Open Science and Experimental Research Lifecycle Management in SLICES-RI

Yuri Demchenko, University of Amsterdam

SLICES Roadshow Norway – 24 January 2023

Outline

- SLICES Research Infrastructure for large scale experimental research
- Open Science and Research Reproducibility
- Experimental research lifecycle and Reproducibility as a Service
  - Experimental research reproducibility study in SLICES-DS/SLICES-PP
- Data Management Infrastructure for full cycle experimental research
  - Variety and Volume of experimental data in SLICES
- Future developments on experimental research reproducibility

SLICES and Open Science

- Open Science is a major initiative by EC and ESFRI
  - Being developed in many projects in H2020 and HE2027
  - Supported by a number of European e-Infrastructure services
  - FAIR (Findable, Accessible, Interoperable, Reusable) data principles commonly accepted for managing research data
- A core objective of the European Open Science Cloud (EOSC) as a federated scientific data infrastructure
- SLICES will benefit and build on the current best practices, recommendations and tools, use services provided by Open Science platforms
- SLICES is actively involved in the EOSC activity
  - Starting from liaison with EOSC Working Groups activities in SLICES-DS to contribution to ongoing EOSC projects in SLICES-PP
  - SLICES Interoperability Framework and services integration with EOSC and Open Science

Open Science Challenges in Experimental Studies

- **SLICES is intended to support large-scale experimental studies on modern/future Digital Infrastructure technologies**
  - **Multi-site, 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 brings its specific of implementing Open Science and FAIR data principles in experimental studies on the Digital Infrastructure technologies
- Important questions in experimenting with new technologies and industry is how open research and experimental data should be
  - IPR and industrial secrets 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

Experimental Research Reproducibility: Main tasks

- Experiment description and automation, including reproducible description and experiment workflow management
- Experiment management infrastructure
- Experimental data/metadata management and FAIR data principles compliance
- Federated Data Management Infrastructure to support experimental research and SLICES infrastructure services operation

3-stages process according to ACM [ref]:

Repeatability: Same team, same experiment setup

Reproducibility: Different team, same experiment setup

Replicability (portability): Different team, different experiment setup

Experimental Research Reproducibility: Study in SLICES-DS

- Reproducible experiment description and orchestration
  - Git and CI/CD iterative experiment design and automation and deployment
  - Jupyter Notebook experiment description and orchestration
  - Common Workflow Language (CWL) for experiment management
- The plain orchestration service (pos) by Technical University Munich
  - Testbed management system and experiment workflow
- Experiment infrastructure deployment and management
  - Cloud native tools using Git CI/CD tools (leveraging DevOps tools and methodology)
  - General infrastructure automation tools Ansible, Terraform, others
- Cloud native Platform Research Infrastructure as a Service (PRIaaS) for full infrastructure, user and data services provisioning

Experiment description: Reproducibility and Portability

- GitHub and GitHub Actions (CI/CD tools)
  - Highly flexible but requires programming and full infrastructure management
  - However, can rely on well developed CI/CD tools
- Jupyter Notebook (Python based) – Popular but limited portability
  - Very popular but often limited to specific experiment environment and infrastructure platform
- Common Workflow Language (CWL)
  - Portable Experiment Description
  - Requires workflow execution environment and infrastructure provisioning platform

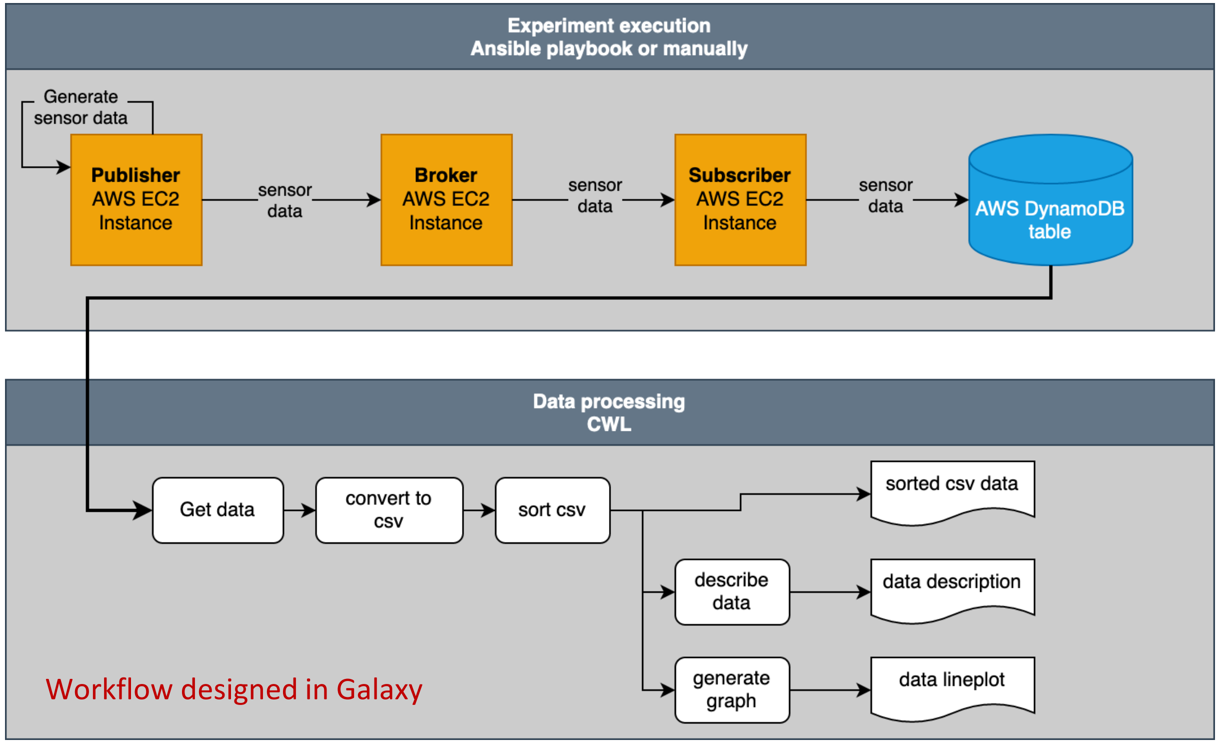
Jupyter Notebook for Experiment Automation and Workflow Description

- Build on other projects experience of using Jupiter Notebooks for experiments automation
  - Grid5000 large-scale infrastructure for experiment-driven research
    - Notebook as experiment drivers and experiment payload
    - Notebook for post-processing and exploratory programming
  - Fed4FIRE+ federation of experimental facilities for Future Internet research
    - Majority testbeds are using Notebooks
- Chameleon (CHI Cloud++) OpenStack based cloud platform to support experimental workflow for Computer Science systems research (US based)
  - Jupyter Notebook integration and experiments management via JupyterLab portal
- Plain Orchestration Services (pos) by Technical University Munich (TUM)

Common Workflow Language (CWL)

- Provides portable platform independent data handling workflow description
    - YAML based
- Requires workflow execution environment
    - Apache AirFlow, StreamFlow, Toil
- Galaxy workflow management and execution platform
    - galaxy.tools.cwl package for Galaxy open-source platform for FAIR data analysis
    - Run code in interactive environments (RStudio, Jupyter, ...) along with other tools or workflows
    - Manage data by sharing and publishing results, workflows, and visualizations
    - Ensure reproducibility by capturing the necessary information to repeat and understand data analyses
    - Recognised as cross EOSC platform supporting FAIR data lifecycle

Example: Ansible playbook and CWL workflow



Workflow designed in Galaxy

```
#!/usr/bin/env cwl-runner
cwlVersion: v1.0
class: Workflow

# The inputs of the workflow as a whole
# These are referenced in the first workflow step
inputs:
  AWS_ACCESS_KEY_ID: string
  AWS_SECRET_ACCESS_KEY: string
  table_name: string

# In the following list the workflow steps are defined
steps:
  # the first step, called "get_data" gets the sensor
data from the DynamoDB table
  get_data:
    run: ../tools/get-dynamodb-data.cwl # the CWL tool
is defined in this file
    # the following list defines the inputs to the CWL
tool
    in:
      AWS_ACCESS_KEY_ID: AWS_ACCESS_KEY_ID
      AWS_SECRET_ACCESS_KEY: AWS_SECRET_ACCESS_KEY
      table_name: table_name
    # the output of this workflow step is defined as
"dynamodb_data"
    out: [dynamodb_data]

  # the second step of the workflow converts the sensor
data from JSON to CSV
  convert_to_csv:
    run: ../tools/json-to-csv.cwl
    in:
      # the input is the output of the previous step,
"dynamodb_data"
      json_file: get_data/dynamodb_data
    out: [csv_file]

  # the third step sorts the sensor data in CSV format
  sort_csv:
    run: ../tools/sort.cwl
    in:
      file_to_sort: convert_to_csv/csv_file
      sort_field:
        default: 2 # which column to sort by
    out: [sorted_file]

  # the 4th step creates a description of the data
  describe_data:
    run: ../tools/describe-csv.cwl
    in:
      # the input is the sorted CSV file from the
previous step
      csv_file: sort_csv/sorted_file
    out: [data_description]

  # the 5th step generates a line plot
  generate_graph:
    run: ../tools/graph-csv.cwl
    in:
      # the input is also the sorted CSV file from the
3rd step
      csv_to_plot: sort_csv/sorted_file
    out: [plot]

# the outputs of the workflow as a whole are the sorted
CSV file from the third
# step, the data description from the 4th step and the
line chart from the 5th
# step
outputs:
  data_csv:
    type: File
    outputSource: sort_csv/sorted_file
  description:
    type: File
    outputSource: describe_data/data_description
  plot:
    type: File
    outputSource: generate_graph/plot
```

Experimental Research Reproducibility as a Service

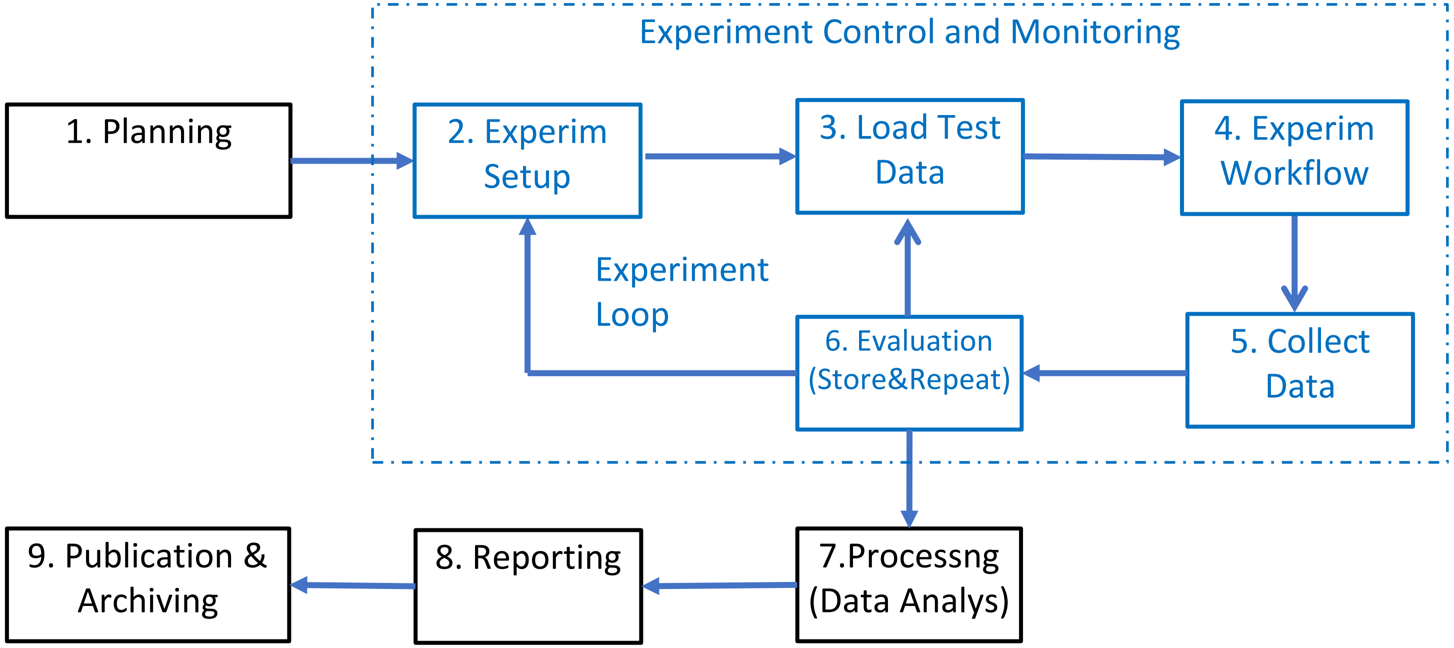
- SLICES to support experiments reproducibility to comply with Open Science
  - Focus on **repeatability** and **reproducibility** with the future support of **replicability**
- Robust, reproducible experiments
  - Documenting all relevant parameters and environment for experiments
  - Automate the documentation 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

slicesRI

Experiment Workflow and Stages


Experiment Control and Monitoring
1. Planning
2. Experim Setup
3. Load Test Data
4. Experim Workflow
5. Collect Data
6. Evaluation (Store&Repeat)
Experiment Loop
7.Processng (Data Analys)
8. Reporting
9. Publication & Archiving

SLICES Data Management Infrastructure (supporting full research lifecycle)

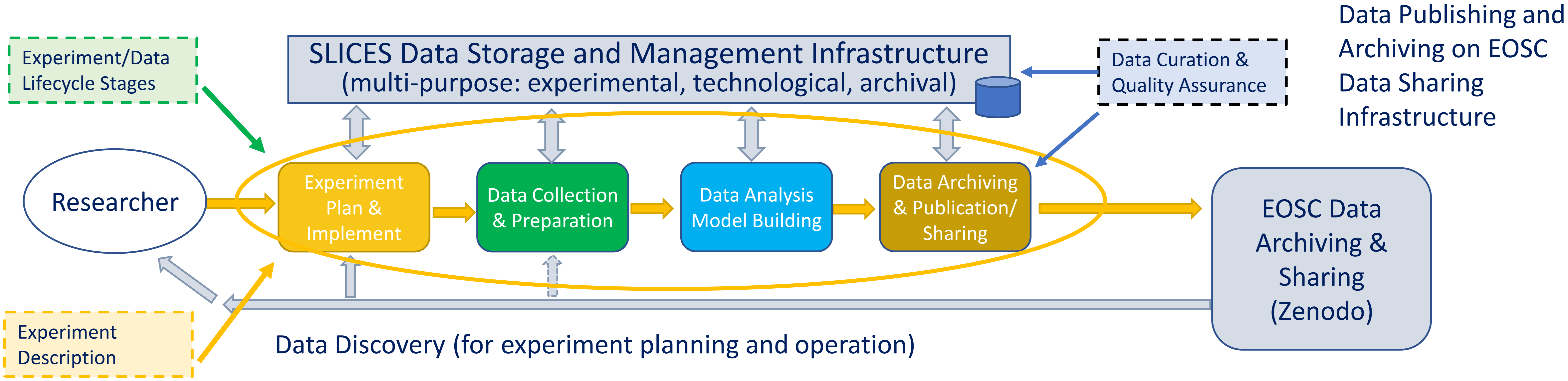
Experimental research stages

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

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

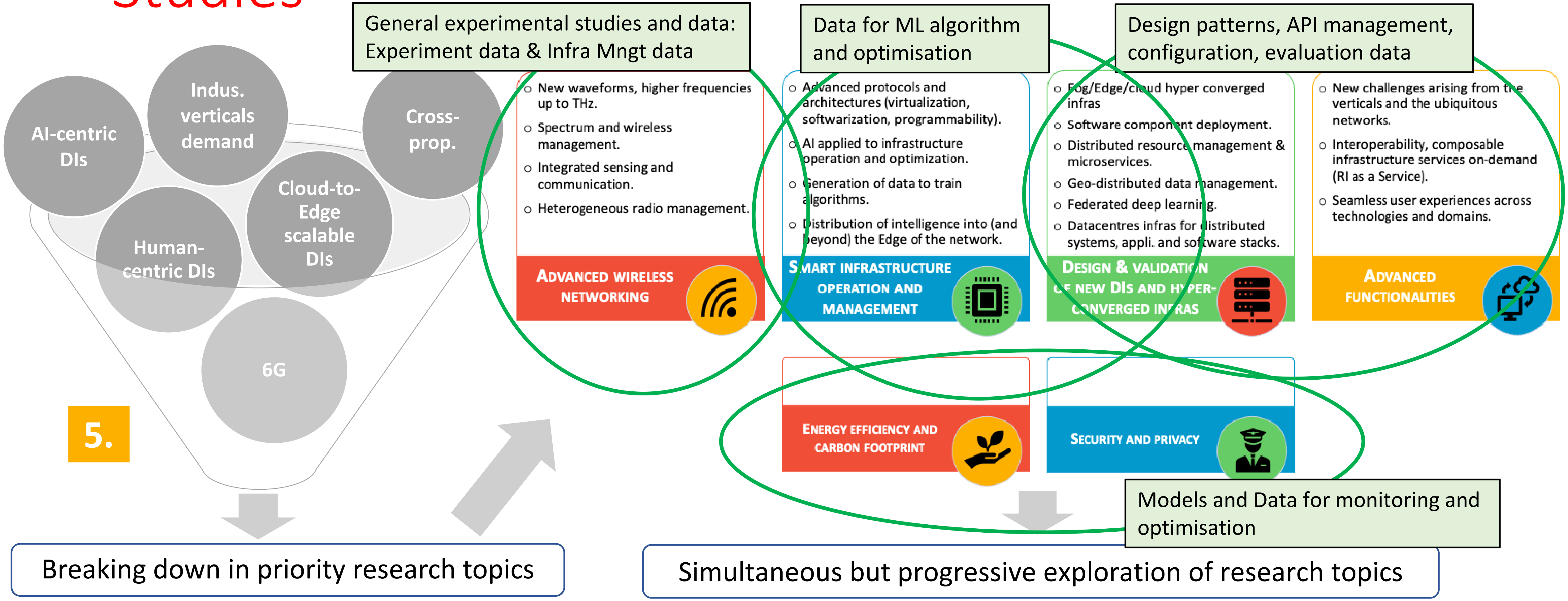
- **Experimental data are big, distributed, domain specific, serving specific communities**
  - **Require effective models and infrastructure services for Research Data Management and secure data sharing**
- Support the whole data lifecycle
  - Connected to research/experiment lifecycle or workflow
- 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


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

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

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

ADVANCED WIRELESS NETWORKING

- 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.

SMART INFRASTRUCTURE OPERATION AND MANAGEMENT

- 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.

DESIGN & VALIDATION OF NEW DIs AND HYPER-CONVERGED INFRAS

- 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.

ADVANCED FUNCTIONALITIES

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

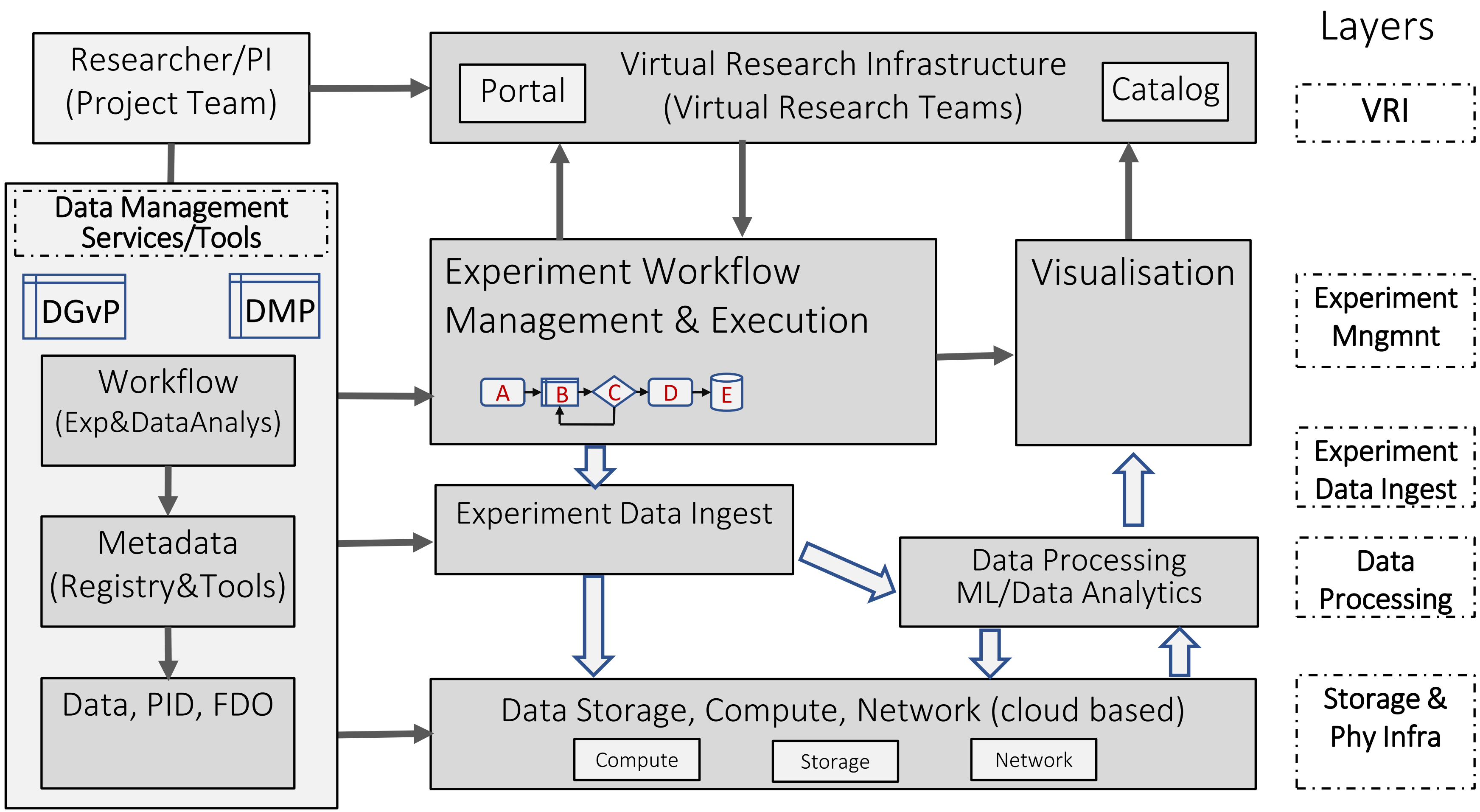
- General experimental studies and data documentation and publication
  - **FAIR (Findable, Accessible, Interoperable, Reusable)** data principles are key for experimental data sharing
  - **Metadata** profiles to be defined for major types of experiments and supported by data and metadata management 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

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

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

DGvP – Data Governance Policy DMP – Data Management Plan

DMI layers

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

Further tasks for Experimental Research Automation in SLICES-RI

- Reproducible experimental research description and infrastructure provisioning tools
  - Platform RI as a Service (PRIaaS) for distributed experimental infrastructure provisioning for virtual researcher teams
  - Adopting Research Object concept (by EOSC and Reliance project)
- Federated multilayer experimental data management infrastructure
  - Experiment data collection, processing and storage
  - Data management policy definition and FAIR compliance
- Metadata as cornerstone for reproducibility of experimental research
  - Metadata profiles definition, extension to support infrastructure management information CIM, MIB, GLUE schemas
- EOSC compliance, interoperability and integration
  - Basis for the future cooperation with European RIs and contribution to EOSC development

Questions and invitation to cooperation

www.slices-ri.eu

slicesRI

www.slices-ri.eu