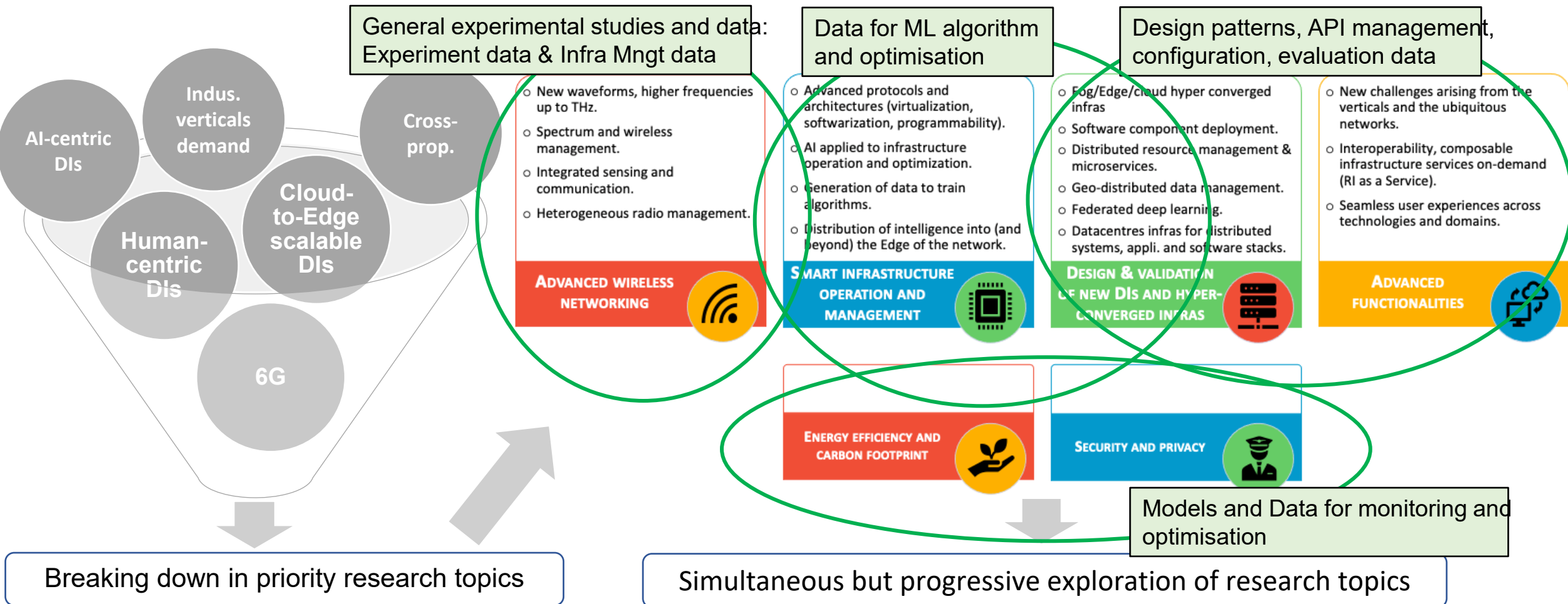
- INRIA, FR
- Sorbonne University (SU), FR
- **Univ of Amsterdam (uvA), NL**
- Univ of Thessaly (UTH), GR
- CNR, IT
- PSNC, PL
- Mandat International (MI), CH
- IoTLAB, FR
- UC3M, ES
- IMEC, BE
- UCLan, CY
- EURECOM, FR
- SZTAKI, HU
- CINI, IT
- CNIT, IT
- Univ Luxemburg, LU
- TUM, DE
- EHU, ES
- KTH,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

Wide range of research topics

What's the methodology behind it? What data produced and processes?



Different Types of Data for Different Experimental Studies



Variety of Data Produced in SLICES

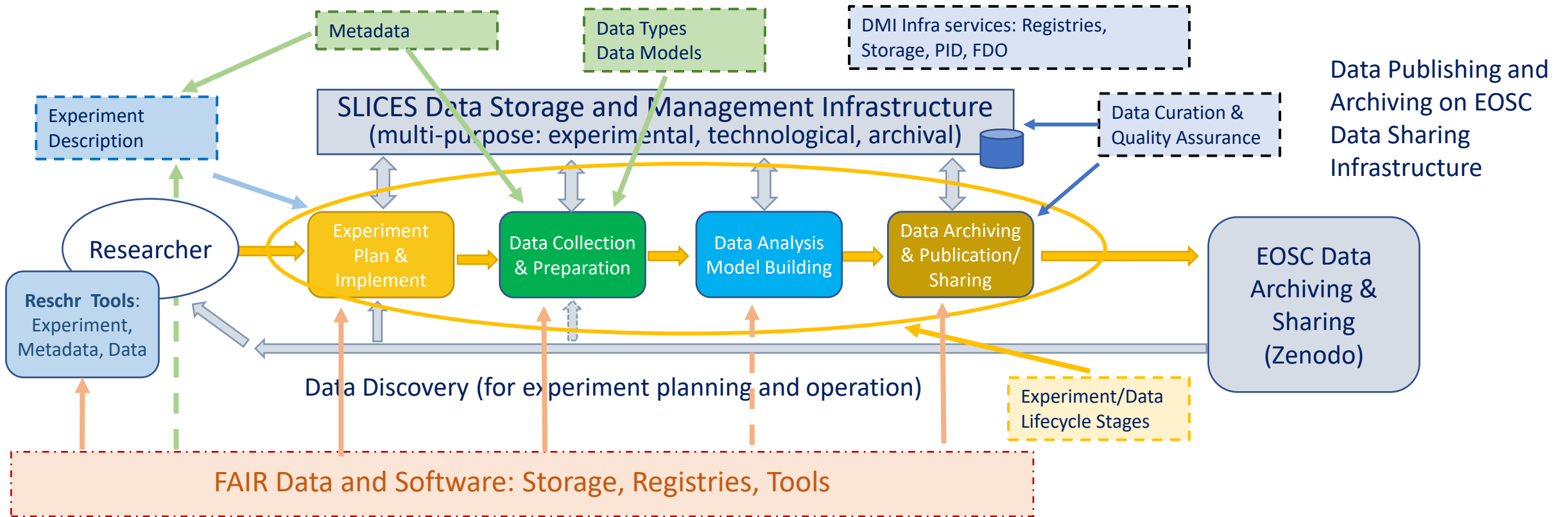
- General experimental studies and data documentation and publication
 - **FAIR (Findable, Accessible, Interoperable, Reusable)** data principles are key for experimental data sharing
 - **Metadata** profiles to be defined for major types of experiments and supported by data and metadata management tools
 - **Infrastructure management information** to be recorded as experiments environment
 - **Research Object (RO)** and FAIR Digital Object (FDO - being developed by EOSC)
- Data produced for and in the process of AI/ML model building - for smart infrastructure management and optimisation (including energy efficiency, performance, resilience)
 - Data modelling and data lineage (staging documenting)
 - AI/ML models serialization and portability
- New Digital Infrastructure architecture elements and design patterns
 - Infrastructure and design patterns + testing results
 - Metadata for API description, identification, composability

SLICES to provide the Robust Data Infrastructure for Experiment/Data Driven Research

- **Experimental data are big, distributed, domain specific, serving specific communities**
 - **Require effective models and infrastructure services for Research Data Management and secure data sharing**
- **Support the whole data lifecycle**
 - **Connected to research/experiment lifecycle or workflow**
- Distributed data storage and experimental data(set) repositories
 - Supporting recognized data interoperability standards (data formats and metadata)
 - Eventually certified: RDA endorsed Maturity and certification practice
 - **Interoperability and integration with EOSC as Federated data infrastructure**
- Data management and data curation and quality assurance
 - **FAIR data principles and SLICES metadata profiles (interoperable with EOSC)**
- Linked data and data discovery using semantic search and knowledge graph
 - **PID (Persistent IDentifier) and FDO (FAIR Digital Object) infrastructure (interoperable with EOSC)**
- (Trusted) Data exchange and secure transfer protocols

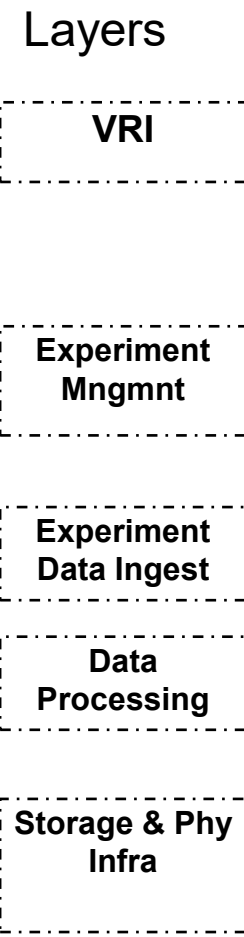
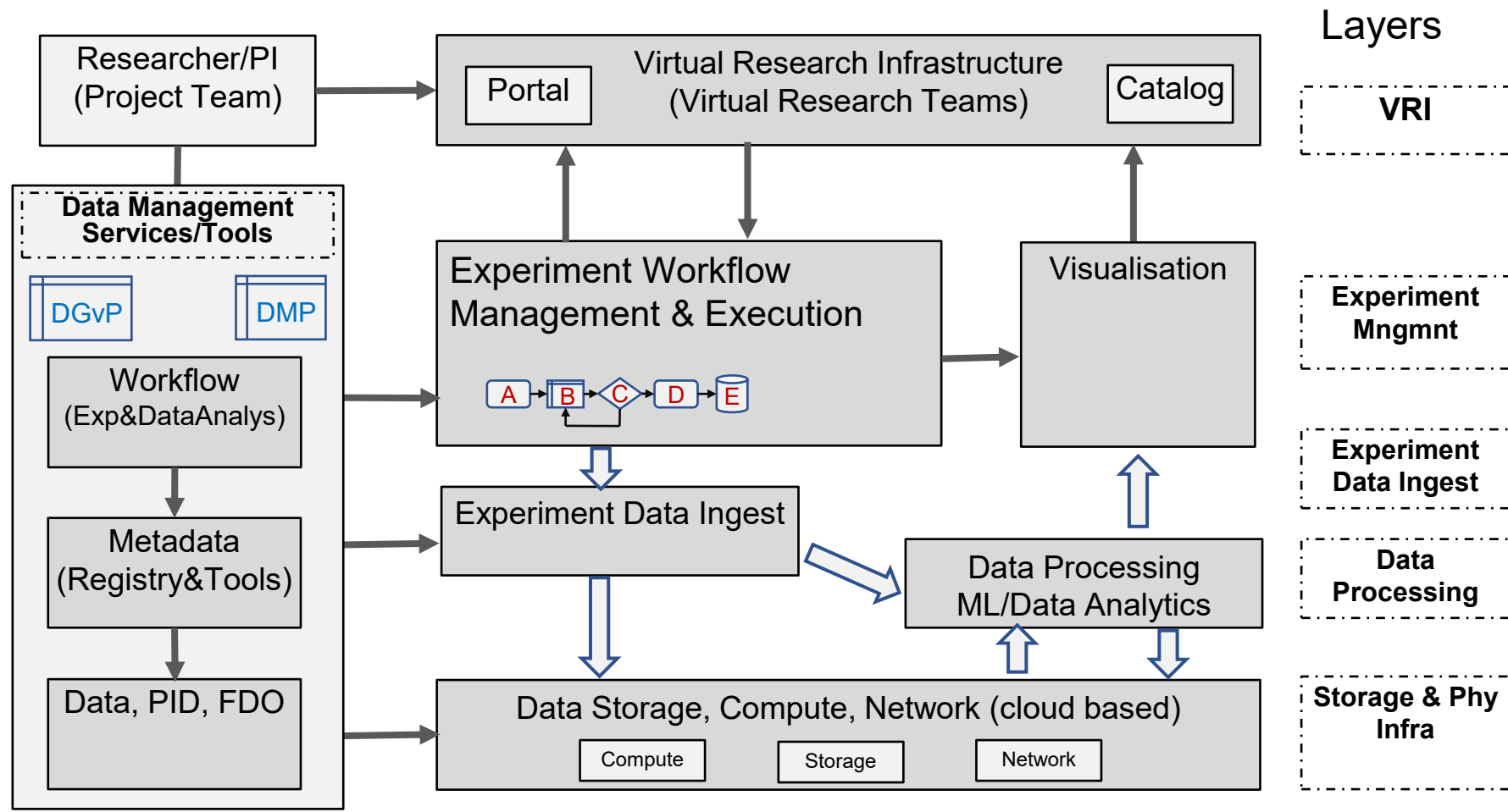


SLICES Experimental Data Lifecycle Model and Dataflow



- **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



Data Management Infrastructure Layers

Data Management Infrastructure Layers to separate data management and governance concerns and actors/roles

- Layer 4 - Experiment Infrastructure configuration and management
- Layer 3 - Experimental data collection/recording
 - Data models, metadata
- Layer 2 - Data processing
 - Data analysis, Process/ML models building, portability
- Layer 1 - Data Storage, Archiving, Exchange
 - Datasets, metadata publication
- Data Management Services and Tools (Data Management Plane)
 - Data Management Plan and Data Quality Assurance, FAIR compliance
 - Metadata registries and tools
 - Data Security and Data protection, GDPR

DMI services

- Metadata and Registries
- Data repositories & Data sharing
- ROCrates for Research publishing – Profile for ERRaaS
- FAIR Digital Object (FDO), PID registries and gateway/proxy – Integrated with EOSC
- Federated AAI – Integrated with EOSC



SLICES DMI Requirements (main)

- **RDM1.** **Distributed data storage and experimental data(set) repositories** should support common data and metadata interoperability standards, in particular, common data and metadata formats.
 - Outsourcing of data storage to the cloud must be protected with appropriate access control and compliant with the SLICES Data Management policies.
- **RDM2.** SLICES DMI should **support the whole research data lifecycle**. It should provide interfaces to experiment workflow and staging.
- **RDM3.** SLICES DMI shall provide PID (Persistent Identifier) and FDO (**FAIR** Digital Object) registration and resolution services to **support linked data and data discovery** that should be **integrated with EOSC** services.
- **RDM4.** SLICES DMI must support (trusted) **data exchange and transfer protocols** that allow policy-based access control to comply with the data protection regulations.
- **RDM5.** SLICES DMI must **enforce user and application access control** and identity management policies adopted by the SLICES community that can be potentially federated with the EOSC Federated AAI.
- **RDM6.** Procedures and policies must be implemented for **data curation and quality assurance**.
- **RDM7.** **Certification of data and metadata repositories** should be considered at some **maturity level** following certification and maturity recommendations by RDA.

FAIR from the technical point of view – Required Infrastructure functionality

- **Findable**
 - Metadata and PDI – infrastructure and tools
 - **Metadata Registries and handles resolution, API**
 - Policies and SLA
- **Accessible**
 - **Repositories and data storage: infrastructure and management**
 - Policy and access control: infrastructure and API management
 - Data access protocols
 - Usage Policy and Sovereignty
 - Data protection, compliance, privacy and GDPR
- **Interoperable**
 - Standard data formats
 - **Metadata Registries and API**
 - FAIR maturity level and certification
- **Reusable**
 - Data provenance and lineage
 - Preservation
 - **Metadata, PID and API – linked or embedded into datasets**

Require comprehensive **data infrastructure** to support

- **Data Storage**
- **Metadata Registries**
- Data publication
- Data discovery
- **Linked data and data lineage (provenance)**
- **Multiple datasets access for analysis**

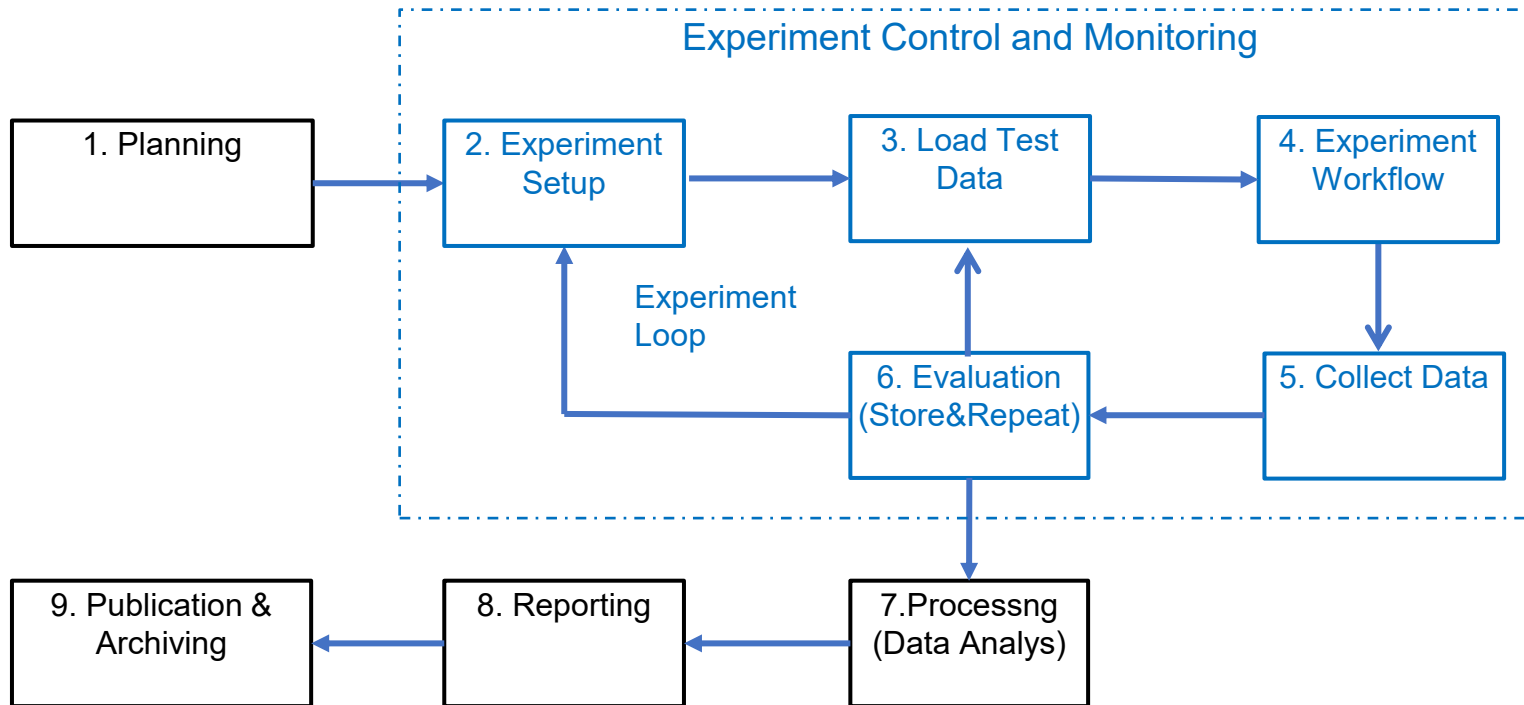
Experimental Research Reproducibility as a Service (ERRaaS)

- **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

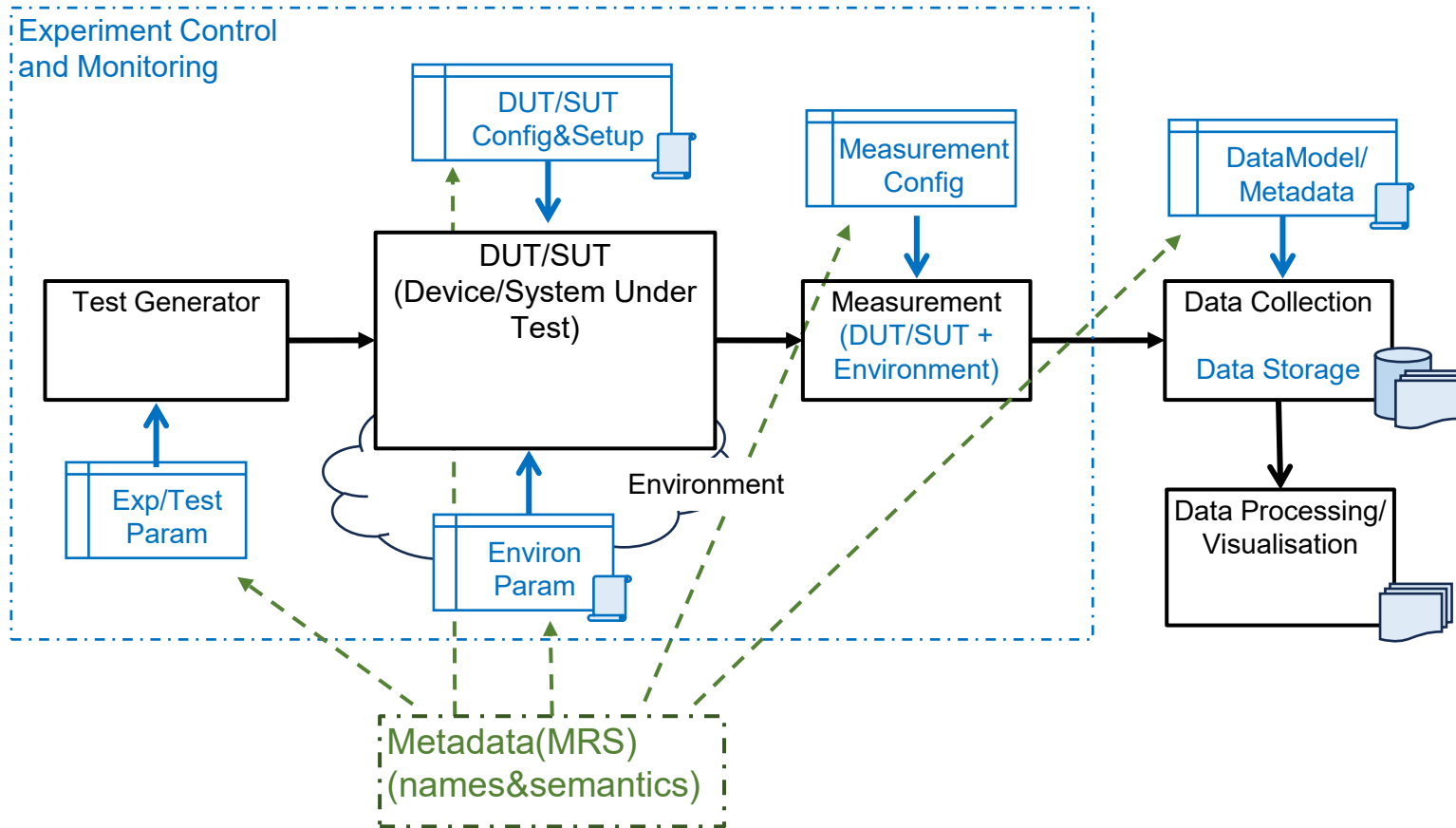
Experiment Workflow and Stages



Experimental research stages

1. Experiment Planning
2. Experiment setup, Equipment configuration
3. Load (test) data
4. Execute workflow
5. Collect data
6. Evaluate and re-run experiment if needed
7. Process/analyse data
8. Produce report
9. Archive/publish data

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)
- Common Workflow Language (CWL)

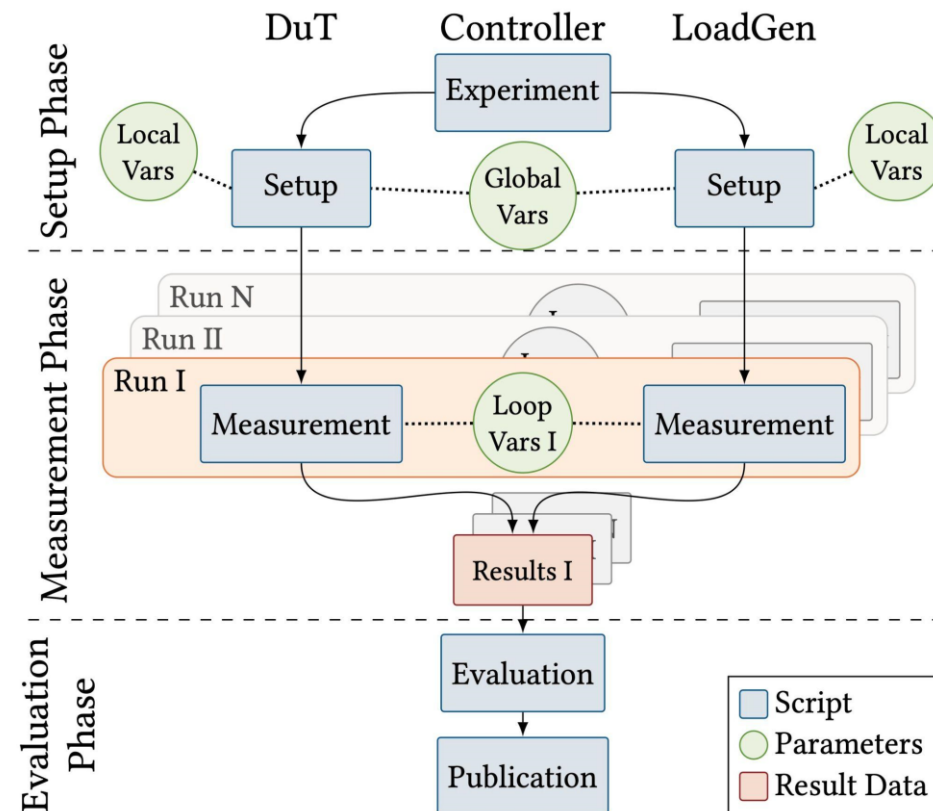
What metadata should describe

- **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

SLICES Platform for ERRaaS - Plain Orchestration Services (POS)

Development by the Technical University Munich (TUM)

- Setup phase
 - Controller manages experiment
 - Controller configures experiment nodes (DuT, LoadGen)
 - Global/local variables (vars) parametrize setup
- Measurement phase
 - Repeated execution of measurement script
 - Loop variables to parameterize each set of measurement run, e.g., changing packet rates data in each run is connected to a specific set of loop vars
- Evaluation phase
 - Collected results/loop vars used for experiment evaluation
 - Automated experiment release (git repository, website)



Structured Experiment Workflow with pos

Example: Pilot #1 with TUM on Metadata definition for POS experiments

EXPERIMENT SETUP

The task of this script is the initialization and preparation of the experiment execution. It is executed on the management host.

- Experiment script

GLOBAL AND LOOP PARAMETERS

List of parameters that were used for this instance of the experiment.

- Global parameters
- Loop parameters

LOAD GENERATOR

The task of this node is the setup and execution of the load generator creating the load for the device under test.

- Local parameters
- Setup script
- Measurement script

DEVICE UNDER TEST

The task of this node is the setup and execution of the investigated packet processing device.

- Local parameters
- Setup script
- Measurement script

EVALUATION

The evaluation script that plots the results.

- Evaluation script call

PUBLICATION

The publication script that created this website.

- Publication script call

[ref] https://gallenmu.github.io/pos-artifacts//web/2020-10-07_23-22-39_868017.html

Types of data in ICT Experiment operation

- **Variables**
- **Configuration (equipment, DUT)**
- Test data
- Orchestration and workflow
- Measurement data (data model, metadata)
- Storage
- Environment



SLICES Metadata Registry Service (MRS) by UCLan

- **Level 1: Domain-agnostic**

- SLICES core FAIR Digital Object, coined S-FDO
- Primary (e.g., persistent identification, description, resource type, creator), Management (e.g., version, metadata profile), Access(access type, access mode), Links, Languages, User Information, Rights (e.g., licence)
- Implemented at the SLICES Central Hub (i.e., SLICES portal) and exposed to all SLICES Nodes and their Sites metadata profile),

- **Level 2: Type-specific**

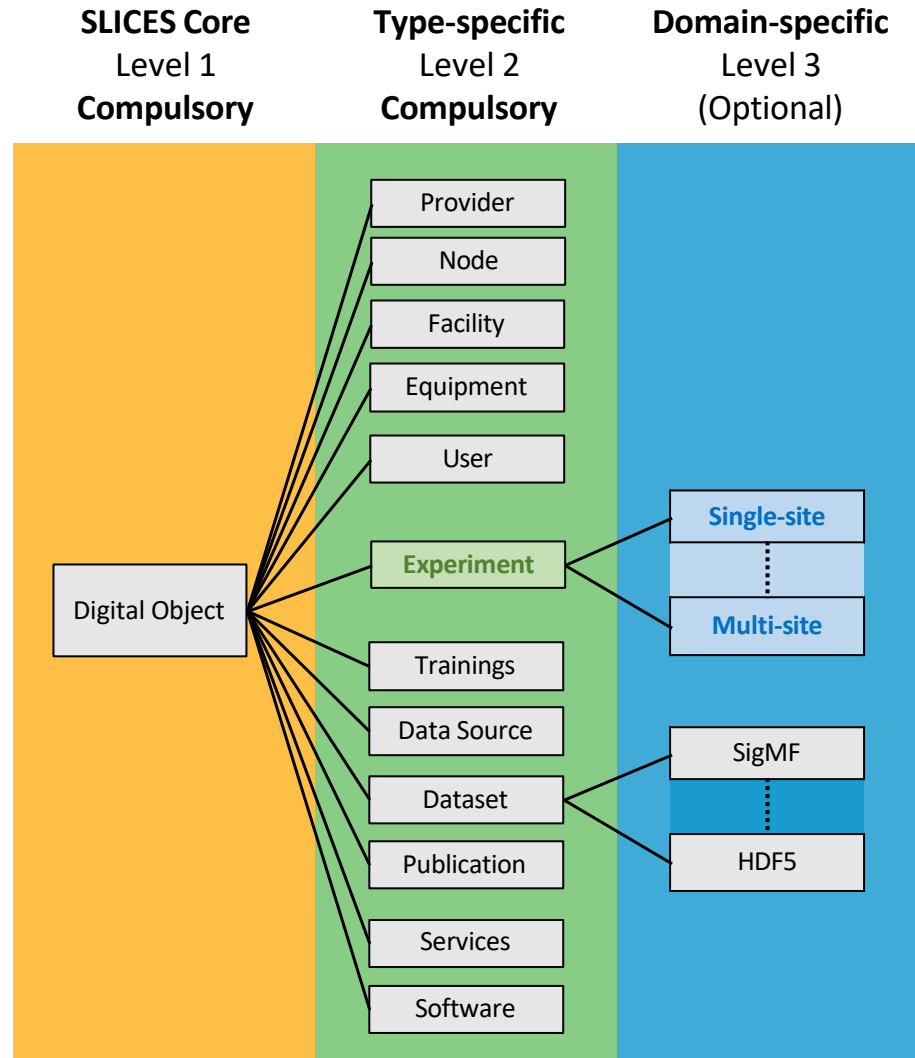
- Extends basic metadata with type-specific metadata
- Experiments have experiment description, classification and workflows
- Primary objective: Human/machine readability, machine-actionability, Easily expandable with new types

- **Level 3: Domain-specific**

- Extends type-specific metadata with domain-specific metadata
- Primary objective: Support specialized domains without polluting the data with attributes that are not applicable to most domains.
- Specialized human/machine readability, machine-actionability

- **Level 2 & Level 3 metadata**

- Stored within SFDOs at central hub
- Base APIs can use SFDOs directly
- Domain-specific APIs may implement orchestration/execution of SFDOs differently



New/emerging technologies and tools for metadata management

- EOSC Core Metadata Tools
- **Research Object (RO) and ROCrate** – Packaged information about and data from experiment – [Experimental Research Profile by SLICES](#)
- 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 \(legacy/unstructured format\)](#)



FAIR Core for EOSC – Tools First Release Announce 2023

1. **EOSC Research Discovery Graph (RDGraph)** to deliver advanced Discovery tools across EOSC resources and communities;
2. **EOSC PID Graph (PIDGraph)** to improve the way of interlinking research entities across domains and data sources on the basis of persistent identifiers (PIDs);
3. **EOSC Metadata Schema and Crosswalk Registry (MSCR)** to support publishing, Discovery and access of metadata schemas and provide functions to operationalize metadata conversions by combining crosswalks;
4. **EOSC Data Type Registry (DTR)** to provide user friendly APIs for metadata imports and access to different data types and metadata mappings;
5. **EOSC PID Meta Resolver (PIDMR)** to offer users a single PID resolving API in which any kind of PID can be resolved through a single, scalable PID resolving infrastructure;
6. **EOSC Compliance Assessment Toolkit (CAT)** to support the EOSC PID policy compliance and implementation;
7. **EOSC Research Activity Identifier Service (RAiD)** to mint PIDs for research projects, allowing to manage and track project related activities;
8. **EOSC Research Software APIs and Connectors (RSAC)** to ensure the long-term preservation of research software in different disciplines;
9. **EOSC Software Heritage Mirror (SWHM)** to equip EOSC with a mirror of the Software Heritage universal source code archive.



ROCrates and Extension for Experiment Description

- SLICES will adopt the RO and RO-Crate frameworks for packaging and managing experimental research products and documenting their evolution (provenance)
 - SLICES will follow the formal RO-Crate procedure to create a new ROE-Crate profile/schema for Experimental RO
 - To support all necessary information required for the [full experiment description and reproducibility](#).
- RO-Crate Specification Version 1.1 allows packaging the following information/entities – all entities can be local or linked.
 - Metadata, workflows, software, models, data, publications, presentations, metadata, logs
- A resource is stored using RO-Crate directory with the following structure:
 - `<RO-Crate root-directory>/`
 - | `ro-crate-metadata.json` # Metadata file **MUST**
 - | `ro-crate-preview.html` # RO-Crate website **MAY**
 - | `ro-crate-preview_files/` # **MAY** be present
 - | | [other RO-Crate website files])
 - | [payload files and directories] # 0 or more
- The metadata file uses JSON format for Linked Data (JSON-LD) and provides information about all entities included in the RO-Crate.
 - Context attributes of the data entity can be used to document equipment or software used to create files, but the description is limited to textual description and serial number.
- Provenance information is limited to CreateAction and UpdateAction attributes of the data entity.

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