SLICES-RI Data Management Infrastructure for Experimental Research

Based on discussion after SLICES Summer School discussion on 15 June 2023 and AH meeting on 5 July 2023

Suggested actions and activities

The following are suggested actions and activities by activities by involved partners.

1. General principles and implementation approach

Data Management is an essential component of the SLICES-RI infrastructure that includes data collection from experiments (including experiment description and measurement data), data storage, data preparation, data lineage and quality assurance, data publication, and data sharing.

A consistent definition of DMI will impose specific requirements to the SLICES Blueprint/Reference Architecture and will require the implementation of special services in the SLICES infrastructure to support data collection, data management, and data sharing.

It is understood thatg DMI creation will have staged process starting from with Blueprint integration, delivering Minimum Viable Product (MVP) and following the SLICES-RI evolution that should also incorporate data and metadata development technologies development, primarily coordinated and facilitated by EOSC. Longterm vision for DMI should incorporate all this factors by adopting sustainable architecture design principles. At some stage it will also require specification of the formal requirements to SLICES DMI, wich will be possibly adopted from the EOSC cluster projects such as FIRCORE4EOSC.

2. SLICES Data Management Infrastructure (DMI)

SLICES must create and maintain its own Data Management Infrastructure that will include central data storage and connected data storage nodes operated by partners and big experimental facilities.

DMI must include all necessary services to support the whole experimental data lifecycle and also include reference datasets such as required for experiment execution, or ML/AI algorithms training.

SLICES DMI may use, where possible, external data storage, registries, metadata services or data discovery services, which should be federated with the SLICES DMI.

DMI must implement federated access control principles and allow integration and federation with EOSC and EGI data management infrastructure and services, in particular for data sharing, publication, and access.

Task 1. Define who will host and operate the SLICES central data storage and an initial set of storage nodes. Decisions should be made on how to federate distributed storage nodes and provide transparent access to the whole DMI. – All, Exec Board

Task 2. Assess options for establishing Metadata Registry to serve SLICES data management purposes, what services should be deployed in the SLICES DMI and which services can be used from external providers/operators in EOSC or EGI infrastructure. Who can host and operate necessary metadata services? - UCLAN

3. General Metadata definition and management

Metadata are an important component of DMI that provide a basis for services interoperability, experimental research reproducibility, effective data sharing and discovery. Effective and consistent metadata management is the foundation of the FAIR data principles implementation.

All data are defined by the data models, metadata, data formats and data types. Metadata are defined as part of the data model.

For SLICES as RI for experimental studies in digital technologies and ICT, metadata includes three main areas:

- General services description: metadata profiles and metadata will be used for publishing SLICES services in EOSC Catalog and the SLICES services catalog
- Description of data collected, produced and handled in SLICES-RI that include experimental data, staged/processed data, archival data, publications, reports, activities, and management data. Additional data categorization is required.
- Experiment description that must include all necessary information required for experiment reproducibility and deployment

Two other categories of metadata that may be required to support SLICES include:

- Infrastructure descriptions that are required for infrastructure management and monitoring (network devices, network traffic, status and events). This type of metadata are well supported by existing network management and service management standards (SNMP MIB-II and related, DMTF CIM and CIMI, TMForum SID)
- Metadata for data processing and lineage, in particular, for data used in ML and AI processes

Defining metadata often includes defining metadata namespaces that will create and basis for unique metadata elements identification and consequently discovery, sharing and integration.

Task 3. Assess existing metadata format and provide recommendations for metadata formats for SLICES services and resources, data produced in SLICES experiments. - UCLAN

Task 4. Assess existing metadata format for infrastructure description and management and their relevance for SLICES use cases. - UvA

Task 5. Assess existing metadata format and provide recommendations for metadata formats for experiment description (see additional information below). – UvA, UCLAN, TUM

4. Experiment description and metadata

Consistent/full experiment description and corresponding metadata must ensure experiment/experimental research reproducibility and FAIR data sharing.

The following data types and metadata are considered essential for consistent experiment description:

- Experiment abstract model with parameters, input variables and variables under test (as it is known at the beginning)
- Experiment setup/infrastructure, including network equipment and the network topology, including VMs/containers
    - Hardware: list of major hardware components (e.g., CPU, network cards, memory, storage)
    - Firmware: version numbers of the installed firmware, e.g., BIOS version, CPU microcode version, firmware version of the network card
    - Software: version of the investigated software and its installed dependencies, version of the installed OS (including version of OS kernel), software installed on the OS
- Configuration of all infrastructure components, deployment sequence (presumably Ansible playbook, Terraform plan, or Jupyter Notebook)
- Test generators, measurement equipment and sensors (and corresponding infrastructure points)
    - Specification of the generated traffic (the content of the traffic)
    - Specification of the patterns used for the generated traffic (e.g., a distribution of the inter-packet gaps or traffic bursts)
- Environment description (hardware/software)
- Experiment workflow (the usage of pos ensures reproducibility of experiment workflow)
- Data ingest process, data preprocessing and assessment
- APIs for experiment setup, monitoring, and data collection

Data models and metadata must be defined for all types of data describing the experiment.

For some well-established experiments, data models maybe defined for the specific data storage and database type, such as data lakes, SQL database, kye-value, document based, or triple storage (for semantic data).

Experiment as a Research Object must be assigned a unique identifier and experiment/object type, optionally registered schema and domain namespace.

Goal of reproducibility: The goal of reproducible experiments is the reproduction of key performance descriptors for a specific experiment. The exact reproduction of these key performance descriptors may depend on specific hardware and software that may change over time. We rather aim for recording a complete picture of the environment an experiment is conducted in. The recorded information is the foundation to find differences between experiments that may not be obvious during the initial experiment.

The usage of pos ensures the collection of many aspects mentioned before. To generate the necessary information for the experiment description, pos relies on standard Linux tools. Typical tools to automate the hardware description would be `lshw` that lists all hardware (and some of the firmware versions). Additional, more specialized tools may be used for other hardware components, such as `ethtool` for network cards. The testbed itself may provide information about the topology, that should be detected regularly (e.g., using tools for topology detection such as `lldp`).

We propose to record the typical format provided by the mentioned tools. A testbed should provide an abstraction layer to access the mentioned information in a more generalized format to simplify further processing. Testbed tools such as pos already provide a common data format (JSON) to unify the output of the different tools.

Task 6. Make an inventory of all data required for full experiment reproducibility (with the target for portability), including infrastructure, environment, variables, used data formats or data models. Such inventory should document a current practice, which further will be used for more formal definitions of data models and metadata. – TUM *(see comments above)*

Task 7. Investigate what and how essential metadata can be extracted from the Ansible playbook or Jupyter Notebook describing experiment setup and orchestration. - UVA

5. SLICES Blueprint Architecture

SLICES Blueprint defines a basic/core/instant infrastructure setup that can be used for running experiments on 5G/6G and related networking technologies. Blueprint is defined in a modular way that allows defining basic building blocks or design patterns that can be composed in a different way to create a platform for running different experiments.

To achieve effective composability and flexible customization of the SLICES experimental setups, the following data and metadata should be defined (similar for the general experiment setup as described above):

- Services deployed in the Blueprint and corresponding APIs
- Input and output data or signals
- Configuration of all elements, including RAN (RU – Radio Units, UE – User Equipment), core 5G network, dedicated network (VPN or VPC), network switches, servers
- Computational nodes/instances type and configuration: hardware (AMD/Intel, RAM, OS, firmware)
- Operational environment that may need to be documented for experiments
- Infrastructure design patterns or templates (in the form of Ansible playbooks or Terraform plans)
- Monitoring or measurement point and corresponding API
- Experiment-specific or other data collected in the infrastructure

Task 8. Make an inventory of all data used or required for the Blueprint infrastructure description, deployment and operation as a part of the experiment setup, as described above. Such inventory should document a current practice, which further will be used for more formal definitions of data models and metadata. - INRIA

6. Metadata Management tools

SLICES must provide metadata management tools that would help researchers and data managers/data stewards to create, combine, transform/map and publish metadata in a consistent way and with minimum efforts and maximum automation.

Metadata tools should support the creation of metadata when publishing research results, papers, datasets, reports.

The proposed metadata registry solution must be based on the already existing Open Source solutions which are preferably adopted (or developed) by the EOSC community.

Task 9. Assess existing metadata registries and metadata management tools for different categories of metadata and use cases/scenarios and provide recommendations to project partners. UCLAN