SLICES Data Management infrastructure services for Experimental Research Reproducibility

SLICES Summer School

13-15 June 2023, Oulu, Finland

SLICES Academy

Outline

- SLICES Initiative on Experimental Research Automation and Reproducibility
    - Reproducible Experimental Research as a Service
- Elements of the Experimental Research Reproducibility
    - Data types produced in SLICES
- FAIR data principles and Metadata Management
- (Prospective) SLICES Data Management Infrastructure

Workshop Materials

- https://drive.google.com/drive/folders/1mfoZs3OXOx_Klhy1r6-YVXlW_4MtadFh?usp=sharing

SLICES-RI Experimental Facilities and Testbeds

- OneLab: Cloud Infrastructure for Researchers (LIP6, Sorbonne University)
- 5TONIC Lab (Uni Carlos III of Madrid)
- NITOS testbed 5G (University of Thessaly)
- Open5G Lab, SOPHIA-NODE: Beyond-5G cloud-native network (INRIA)

- imec testbed for networking, cloud, AI and IoT research (Ghent Uni)
- LeonR&Do Lab (COSMOTE, GR)
- SN4I Lab Smart Network for Industry 4.0 (Uni Basque Country)
- IoT Lab (Mandat International, CH)

Open Science Challenges in Experimental Studies

- **SLICES is intended to support large-scale experimental studies on modern/future Digital Infrastructure technologies**
  - **Multi-site, multi-scale, cross-domain, federated, experiment driven, researcher/user centric**
- Scientific value of experimental research is in the reproducibility of experiments, sharing and (re)usability of data
- SLICES-RI adds its specifics of implementing **Open Science** and **FAIR** (Findable – Accessible - Interoperable – Reusable) data principles for experimental studies on the Digital Infrastructure technologies
- (Also) Important questions in experimenting with new technologies and cooperation with industry is how open research and experimental data should be
  - IPR and industrial KnowHow must be protected by Data Governance policies and enforcement
  - General infrastructure management data must be handled with responsibility
  - Compliance with the European Cybersecurity Assurance Act to be considered
  - Compliance with GDPR

Experimental Research Reproducibility as a Service

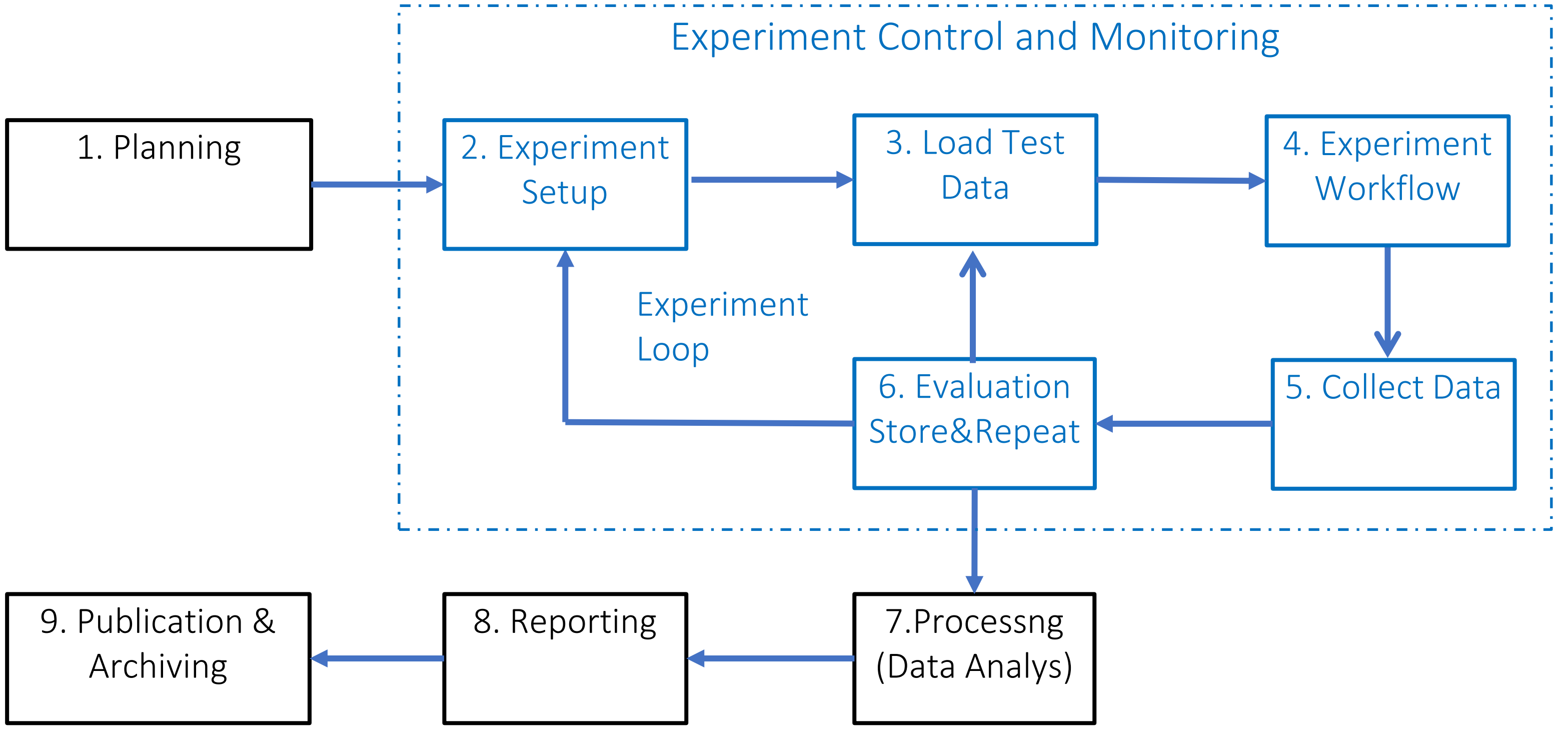
- SLICES to support experiments reproducibility to comply with Open Science
  - Focus on **repeatability** and **reproducibility** with the future support of **replicability**
- Robust, reproducible experiments
  - Documenting all relevant parameters and environment for experiments
  - Automate the documentation of experiments
  - ➢ **Well-structured experiment workflow may serve as documentation**
- Benefits for research community
  - Reduce amount of work for experimenters to create reproducible experiments
  - Reduce amount of work for other researchers to recreate and re-run experiments
  - Make reproducibility an integral part of experiment design
  - ➢ **Automate entire experiment (setup, execution, evaluation)**

Experimental research stages

- Experiment Planning
- Experiment setup, Equipment configuration
- Load (test) data
- Execute workflow
- Collect data
- **Evaluate and re-run experiment if needed**
- Process/analyse data
- Produce report
- Archive/publish data

slicesPP

Experiment Workflow and Stages



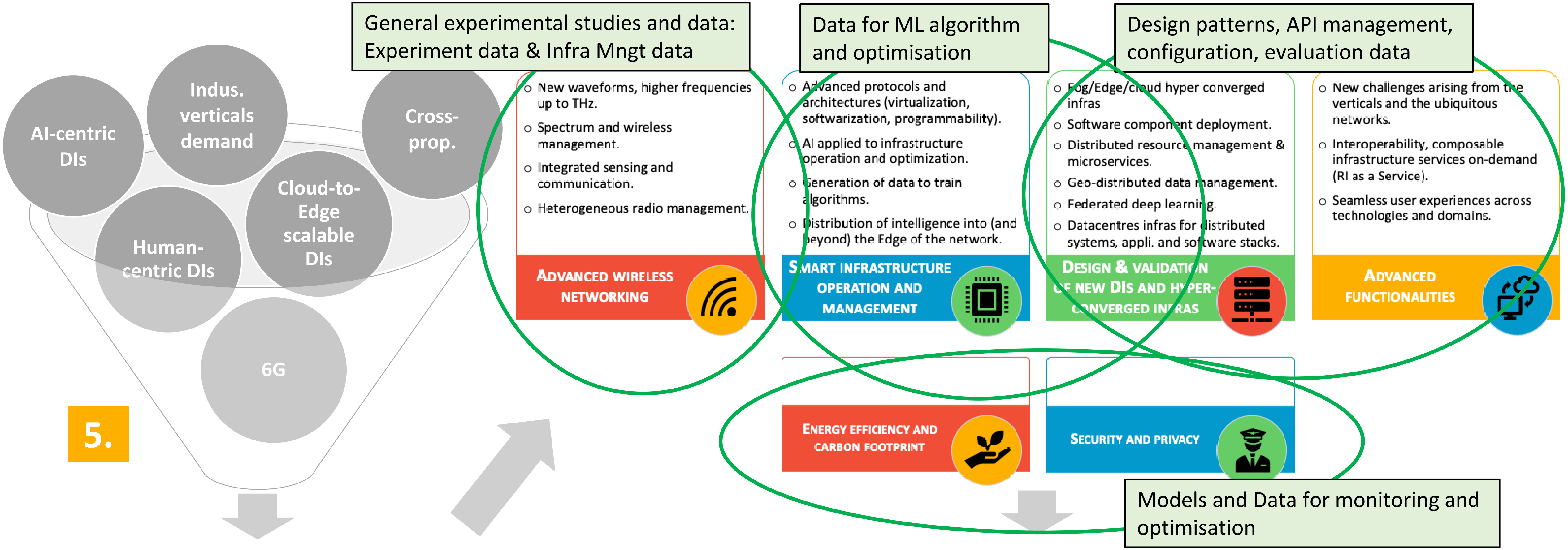
Experimental research stages

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

Plain Orchestration Service (POS) by TU Munich

- The plain orchestrating service (pos) provides two components:
  - Testbed controller and Experiment workflow
- The testbed controller takes care of the allocation and management of experimental resources
  - It provides bare-metal access to the experiment nodes
  - Images for the experiment nodes are provided in the form of live Linux images
- Using **live images** for experiments has two benefits:
  - First, rebooting an experiment node helps reset the system to a well-defined state.
  - Second, testbed users are aware of the non-permanence of their configuration, gently pushing users towards documenting and automating experiment configuration.

Different Types of Data for Different Experimental Studies



General experimental studies and data: Experiment data & Infra Mngt data

Data for ML algorithm and optimisation

Design patterns, API management, configuration, evaluation data

ADVANCED WIRELESS NETWORKING

- New waveforms, higher frequencies up to THz.
- Spectrum and wireless management.
- Integrated sensing and communication.
- Heterogeneous radio management.

SMART INFRASTRUCTURE OPERATION AND MANAGEMENT

- Advanced protocols and architectures (virtualization, softwarization, programmability).
- AI applied to infrastructure operation and optimization.
- Generation of data to train algorithms.
- Distribution of intelligence into (and beyond) the Edge of the network.

DESIGN & VALIDATION OF NEW DIS AND HYPER-CONVERGED INFRAS

- Fog/Edge/cloud hyper converged infras
- Software component deployment.
- Distributed resource management & microservices.
- Geo-distributed data management.
- Federated deep learning.
- Datacentres infras for distributed systems, appli. and software stacks.

ADVANCED FUNCTIONALITIES

- New challenges arising from the verticals and the ubiquitous networks.
- Interoperability, composable infrastructure services on-demand (RI as a Service).
- Seamless user experiences across technologies and domains.

ENERGY EFFICIENCY AND CARBON FOOTPRINT

SECURITY AND PRIVACY

Models and Data for monitoring and optimisation

Breaking down in priority research topics

Simultaneous but progressive exploration of research topics

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 (being developed by EOSC)
- Data produced for AI/ML algorithms training for smart infrastructure optimisation and management (including energy efficiency, performance, resilience, sustainability)
  - 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
  - Metadata for API description, identification, composability

slicesPP

FAIR Data Principles are realized via Metadata Management (GO FAIR recommendations)

Findable:

- F1 (meta)data are assigned a **globally unique and persistent identifier**;
- F2 data are **described with rich metadata**;
- F3 metadata clearly and explicitly include the **identifier** of the data it describes;
- F4 (meta)data are **registered or indexed** in a searchable resource;

Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles;
- I3. (meta)data include qualified references to other (meta)data;

Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
  - A1.1 the protocol is open, free, and universally implementable;
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary;
- A2 metadata are accessible, even when the data are no longer available;

Reusable:

- R1 meta(data) are richly described with a plurality of accurate and relevant attributes;
- R1.1 (meta)data are released with a clear and accessible data usage license;
- R1.2 (meta)data are associated with detailed provenance;
- R1.3 (meta)data meet domain-relevant community standards;

slicesPP

FAIR from the technical point of view

- Findable
  - Metadata and PDI – infrastructure and tools
  - 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 and Registries**
- Data publication
- Data discovery
- Linked data and data lineage (provenance)
- Multiple datasets access for analysis

Research Objects for Metadata definition

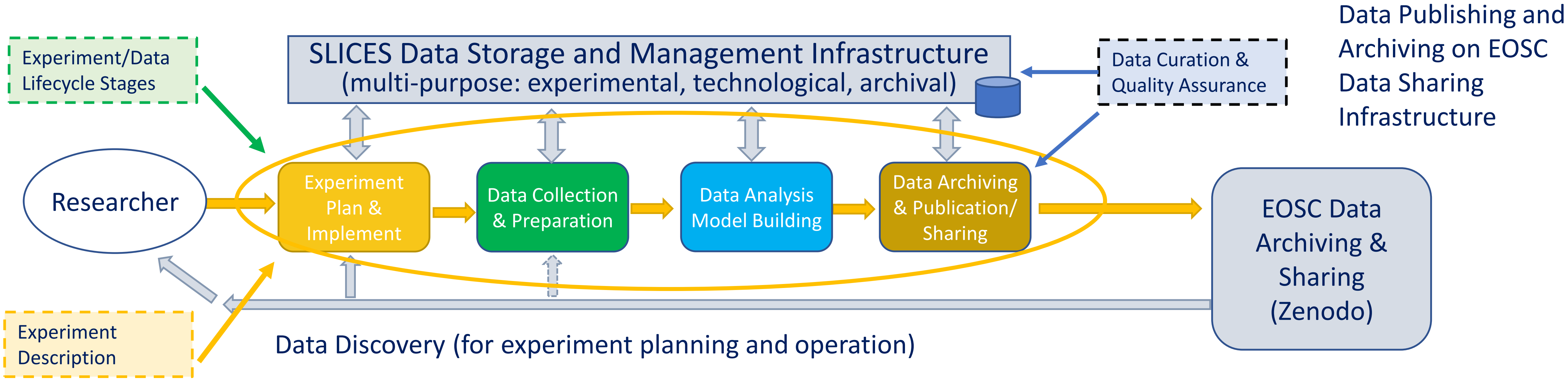
- Data sets: datamodels/schemas, databases, storage
- Experiment
  - Orchestration; configuration; equipment: DUT, test generators, measurement; data storage; data models/metadata
- Workflow: Stages, Operations/conditions, workstations
- Dataflow: Stages, transformations, lineage/provenance, data models

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 a European 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

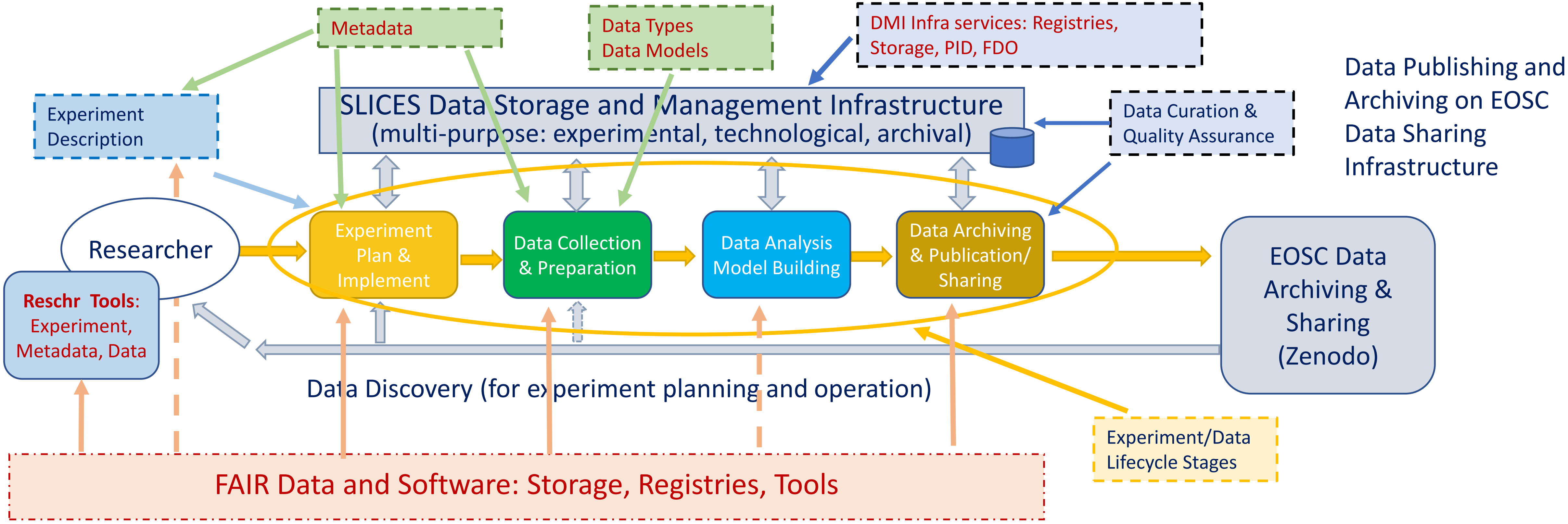
slicesPP

SLICES Experimental Data Lifecycle Model and Dataflow


Experiment/Data Lifecycle Stages
SLICES Data Storage and Management Infrastructure (multi-purpose: experimental, technological, archival)
Data Curation & Quality Assurance
Data Publishing and Archiving on EOSC Data Sharing Infrastructure
Researcher
Experiment Plan & Implement
Data Collection & Preparation
Data Analysis Model Building
Data Archiving & Publication/ Sharing
EOSC Data Archiving & Sharing (Zenodo)
Experiment Description
Data Discovery (for experiment planning and operation)

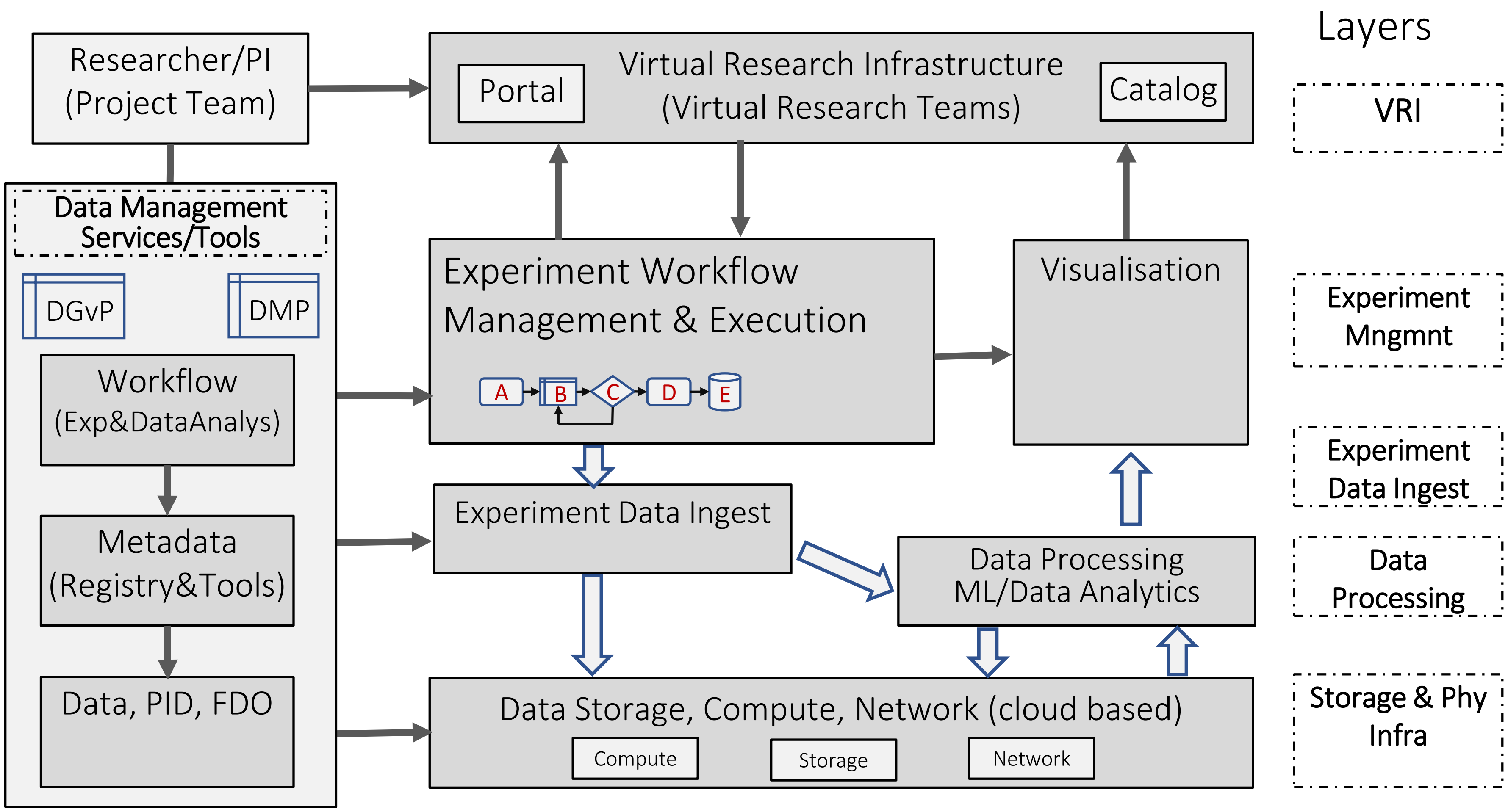
- **Each Data Lifecycle stage** – experiment, data collection, data analysis, and finally data archiving, works with own **data set,** which must be **linked**.
  - All data sets need to be stored and possibly re-used in later processes.
- Many experiments and research require already existing datasets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

SLICES Experimental Data Lifecycle Model and Dataflow


Metadata
Data Types
Data Models
DMI Infra services: Registries, Storage, PID, FDO
Data Publishing and Archiving on EOSC Data Sharing Infrastructure
Experiment Description
SLICES Data Storage and Management Infrastructure
(multi-purpose: experimental, technological, archival)
Data Curation & Quality Assurance
Researcher
Experiment Plan & Implement
Data Collection & Preparation
Data Analysis Model Building
Data Archiving & Publication/ Sharing
EOSC Data Archiving & Sharing (Zenodo)
Reschr Tools: Experiment, Metadata, Data
Data Discovery (for experiment planning and operation)
Experiment/Data Lifecycle Stages
FAIR Data and Software: Storage, Registries, Tools

- **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 Data Management Infrastructure



Layers
Researcher/PI (Project Team)
Virtual Research Infrastructure (Virtual Research Teams)
Portal
Catalog
VRI
Data Management Services/Tools
DGvP
DMP
Workflow (Exp&DataAnalys)
Metadata (Registry&Tools)
Data, PID, FDO
Experiment Workflow Management & Execution
A
B
C
D
E
Visualisation
Experiment Mngmnt
Experiment Data Ingest
Experiment Data Ingest
Data Processing ML/Data Analytics
Data Processing
Data Storage, Compute, Network (cloud based)
Compute
Storage
Network
Storage & Phy Infra

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
  - FAIR Digital Object (FDO), PID registries and gateway/proxy
- 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, Access control, GDPR

New/emerging technologies to consider

- FAIR Data Object (FDO) => SLICES FDO (SFDO)
- Research Object (RO) => Experimental RO (ExRO)
- EOSC Catalog => SLICES Federated Catalog (federated with EOSC)
- PID => SLICES subdomain/SLICES Data Space
- Machine Actionable DMP (maDMP)

Additional information

FAIR is an Overloaded Concept

Findable – Accessible – Interoperable - Reusable

- Primarily, FAIR is (set of) principles for sustainable Research Data Management (RDM) and Open Science
  - Findable – Accessible – Interoperable – Reusable
- FAIR is an initiative
- FAIR is a key policy area of EOSC
- FAIR data management is part of Data Management Plan (DMP) and required by Horizon Europe and many national funding bodies
- FAIR impose a number of requirements to Research Infrastructure
- Existing RIs run dedicated projects on FAIR adoption: ENVRI-FAIR, ELIXIR

- Universities should play important role in FAIR and RDM adoption
  - Still slow adoption at all levels: Bachelor, Master, Doctoral, teachers

EOSC and RDA activities of interest to SLICES-RI

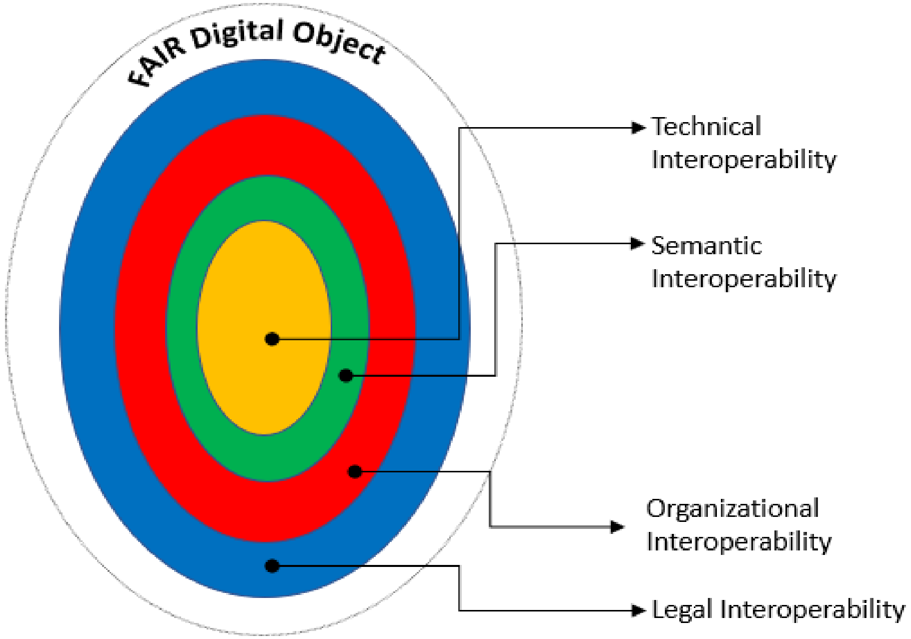
- RDA FAIR maturity WG
  - https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
- RDA FAIR for AI
- RDA machine actionable DMP
  - https://www.rd-alliance.org/groups/active-data-management-plans.html
- RDA Computational reproducibility
  - https://www.rd-alliance.org/computational-reproducibility-what%E2%80%99s-next-rda
- RDA FAIR Digital Object (FDO) and PID infrastructure
  - Special SLICES-PP WP7 focus on the topic of Data Management Infrastructure elements
- EOSC Registry and EOSC Services Portal
- EOSC Interoperability Framework
- EOSC semantic interoperability and Research Object

EOSC-IF is about Data – To support FAIR data sharing

- FAIR Digital Object (FDO) is a key concept
- Technical Interoperability:
  - Artefact Common Protocols and Data Formats
- Semantic Interoperability:
  - Contextual Semantics related to Common Semantic resources
- Organisational Interoperability:
  - Contextual Semantics related to Common process resources
- Legal Interoperability:
  - Contextual licenses related to Common Licenses resources

- FDO is actively promoted by GO FAIR Technical Center and Peter Wittenburg
  - Recent presentation at e-IRG meeting (e-Infrastructure Reflection Group - EC policy consulting body)

FAIR Digital Object – A core for EOSC-IF


FAIR Digital Object
Technical Interoperability
Semantic Interoperability
Organizational Interoperability
Legal Interoperability

- In EOSC, a digital object can be research data, software, scientific workflows, hardware designs, protocols, provenance logs, publications, presentations, etc
- FAIR Digital Object Extends Digital Object concept for better FAIRness
- FAIR Digital Object (FDO) is a core building block of EOSC-IF
  - Four interoperability layers applied
  - Requires infrastructure support

FDO (FAIR Digital Object) and PID Infrastructure Requirements

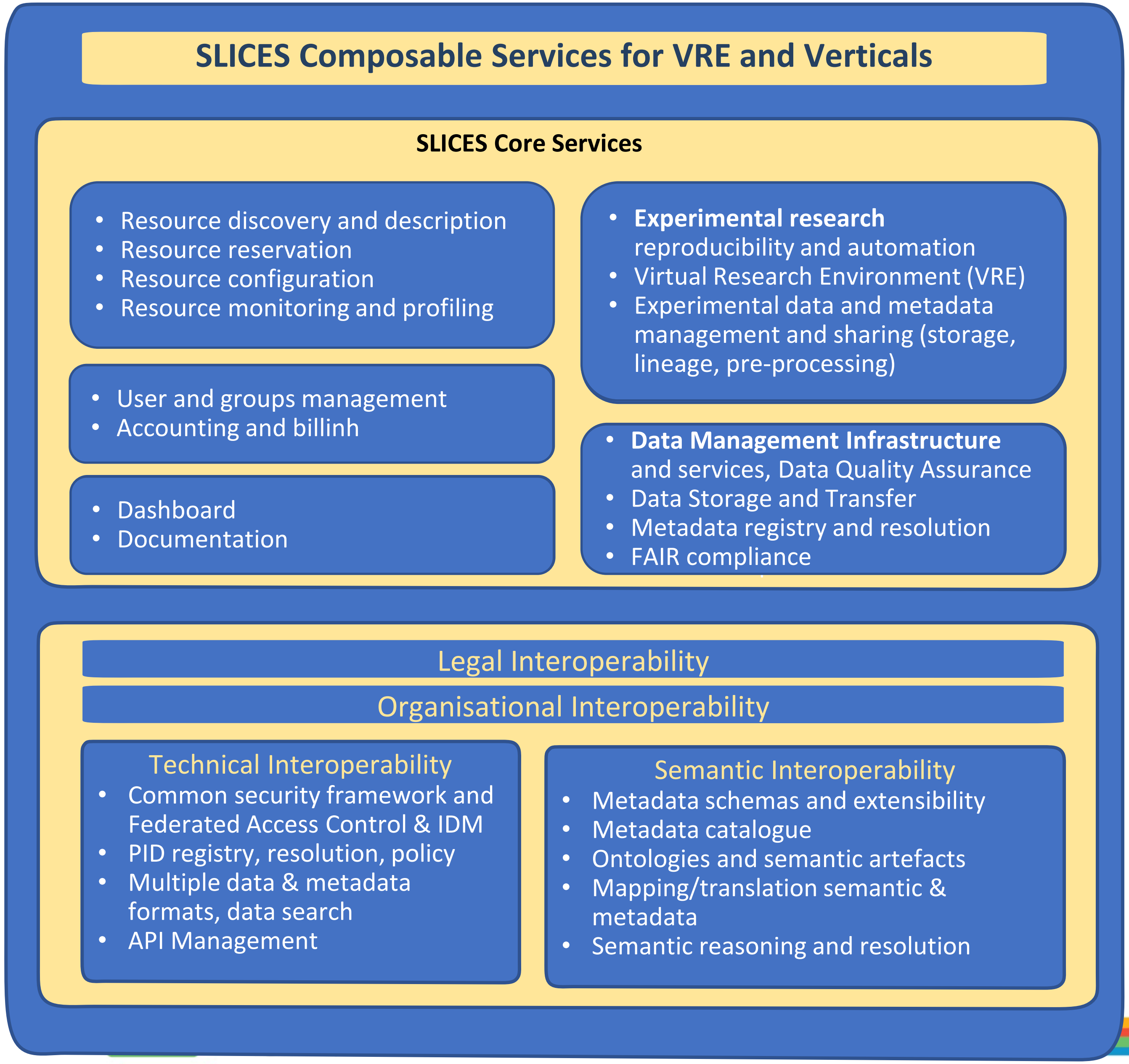
- General requirements include **machine actionability**, technology independence, **persistent binding,** abstraction and structured hierarchical encapsulation, compliance with standards and community policies (as specified in the FDO general requirements G3-G9);
- ***FDO is identified by PID***; there are possible multiple PID frameworks defined by PDI scheme, namespaces, ontologies or controlled vocabularies (FDOF1);
- A **PID resolves to a structured record (PID record)** with attributes that are semantically defined within a (data) type ontology (which may be defined for different application or science domains) (FDOF2);
- PID record may include other attributes that are important to characterize specific types of FDO or that are required by applications. Additional attributes must be registered in a ***data types registry*** (FDOF4);
- Metadata used to describe FDO properties should use standard semantics and ***registered schemes*** to allow machine readability and actionability (FDOF8-FDOF10).

- FAIR Digital Object Framework, Technical implementation guideline, version 1.02, [online] https://datashare.rzg.mpg.de/s/RTeYZGe3QMgEciH/download?path=%2F&files=FAIR%20Digital%20Object%20Framework-v1-02.pdf

Interoperability and integration with EOSC

- Registration SLICES Provider Profile finalized – INRIA is assigned as admin/legal contact
- SLICES Interoperability Framework
  - Continue work started in SLICES-DS (design templates definition with Ansible, Terraform)
  - Define core API for example testbeds/experimental facilities
- Working contacts between EOSC & EGI established but need to be formalized
  - Learning from EGI experience in building their distributed infrastructure
  - Communicating to EGI specific requirements from SLICES

Conceptual view of SLICES Interoperability Architecture



SLICES Composable Services for VRE and Verticals

SLICES Core Services

- Resource discovery and description
- Resource reservation
- Resource configuration
- Resource monitoring and profiling

- User and groups management
- Accounting and billinh

- Dashboard
- Documentation

- **Experimental research** reproducibility and automation
- Virtual Research Environment (VRE)
- Experimental data and metadata management and sharing (storage, lineage, pre-processing)

- **Data Management Infrastructure** and services, Data Quality Assurance
- Data Storage and Transfer
- Metadata registry and resolution
- FAIR compliance

Legal Interoperability

Organisational Interoperability

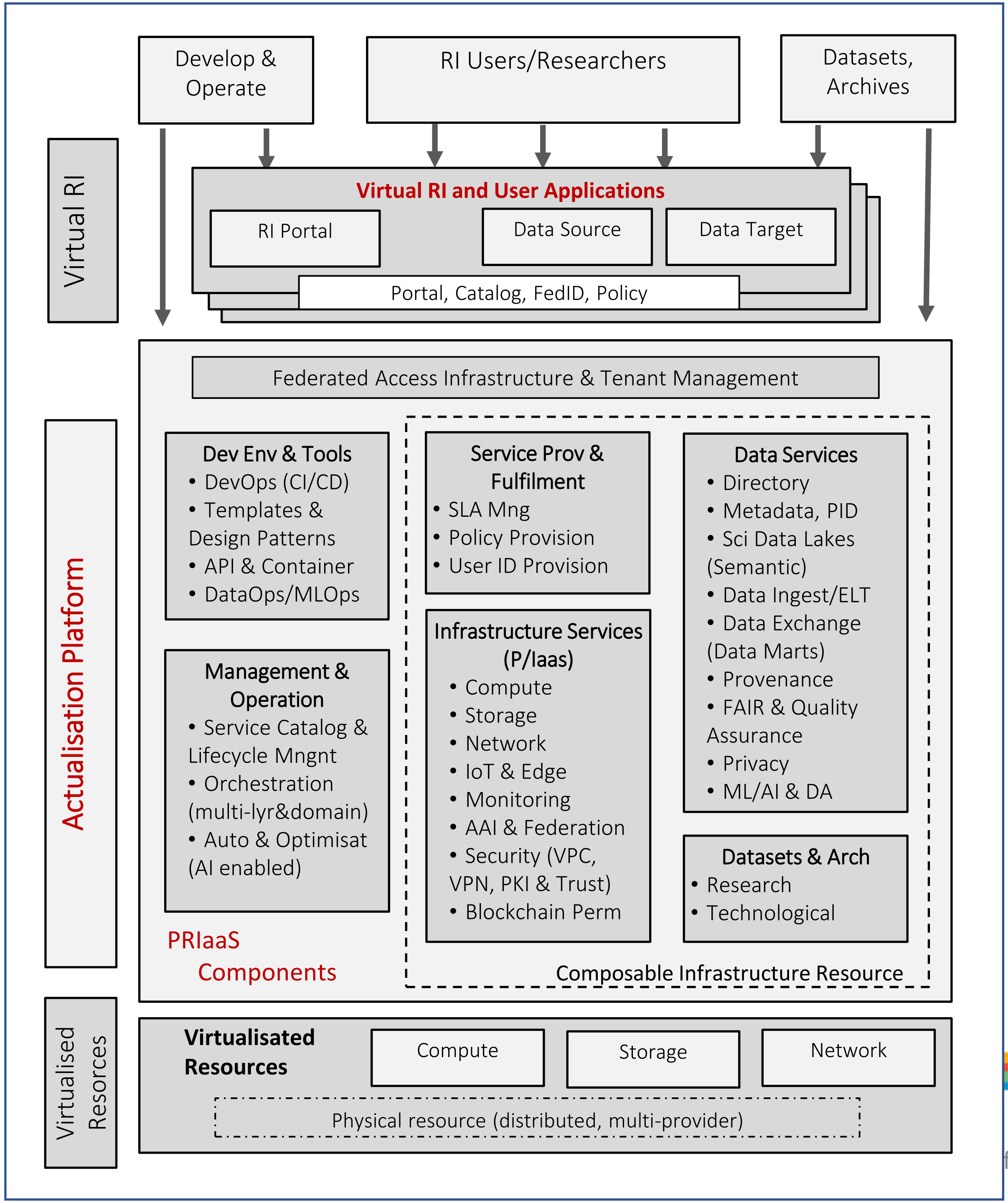
Technical Interoperability

- Common security framework and Federated Access Control & IDM
- PID registry, resolution, policy
- Multiple data & metadata formats, data search
- API Management

Semantic Interoperability

- Metadata schemas and extensibility
- Metadata catalogue
- Ontologies and semantic artefacts
- Mapping/translation semantic & metadata
- Semantic reasoning and resolution

- Provides vision and roadmap to achieving interoperability with EOSC
- Some services can be used from EOSC, some services will require API with EOSC services of metadata mapping
- Data Interoperability and sharing is an important component of SLICES-RI
  - Compliance with the Open Science and FAIR data principles
  - Semantic interoperability
  - Supported by robust data infrastructure
  - Data Management and Governance



Develop & Operate

RI Users/Researchers

Datasets, Archives

Virtual RI

Virtual RI and User Applications

RI Portal

Data Source

Data Target

Portal, Catalog, FedID, Policy

Actualisation Platform

Federated Access Infrastructure & Tenant Management

Dev Env & Tools
- DevOps (CI/CD)
- Templates & Design Patterns
- API & Container
- DataOps/MLOps

Service Prov & Fulfilment
- SLA Mng
- Policy Provision
- User ID Provision

Data Services
- Directory
- Metadata, PID
- Sci Data Lakes (Semantic)
- Data Ingest/ELT
- Data Exchange (Data Marts)
- Provenance
- FAIR & Quality Assurance
- Privacy
- ML/AI & DA

Management & Operation
- Service Catalog & Lifecycle Mngnt
- Orchestration (multi-lyr&domain)
- Auto & Optimisat (AI enabled)

Infrastructure Services (P/Iaas)
- Compute
- Storage
- Network
- IoT & Edge
- Monitoring
- AAI & Federation
- Security (VPC, VPN, PKI & Trust)
- Blockchain Perm

Datasets & Arch
- Research
- Technological

PRIaaS Components

Composable Infrastructure Resource

Virtualised Resources

Virtualisated Resources

Compute

Storage

Network

Physical resource (distributed, multi-provider)

PRIaaS Architecture Model (2021 - in progress)

Actualisation Platform Components [ref]

- Core Infrastructure Services (IaaS & PaaS)
- Data Services
- Management and Operation
- Development Environment and Tools
  - DevOps
  - Templates and Patterns
- Service Provisioning and Fulfilment
- Datasets and Archives
- Federated Access Infrastructure + IoT Edge and Tenants Management
- Virtual RIs and Portal

[ref] IG1157 Digital Platform Reference Architecture Concepts and Principles v5.0.1, 21 July 2020