Experimental Research Reproducibility and Experiment Workflow Management

Yuri Demchenko, University of Amsterdam

on behalf of

Sebastian Gallenmüller (Technical University of Munich, Germany), Serge Fdida (Sorbonne University, France), Panayiotis Andreou (University of Central Lancashire), Mathias Kirkeng (UvA)

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
  - pos – plain orchestration service
- Data Management Infrastructure for full cycle experimental research
  - Variety and Volume of experimental data in SLICES
- Future developments on experimental research reproducibility

SLICES-RI: Scientific Large-scale Infrastructure for Computing/ Communication Experimental Studies

- SLICES is a distributed Digital Infrastructure to support large-scale experimental research focused on networking protocols, radio technologies, services, data collection, parallel and distributed computing
- SLICES will integrate multiple experimental facilities and testbeds operated by partners providing a common services access and integration platform
- SLICES will allow academics and industry to experiment and test the whole spectrum of digital technologies
  - Experiment automation to support design, experiment, and operate the full research lifecycle management.

slicesRI

SLICES and Open Science: Challenges

- **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**
- SLICES-RI brings its specific of implementing Open Science and FAIR data principles in experimental studies on the Digital Infrastructure technologies
- Scientific value of experimental research is in the reproducibility of experiments, sharing and (re)usability of data
- 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

slicesRI

Experimental Research Reproducibility: Main aspects

3-stages process according to ACM [ref]:

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 description and automation, including reproducible description and experiment workflow management
- Experimental infrastructure services operation
- Experimental data and metadata management
- Federated Data Management Infrastructure to support complex experimental research
- FAIR data principles compliance (core for Open Science)

[ref] https://www.acm.org/publications/policies/artifact-review-and-badging-curr

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 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

Experimental Research Reproducibility: Study in SLICES-DS/SLICES-PP

- 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
- 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
- The plain orchestration service (pos) by Technical University Munich
  - Testbed management system and experiment workflow
- 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

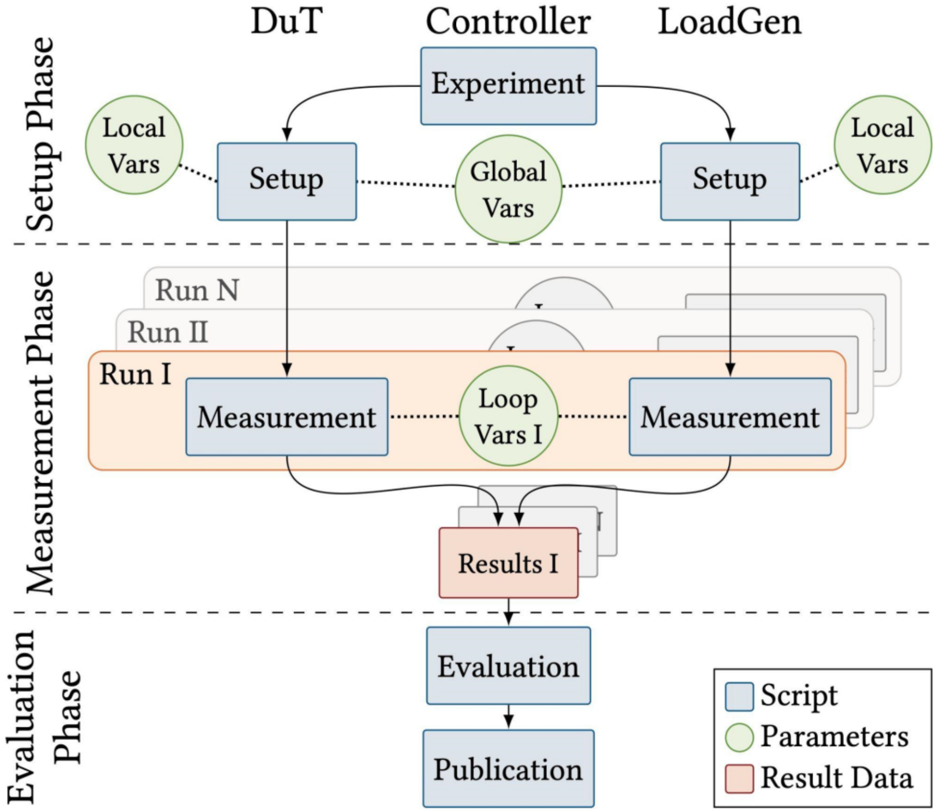
- Build on other projects experience of using Jupiter Notebooks for experiments automation
  - Grid5000 large-scale infrastructure for experiment-driven research
    - Notebook as experiment drivers and experiment payload
    - Notebook for post-processing and exploratory programming
    - Notebooks as tutorials
  - 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)
  - JupyterNotebook integration and experiments management via JupyterLab portal
- Plain Orchestration Services (pos) by Technical University Munich (TUM)

Plain Orchestration Service (pos)

- 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.

pos Experiment Workflow Management

- 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)

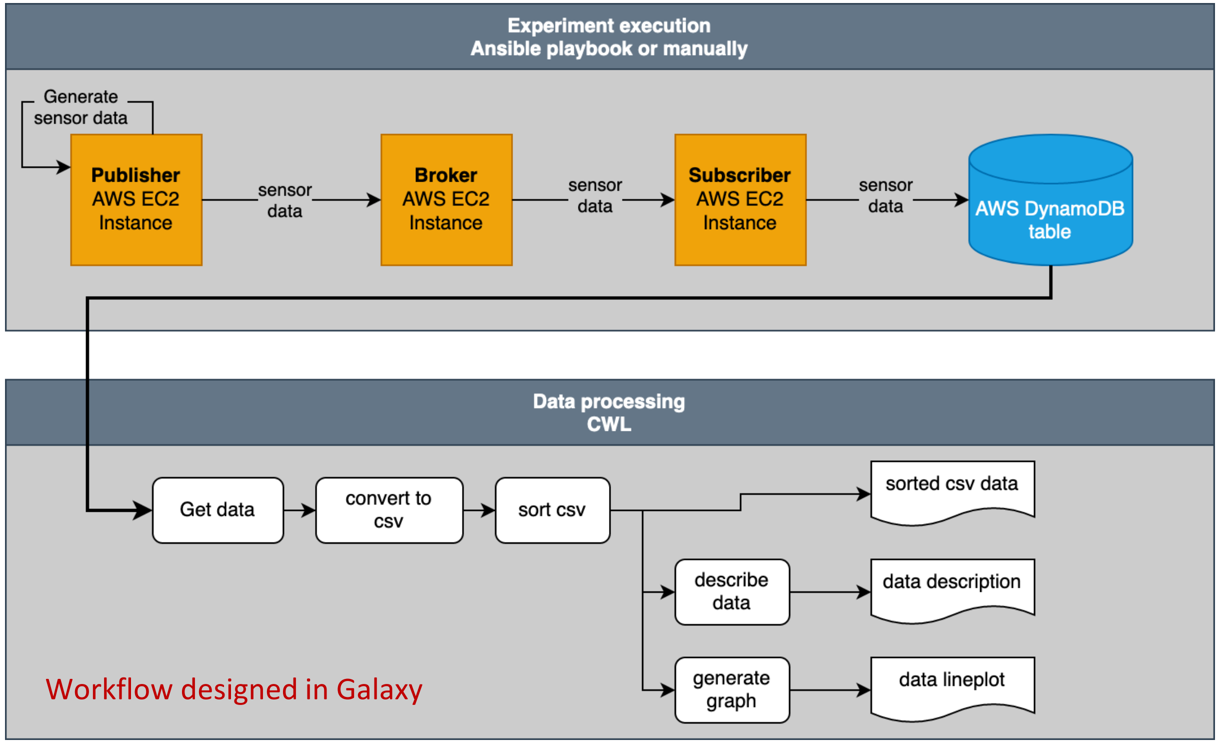


Structured Experiment Workflow with pos

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



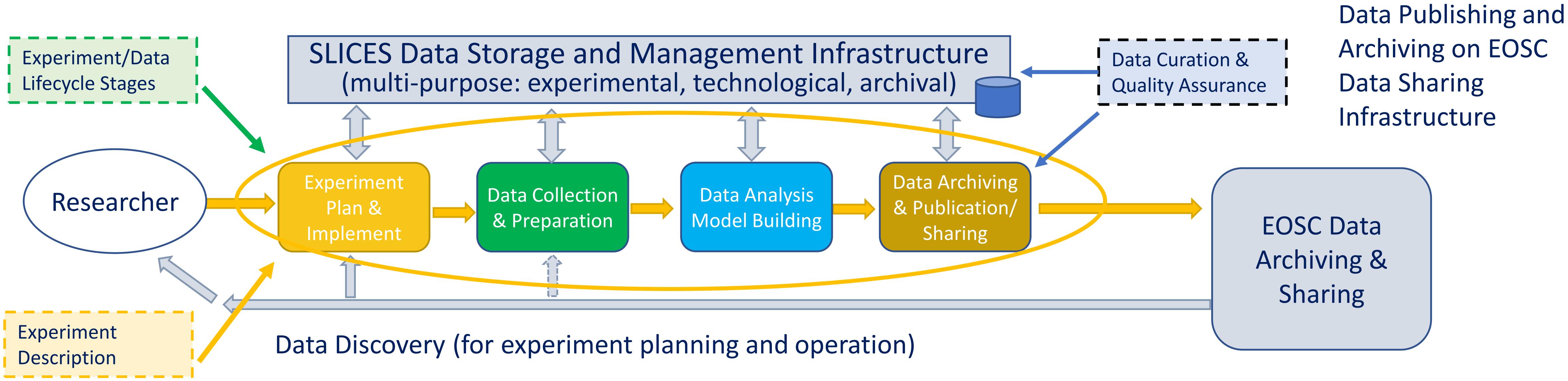
Workflow designed in Galaxy

```
#!/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
```

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

- **Experimental data are big, distributed, domain specific, serving specific community**
  - **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
- (Trusted) Data exchange and secure transfer protocols

SLICES Experimental Data Lifecycle Model and Dataflow

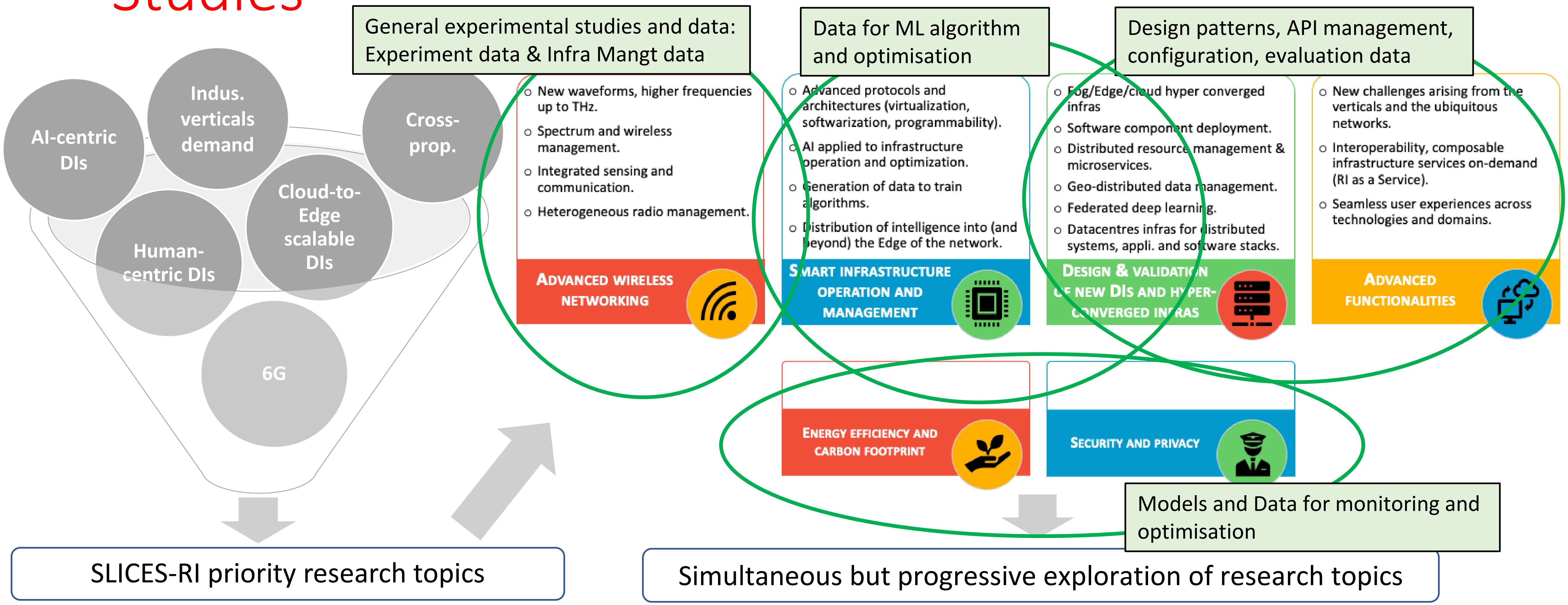

Experiment/Data Lifecycle Stages
SLICES Data Storage and Management Infrastructure (multi-purpose: experimental, technological, archival)
Data Curation & Quality Assurance
Data Publishing and Archiving on EOSC Data Sharing Infrastructure
Researcher
Experiment Plan & Implement
Data Collection & Preparation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

Variety of Data produced in SLICES

- General Digital Infrastructure experimental studies and data documentation and publication
  - 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
  - **FAIR (Findable, Accessible, Interoperable, Reusable)** data principles are key for experimental data sharing
- 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
  - Metadata for API description, identification, composability
  - Research Object (RO) and FAIR Digital Object (being developed by EOSC)

Different Types of Data for Different Experimental Studies



General experimental studies and data: Experiment data & Infra Mangt 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.

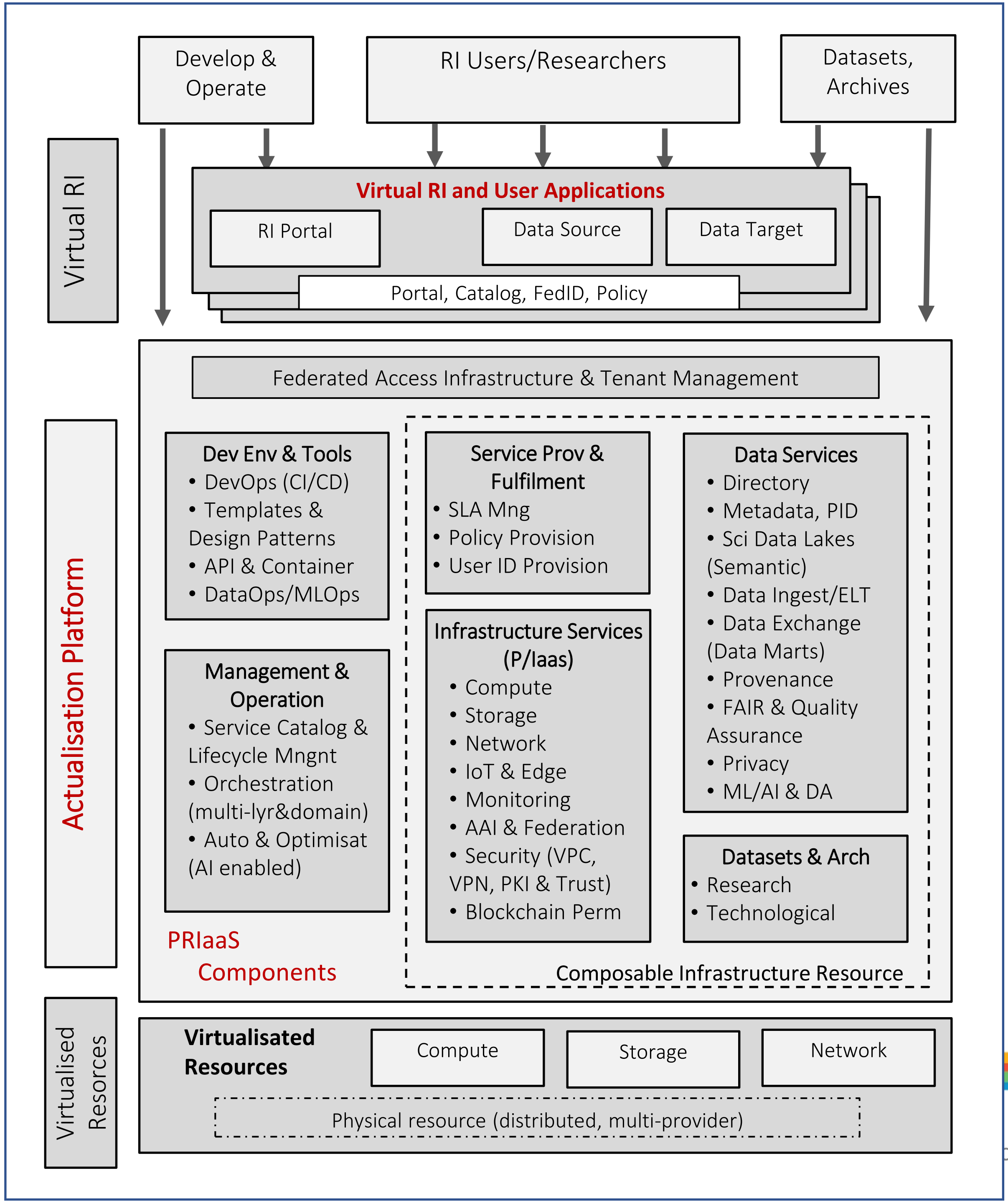
ENERGY EFFICIENCY AND CARBON FOOTPRINT

SECURITY AND PRIVACY

Models and Data for monitoring and optimisation

SLICES-RI priority research topics

Simultaneous but progressive exploration of research topics



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

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

PRIaaS Components

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

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

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

Datasets & Arch
- Research
- Technological

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

Further tasks for Experimental Research Automation in SLICES-RI

- Reproducible experimental research description and infrastructure tools
- PRIaaS for distributed experimental infrastructure provisioning
  - Experiment and data management
  - Virtual researcher teams support and federated identity management (user provisioning and access management)
- Metadata as a cornerstone for reproducibility of experimental research
  - Metadata profiles definition
- EOSC compliance, interoperability and integration
  - Basic for future cooperation

Questions and invitation to cooperation

www.slices-ri.eu

slicesRI

www.slices-ri.eu