

HORIZON 2020 H2020 - INFRAIA-2020-1

D3.2 Data Management and Reproducibility Methods Report

Acronym	SLICES-SC
Project Title	Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies – Starting Community
Grand Agreement	101008468
Project Duration	42 Months (01/03/2021 – 31/08/2024)
Due Date	30 June 2024 (M40)
Submission Date	17 July 2024 (M41)
Authors	Cédric Crettaz (MI), Vasiliki Tsiompanidou (MI) Adrian Quesada Rodriguez (MI) Sébastien Ziegler (MI), Renata Radocz (MI), Ana María Paccheco (MI), Lisa Sieker (MI), Maria Roglekova (MI), Iida Lehto (MI), Dolina Tsiotzora (MI), Brecht Vermeulen (imec), Wim Van der Meerssche (imec), Thijs Walcarius (imec), Lucas Nussbaum (Inria), Luke Bertot (inria), Houssam ElBouanani (inria), Chadi Barakat (inria), Walid Dabbous (inria), Thierry Turletti (inria), Antti Pauanne (Oulu), Akos Hajnal (Sztaki)
Reviewers	Serge Fdida (SU), Raffaele Bruno (CNR)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101008468. The information, documentation and figures available in this deliverable, is written by the SLICES-SC project consortium and does not necessarily reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information contained herein.





Executive Summary

This deliverable is the second version of the data management plan for SLICES-SC, thus, following deliverable 3.1, however, it also contains reporting activities in regard to the reproducibility methodology of SLICES. The aim of a data management plan is to detail the handling of personal data that is involved in a research project, such as SLICES, and, as such, provide integrity and accountability for partners and stakeholders throughout the project. As the second version, the aim of D3.2 is to build upon the work that has been done within the first version of the data management plan. Therefore, this version will solely contain updates I regard to regulatory compliance and the management of data by the consortium partners.

Furthermore, this document will have another focal point, next to providing updates for the data management, which is reproducibility. As such, the importance of reproducibility, especially in regard to the avoidance of errors in repetition of the experiment, is detailed, as well as best practices are suggested in order to attain to common standards of reproducibility.

In regard to the updates of the data management plan, this document focusses less on the ethical components, as those are provided for, in detail, in the ELES report. However, an overview of the partner's handling of personal data is given as well. Thus, an extensive data summary is given detailing aspects such as type, size and format of data, as well as, the purpose of processing or reusing the data. An overview of the collection and processing activities of personal data by the partners is given as well, and safeguards are explained that are being taken in order to comply with relevant legislation on data protection.

Additionally, data monitoring is discussed and new ways of enhancing coordination in regard to data protection are introduced as well. In regard to the latter, the tasks and objectives of the Data Protection Office are described, including the website, which enables partners to easily update information if something has changed in regard to their data management, and it enables data subjects to be able to contact a central point to complain about the exercise of their rights in regard to data protection.

Lastly, this deliverable produces recommendations addressing the general data management activities conducted throughout the project, but also specific recommendations on how reproducibility requirements can be adhered to accordingly.



Table of Contents

EXECUTIVE SUMMARY	2
TABLE OF CONTENTS	3
LIST OF FIGURES.....	6
LIST OF TABLES	7
LIST OF ABBREVIATIONS.....	8
1. SLICES-SC IN A NUTSHELL	9
1.1. OBJECTIVES OF WP3 AND ITS RELEVANT TASKS	9
1.2. OBJECTIVES OF D3.2.....	9
2. OVERALL STRUCTURE OF THE DELIVERABLE	10
3. IMPORTANCE OF REPRODUCIBILITY FOR SLICES	10
3.1. EUROPEAN AND ESFRI POLICY.....	11
4. BEST PRACTICES AND TOOLS FOR REPRODUCIBILITY.....	11
4.1. BEST PRACTICES FOR REPRODUCIBILITY.....	11
4.2. TOOLS FOR REPRODUCIBILITY.....	12
4.2.1. <i>Reproducibility in the Fed4FIRE project</i>	12
4.2.2. <i>Jupyter notebooks for reproducible experiments</i>	16
4.2.3. <i>Distrinet (elay-based fidelity monitoring of network emulation)</i>	20
5. REPRODUCIBILITY TEST AND VALIDATION	21
5.1. REPRODUCIBILITY TESTS AND VALIDATION APPROACH	21
5.2. TECHNICAL VALIDATION	22
5.2.1. <i>User experience</i>	22
5.3. TECHNICAL ASPECTS.....	22
5.3.1. <i>Command-line interface</i>	22
5.3.2. <i>Links to data</i>	23
5.3.3. <i>ACL and user management</i>	24
5.3.4. <i>Data size limits</i>	24
5.3.5. <i>Data size display</i>	24
5.4. FAIR ASPECTS	26
5.4.1. <i>Findability and Accessibility</i>	26
5.4.2. <i>Interoperability, Reusability, and Metadata</i>	26
6. REGULATORY AND ETHICAL COMPLIANCE UPDATES	27
6.1. REGULATORY COMPLIANCE UPDATES	27
6.2. ETHICAL COMPLIANCE UPDATES.....	28





7.	DATA MANAGEMENT PLAN	28
7.1.	DATA SUMMARY	28
7.1.1.	<i>Types of Data</i>	29
7.1.2.	<i>Formats of Data</i>	29
7.1.3.	<i>Purpose of data generation or reuse</i>	29
7.1.4.	<i>Data Size</i>	30
7.1.5.	<i>Origin of generated or reused data</i>	30
7.1.6.	<i>User Groups</i>	30
7.1.7.	<i>Dataset License Types</i>	30
7.1.8.	<i>Interaction with Other Infrastructures/Systems</i>	30
7.2.	CATEGORIES OF DATA	31
7.3.	OVERVIEW OF PERSONAL DATA COLLECTION, PROCESSING, AND SAFEGUARDS	42
7.4.	DATA SECURITY	48
8.	FAIR DATA ACCESS POLICY	49
8.1.	FAIR PRINCIPLES OVERVIEW	49
8.2.	FINDABILITY	50
8.3.	ACCESSIBILITY	51
8.4.	INTEROPERABILITY	52
8.5.	REUSABILITY	52
8.6.	OPEN DATA REPOSITORIES	53
8.6.1.	<i>Open CKAN Server</i>	53
8.6.2.	<i>GitLab</i>	58
8.6.3.	<i>Dataverse</i>	59
8.6.4.	<i>EOSC Integration</i>	61
8.7.	ANALYSIS FOR THE OPEN DATA STORAGE	61
9.	INTELLECTUAL PROPERTY RIGHTS MANAGEMENT	63
10.	ALLOCATION OF RESOURCES	63
11.	DATA PROTECTION OFFICE AND ORGANISATION	63
11.1.	COORDINATION WITH SLICES-DS AND SLICES-PP	63
11.2.	DATA PROTECTION COORDINATION COMMITTEE	64
11.3.	DATA PROTECTION OFFICE	65
11.4.	PUBLIC INFORMATION FOR DATA MANAGEMENT AND PROTECTION	70
11.4.1.	<i>SLICES-SC website</i>	70
11.4.2.	<i>SLICES Portal</i>	70
12.	DATA PROTECTION MONITORING	72



12.1.	DATA PROTECTION COORDINATION	72
12.2.	DATA PROTECTION WORKSHOP	72
12.3.	DATA PROTECTION MONITORING	73
12.4.	DATA PROTECTION AND COMPLIANCE TOOLS FOR SLICES	73
13.	RECOMMENDATIONS.....	74
14.	CONCLUSION	76
	ANNEX A: DATA MANAGEMENT PROCESSING FORM	77
	ANNEX B: DATA PROCESSING AGREEMENT	83
	ANNEX C: DATA PROTECTION COORDINATION AND MONITORING SURVEY.....	84





List of Figures

FIGURE 1. ESPEC EXAMPLE WITH THE NEW FUNCTIONALITY OF 'DIRECT' FOR LITERAL BLOCKS OF DATA.....	13
FIGURE 2. ESPEC EXAMPLE WITH ANSIBLE SCRIPTING, SUPPORTING THE NEW ITEM 'GALAXY' FOR ANSIBLE GALAXY (PRE-PACKAGED UNITS OF WORK)	14
FIGURE 3. EXAMPLE OF AN EXPO ORCHESTRATION DEFINITION (INCLUDING THE NEW 'ENVIRONMENT')	15
FIGURE 4. DOCUMENTATION OF THE EXPO ORCHESTRATION DEFINITION FILE	15
FIGURE 5. JUPYTER NOTEBOOK AT IMEC'S GPULAB TESTBED.....	16
FIGURE 6. GRID'5000 METADATA BUNDLER USAGE AND EXAMPLE	18
FIGURE 7. SLICES-SC CKAN SERVER WELCOME-PAGE	22
FIGURE 8. URL TO DATA AVAILABLE.....	23
FIGURE 9. INTERFACE FOR UPLOADING DATA OR CREATING A LINK	23
FIGURE 10. METADATA FIELDS INCLUDING SIZE INFORMATION.....	25
FIGURE 11. METADATA FIELDS IN CKAN.....	27
FIGURE 12. INTERACTIONS WITH OTHER WPS	54
FIGURE 13. DATASET FOR 5G EXPERIMENTS	55
FIGURE 14. EXAMPLE OF A FILE STORED ON GITHUB.....	56
FIGURE 15. CKAN AVAILABLE THROUGH EOSC.....	57
FIGURE 16. EXPERIMENT DONE ON SLICES-SC TESTBEDS AND PUBLISHED ON GITLAB.....	58
FIGURE 17. DATA PROTECTION AND COORDINATION COMMITTEE	65
FIGURE 18. HOME PAGE OF THE DATA PROTECTION OFFICE WEBSITE	66
FIGURE 19. ONLINE FORM TO UPDATE SLICES PARTNERS' DATA PROTECTION ACTIVITIES	67
FIGURE 20. ONLINE FORM FOR DATA SUBJECTS	68
FIGURE 21. DATA PROTECTION OFFICE RULES AND OPERATION	69
FIGURE 22. SLICES DATA PROTECTION OFFICE REPORTS.....	70
FIGURE 23. SLICES PORTAL HOME-PAGE	71
FIGURE 24. A LOOK AT THE 'DISCOVER OUR TESTBEDS' SITE ON THE SLICES PORTAL.....	71



List of Tables

TABLE 1. CATEGORIES OF DATA AND DATASETS	32
TABLE 2. DATASET EXPLANATION TABLE	32
TABLE 3. MI DATASET(S)	33
TABLE 4. MI DATASET(S)	34
TABLE 5. UTH DATASET(S).....	35
TABLE 6. UTH DATASET(S).....	36
TABLE 7. UTH DATASET(S).....	37
TABLE 8. UTH DATASET(S).....	37
TABLE 9. UTH DATASET(S).....	38
TABLE 10. IMDEA DATASET(S).....	39
TABLE 11. SZTAKI DATASET(S).....	40
TABLE 12. PSNC DATASET(S)	41
TABLE 13. IMEC DATASET(S).....	41
TABLE 14. MI'S RESPONSE TO PERSONAL DATA COLLECTION, PROCESSING, AND SAFEGUARDS.....	43
TABLE 15. UTH'S RESPONSE TO PERSONAL DATA COLLECTION, PROCESSING, AND SAFEGUARDS.....	44
TABLE 16. IMDEA'S RESPONSE TO PERSONAL DATA COLLECTION, PROCESSING, AND SAFEGUARDS.....	45
TABLE 17. SZTAKI'S RESPONSE TO PERSONAL DATA COLLECTION, PROCESSING, AND SAFEGUARDS.....	46
TABLE 18. INRIA'S RESPONSE TO PERSONAL DATA COLLECTION, PROCESSING, AND SAFEGUARDS.....	47
TABLE 19. IMEC'S RESPONSE TO PERSONAL DATA COLLECTION, PROCESSING, AND SAFEGUARDS.....	48
TABLE 20. TECHNICAL AND ORGANISATIONAL MEASURES.....	49
TABLE 21. DATA FINDABILITY ACCORDING TO PARTNER RESPONSES.....	50
TABLE 22. DATA ACCESSIBILITY ACCORDING TO PARTNER RESPONSES	51
TABLE 23. DATA INTEROPERABILITY ACCORDING TO PARTNER RESPONSES.....	52
TABLE 24. REUSABLE DATA ACCORDING TO PARTNER RESPONSES	53



List of Abbreviations

- AI** (Artificial Intelligence)
- API** (Application Programming Interface)
- BS** (Base Station)
- CQI** (Channel Quality Indicator)
- DMI** (Data Management Infrastructure)
- DMP** (Data Management Plan)
- DOI** (Digital Object Identifier)
- DPA** (Data Protection Authority)
- DPIA** (Data Protection Impact Assessment)
- DPO** (Data Protection Officer)
- EOSC** (European Open Science Cloud)
- ESFRI** (European Strategy Forum on Research Infrastructures)
- ESpec** (Experiment Specification)
- FAIR** (Findable, Accessible, Interoperable and Reusable)
- GDPR** (General Data Protection Regulation)
- IQSS** (Institute for Quantitative Social Science)
- IP** (Intellectual Property)
- IPR** (Intellectual Property Rights)
- MRS** (Metadata Registry System)
- NAS** (Network Attached Storage)
- OAI-PMH** (Open Archives Initiative Protocol for Metadata Harvesting)
- RI** (Research Infrastructure)
- RO** (Research object)
- SFDO** (SLICES Fair Digital Object)
- TOMs** (Technical and Organizational Measures)
- UI** User Interface
- UTH** (University of Thessaly)
- WP** (Work Package)
- WSGI** (Web Server Gateway Interface)



1. SLICES-SC in a nutshell

The SLICES Research Infrastructure aims to develop and provide services related to experimentation in the context of digital sciences such as 5G, 6G, NFV, Internet of Things, and cloud computing. The SLICES-SC project is currently building a community of researchers around SLICES-RI, which will offer the necessary solutions to create and manage efficiently the experiments. Among the features to be implemented by the SLICES-RI for the experimenters, the SLICES-SC project has investigated and developed a facilitated access procedure for the experiments, ensuring the reproducibility of the research experiments, the validation of the experiment results, and, finally, the publication of the results in an open data access format. This process must be correlated with appropriate protection to data subject rights in line with European requirements and regulations. The following sections introduce the baseline elements to consider in this task.

1.1. Objectives of WP3 and its relevant tasks

The goals of WP3 are to ease the virtual access to the data collected during the different experiments and to create the required policy and guidelines on data management. They ensure a correct collection of the data from the different testbeds, efficient data management and a data protection framework that is compliant with the regulations and the standards related to security and privacy. Similarly, a framework to facilitate the reproducibility of the experiments was designed in the context of WP3. An open data server has been installed for the publication of open data generated during the project.

The first task of WP3 is task T3.1 which is in charge of the Data Management Plan (DMP). This implies taking into account European regulation on data, in particular considering the GDPR (General Data Protection Regulation) and the ePrivacy Directive. The FAIR (Findable, Accessible, Interoperable, Reusable) principles have been integrated into the project's activities, as described in the Data Management Plan to guarantee the sustainability of data generated in the diverse experiments. T3.1 was completed during the first semester of the project and delivered the first iteration of the data management plan, providing the initial framework and guidelines to be followed by the consortium.

Task T3.2 concerns the reproducibility of the results, notably through the provisioning of the data created during previous experiments. Guidelines on how to publish the experiment results have been defined to increase the scientific interactions in the different disciplines involved in the SLICES activities. Benchmarking has been also investigated in this task to improve repeatability.

Finally, task T3.3 has established a data protection office that monitors the application of the data management policy and the practices related to data protection and privacy. The office has also created a set of guidelines concerning the management of SLICES Intellectual Property Rights (IPR). A number of repositories, including a CKAN server, have been installed in the context of this task to publish the open datasets.

1.2. Objectives of D3.2

The present deliverable D3.2 reports specifically on the work carried out in the tasks T3.2 and T3.3. Indeed, this document presents the solutions implemented to ease the reproducibility of experiments in the SLICES Research Infrastructure and ensure adequate dissemination of the research outcomes, as well as their validation. The results concerning the investigations on benchmarking are also described in this deliverable.



Concerning task T3.3, the data protection office is presented more in detail, explaining its operation methods, rules, and guidelines, as well as its role within the SLICES infrastructure. Additionally, this deliverable provides an overview of the processes that can be completed through the website of the Data Protection Office, highlighting its importance for data protection within the SLICES ecosystem.

Finally, this document describes the tools put in place to facilitate data management, in particular for the different kinds of open data encountered in the context of the SLICES-SC project, focusing particularly on compliance with the FAIR principles. It also provides the necessary updates to the first iteration of the Data Management Plan, explaining partners' data-related activities, as well as the safeguards they have put in place to ensure data security and privacy.

2. Overall Structure of the Deliverable

This deliverable is divided on two main parts:

- Reproducibility framework for SLICES (corresponding to task T3.2);
- Data management methods of the SLICES project (corresponding to task T3.3).

It is followed by a set of recommendations for future works and a conclusion.

More specifically, the deliverable introduces the importance of reproducibility in the context of SLICES (section 3). The best practices and tools to ensure the reproducibility of experiments are then presented in section 4 with the aim to provide a practical perspective to this issue. Both these sections are then enriched with a methodology for reproducibility and related guidelines for SLICES-SC reproducibility (section 5). Lastly, the enhancement of the experimentation as a service is presented and followed by a discussion on reproducibility test and validation.

The document then introduces data management, where it presents a high-level overview of ethical and regulatory compliance (section 6), which refers to the ELES report (SLICES-SC D7.2). Section 7 introduces the updates to the data management plan, particularly as related to individual consortium partners. Updates are provided in regard to their data management and data security, as well as general progress on data management solutions. Section 8 provides an update on the FAIR data management principles and how the project is realizing these principles in practice, while section 9 briefly discusses other issues of relevance to data management.

Finally, Section 10 introduces the Data Protection Office and explains its key features, and section 11 details the project's personal data monitoring efforts and coordination. Section 12 provides associated recommendations for the future of the SLICES infrastructure, both on specific reproducibility issues and on future data management activities.

3. Importance of Reproducibility for SLICES

Reproducibility has been one of the cornerstones of scientific research for a long time and it is one of the main elements of the European Code of Conduct for Research Integrity. As such, it is equally important in the scope of the SLICES-SC project and beyond. That being said, it is important to first define and make the necessary distinctions between the terms of reproducibility and replicability, as sometimes they are understood in different ways.



These terms have very creditably been analysed in “Reproducibility and Replicability in Science”,¹ defining reproducibility as “obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis”. Reproducibility in this context does not entail using new data but the ones that were already obtained in the original research that is being reproduced. It also includes using the same methods and analyses.

Replicability, on the other hand, is defined as “obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data”. Unlike reproducibility, in replicability new data is being generated in research that does not use the same methods or conditions of analysis and still reaches the same results when answering the same research question.

The importance of reproducibility is included in its own definition, namely providing the ability to produce the same results using the same data when performing the experiment. Through this procedure, researchers can repeat previous tests, observe and acquire more information not only on the results but also the used methodology. In this way, the experiment, methodology, and results can be validated and reviewed, in line with the basic scientific principles. Reproducibility is essential not only for further research but also for the original experiment, as it enables its repetition and, thus, helps avoid errors.

3.1. European and ESFRI Policy

The ESFRI (European Strategy Forum on Research Infrastructures) policy is aligned with the European policy concerning open research data and reproducibility. Indeed, according to said policy, reproducibility should be ensured through the high quality of open research data. To that end, the ESFRI is supporting the development of the EOSC (European Open Science Cloud), a space where the various Research Infrastructures (RIs) can easily and freely share their open data. The basic idea is that each Research Infrastructure should follow the FAIR principles for the open data generated and used in the different nodes and sites, in line with the overall European strategy on research. Then, the data are publicly shared on EOSC where the other Research Infrastructures can reuse them with ease.

In this context, the interoperability of open research data is critical when the data are used in the context of multidisciplinary experimentations. Furthermore, the ESFRI policy encourages Research Infrastructures to work on ensuring the transparency and quality of their data, as well as providing adequate research acknowledgement and training resources in the context of open research data.

4. Best Practices and Tools for Reproducibility

4.1. Best Practices for Reproducibility

Adopting and implementing widely recognised best practices in the context of experiment reproducibility are crucial to ensure the correct sharing of open research data. To achieve this goal, the FAIR principles have been incorporated from the design of the SLICES activities and are used in the

¹ National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. <https://doi.org/10.17226/25303> [Last accessed 17 July 2024]



entire lifecycle of the open research data generated by the SLICES-SC project and its connected projects, such as SLICES-PP.

In this regard, reproducibility is closely connected to adequate data management practices, in order to ensure the dissemination of research results of high quality, that would inspire trust in the SLICES infrastructure, thus benefiting both the research community and society as a whole. As such, it has been essential from the start for the SLICES Research Infrastructure to put in place all the components, services, and processes guaranteeing proper data management of each research linked to ICT.

Taking the above into consideration, a number of elements were addressed within SLICES in order to improve the reproducibility of experiments conducted therein. In particular, the first element considered is the need to facilitate access to and the discovery of open research data at such a large scale. In order to address this need, a sufficient number of storage resources have been made available in the SLICES Research Infrastructure to allow the parallelisation of experiments using the open research data in the SLICES-RI architecture, as will be detailed below.

In addition to the above, data quality assurance has been highlighted as a major factor in the research infrastructure's success and wide-spread adoption. As such, data quality assurance measures have been adopted both within the project and within the upcoming SLICES Research Infrastructure, aiming at verifying, among others, the data integrity and accuracy.

Additionally, cybersecurity and data protection were particularly considered in the context of reproducibility, in order to protect the integrity of the open research data. The present deliverable will delve deeper into some of the practices that were implemented to that end both at a partner and a project level.

Finally, the application of metadata profiles has also been noted as a great practice to ensure the proper sharing and interoperability of data used in different experiments. Metadata can also indicate the provenance and the lineage of the associated open research data, further enhancing its reproducibility aspects.

In order to ensure reproducibility in line with the above elements, a number of tools for reproducibility were studied and used in the context of the SLICES-SC project to ensure the reproducibility of the experiments conducted within and beyond the project, as further analysed below.

4.2. Tools for Reproducibility

4.2.1. Reproducibility in the Fed4FIRE project

4.2.1.1. Introduction

In the Fed4FIRE+ project (2017-2022), a specific task 3.2 was dedicated to the topic of Experiment-as-a-Service, data retention and reproducibility of experiments. The deliverables prepared by WP3 have more details on this. In this section, the relevant work performed in Fed4FIRE on this topic that were integrated into the SLICES ecosystem and methodologies will be summarized. Most notably, the following tools and frameworks will be discussed:

- ESpec (Experiment Specification): a framework to provision resources and (complex) software/start up scripts. As flexible to e.g. install OpenStack on a bunch of bare metal nodes;
- Expo: a lightweight experiment orchestration tool to be used during the experiment phase to coordinate an experiment on multiple nodes;



- Jupyter notebooks to help in reproducing experiments;
- A metadata bundler to make it easy to collect a lot of information on the used resources;
- Distrinet: a tool to scale up network experiments while making them reproducible as well

4.2.1.2. ESPEC (Experiment Specification) and jFed

The Experiment Specification format is not a replacement for the RSpec format (defined by the GENI/Fed4FIRE APIs). An ESPEC contains an RSpec and combines it with other files.

- The purpose of an RSpec is to define which resources are needed;
- The purpose of an ESPEC is to additionally define which files should be placed where, and which scripts should be started.

The current ESPEC specification already allows some complex ESPECs. However, the base idea when using an ESPEC should be to keep it simple, and to put as little as possible in the ESPEC. It is meant for easily bootstrapping an experiment, but not for running experiment logic.

In contrast, it is preferable to keep ESPECs so simple that a user not familiar with the format will be easily able to manually execute the experiment using tools that do not support ESPEC.

The ESPECs can be used with the jFed client tool (<https://jfed.ilabt.imec.be>), both in the GUI version and the command line version.

We defined e.g. an ESPEC that deploys OpenStack via the EnOS framework.

For a full description, see <https://jfed.ilabt.imec.be/espec>, while below can be found some complex examples:

```
version: 1.0-basic
rspec:
  - bundled: 3-nodes.rspec
upload:
  - exp-files-set1.tar.gz
  - bundled: exp-files-set2.tar.gz
    path: /tmp
    nodes: [central, exp1]
  - download: http://example.com/exp-files-set3.tar.gz
  - direct: |
    You can also directly specify the content of a file. This text will thus be stored on all nodes in /tmp,
    Check the yaml syntax of "literal-blocks" for details about syntax and removing indentation
    path: /tmp/demo.txt
execute:
  - bundled: setup-central-node.sh
    nodes: central
  - bundled: setup-exp-node.sh
    nodes: [exp1, exp2]
  - local: /work/repo/start-exp.sh
    nodes: [exp1, exp2]
```

Figure 1. ESPEC example with the new functionality of 'direct' for literal blocks of data



```
version: 1.0-basic
rspec: my-experiment.rspec
dir:
  - path: /work/ansible/
    content: ansible
ansible:
  host:
    type: EXISTING
    name: control
    galaxy-command: /usr/local/bin/ansible-galaxy
    playbook-command: /usr/local/bin/ansible-playbook
    execute:
      - my-custom-ansible-install.sh
  galaxy:
    - download: http://example.com/ansible-requirements.yml
    - my-ansible-requirements.yml
  playbook:
    - bundled: setup-software.yml
    debug: 2
    - run-1st-experiment.yml
    - run-2nd-experiment.yml
  group:
    servers:
      - server1
      - server2
    clients:
      - client1
      - client2
```

Figure 2. ESPEC example with Ansible scripting, supporting the new item 'galaxy' for Ansible galaxy (pre-packaged units of work)

4.2.1.3. Experiment Orchestration Tool (ExpO)

ExpO is short for "Experiment Orchestrator". It allows you to run time-sensitive experiments over multiple machines in a very light way and can be used after the provisioning phase (manual provisioning through a tool/scripts or automated e.g. with ESPEC).

ExpO consists of two pieces of lightweight software:

- the **ExpO slave** which is present on all machines participating in the experiment, waiting for instructions on when to execute commands
- the **ExpO director** which executes experiments defined in an Experiment Orchestration definition

See for the full details at: <https://gitlab.ilabt.imec.be/ilabt/expo>

An example of an orchestration definition can be found below:



```
version: 1.0
nodes:
  node1: [groupA]
  node2: groupB
  node3:
    - groupA
    - groupB
  node4
commands:
  - command: "iperf -s"
    groups: [groupA]

  - after: 10
    id: iperf_client
    command: "iperf -c server"
    groups: [groupB]

  - command: echo "Hello $EXAMPLE_NAME"
    args:
      chdir: /tmp
      environment:
        EXAMPLE_NAME: 'World'
    nodes: node4
```

Figure 3. Example of an ExpO orchestration definition (including the new 'environment')

The file format

version Currently always 1.0

nodes List of nodes expected in the experiment. Optionally you can specify to which groups the node belongs in this experiment

commands: List of commands to be executed

Command format

command: the command to execute

after: number of seconds after the previous command that this command must be executed (optional, if omitted, it will be executed immediately after the previous command)

daemon: whether to keep this command running in the background. Used by the CLI to know if it should wait for the command to finish or not.

groups: the groups which must execute the command (optional if **nodes** has been specified)

nodes: the nodes which must execute the command (optional if **groups** has been specified)

id: a custom ID to identify MQTT-messages which relate to this command, such as the result, stdout, stderr, stdin (optional, if omitted, the index of this command in the **commands**-list is used as an id)

stderr: whether to stream the stderr as MQTT-messages with topic `expo/<experiment_id>/node/<node_id>/cmd/<command_id>/stderr` (defaults to True)

stdout: whether to stream the stdout as MQTT-messages `expo/<experiment_id>/node/<node_id>/cmd/<command_id>/stdout` (defaults to False)

stdin: whether to stream data sent to MQTT topic `expo/<experiment_id>/node/<node_id>/cmd/<command_id>/stdin` or `expo/<experiment_id>/group/<group_id>/cmd/<command_id>/stdin` to the stdin of the process (defaults to False)

Warning: the order in which data is streamed to the stdin of a process is indeterminate when mixing both topics

Figure 4. Documentation of the ExpO orchestration definition file

4.2.2. Jupyter notebooks for reproducible experiments

Some of the testbeds, in particular, imec’s GPU Lab and Inria’s GRID’5000 offer Jupyter notebooks to the experiments which help in reproducing and sharing experiments.

4.2.2.1. Jupyter notebooks at imec’s GPU Lab testbed

As per the official website (<https://doc.ilabt.imec.be/ilabt/jupyter/index.html>), it is possible to use GPUs or only CPUs to use the Jupyter notebooks. This is typically used for the first steps in machine learning, classes and quick experimentation (GPU Lab with a job-based workflow is then used for longer running more advanced jobs).

Below is provided a screenshot of the environment, with an example class on machine learning that has been used to that end.

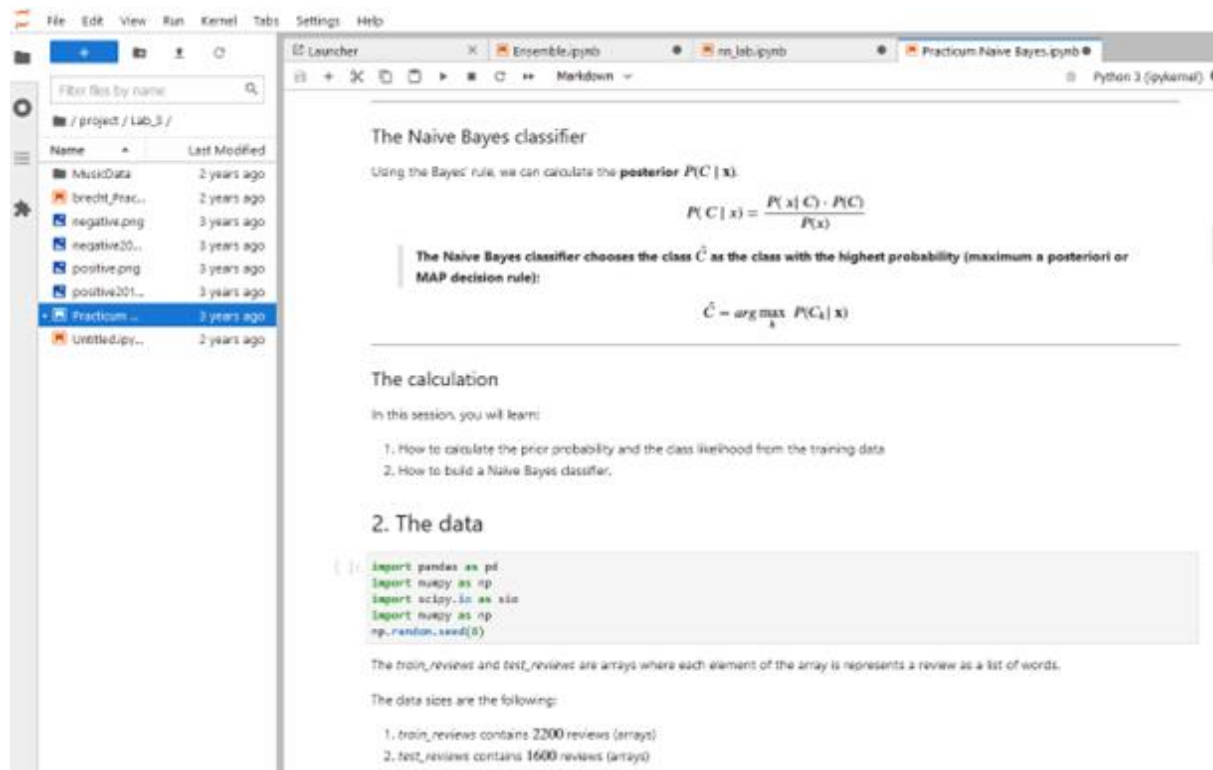


Figure 5. Jupyter notebook at imec’s GPU Lab testbed

4.2.2.2. Jupyter notebooks on Inria’s GRID’5000

Computer science testbeds often require extensive experiment automation to be used efficiently. Jupyter notebooks can contain scientific reasoning, experiment orchestration, and experiment results in an easy-to-read, easy-to-execute format. However, some level of support is still required from the testbeds to facilitate the notebooks’ functioning on their platforms. Additionally, this format can be used for more than driving experiments, while it is worth noting that different uses have different requirements in terms of support.



During research, a lot of code is often generated to collect and process data. Often, this code used to remain unpublished, while researchers simply described the process in the corresponding article, if any. Even when it was published, the code was often relatively undocumented and hard to follow. In proposing their notebook format integrating code and prose, the Jupyter team hoped to foster readable code for reproducibility.

Jupyter notebooks combine ideas of literate programming, interactive programming, and laboratory notebooks. Notebook files contain code cells interspersed with text. Additionally, outputs generated by the executed code cells are also saved in the notebook. The Jupyter notebook application is designed to view, edit, and run notebooks and the execution of code cells is handled by Jupyter kernels.

4.2.2.3. General setup within SLICES

Within SLICES, it was decided to deploy a single JupyterHub instance for the whole of g5k. This instance is placed on a service machine in the internal g5k network, with a reconfigurable proxy executed on the same machine as the hub. To allow users' access to the reconfigurable proxy, and, thus, the hub and labs, a route is added to g5k's pre-existing Apache proxies. This frontline proxy is already used for other g5k services, can handle user authentication, and is meant to mitigate any weakness the reconfigurable proxy might have.

The authentication module used in the hub is `jhub_remote_user_authenticator`, which removes the authentication page and instead relies on incoming HTTP headers for authentication. This allows the frontline proxy to perform authentication and pass on authentication information to the hub through HTTP headers. This is advantageous as it uses an already established g5k authentication infrastructure without involving a new service.

The spawner module is a custom module implemented specifically to match g5k's needs, called `G5kSpawner`. From the users' point of view, the spawner requires the user to select a site and decide whether to start the lab on a front end or a node. If a node is selected, the users can specify the OAR resource tree to request, the wall time of the OAR job and other information required by OAR to service the request.

For the execution of lab instances on compute nodes, the hub delegates the execution of the Jupyter-labhub program to OAR, which it controls through g5k's REST API.

The REST API allows the hub to interact with OAR instances of every site and as a service operated by g5k, the hub is provided an SSL client certificate allowing it to make calls to the API in the name of the user requesting the lab instance. When requesting resources from OAR the hub provides the lab command to execute. OAR will execute the command automatically once the requested resources become available on the main node of the request. If the Jupyter-labhub command ends before the end of the OAR job, the job will be ended immediately by OAR. As such, the lifecycle of the OAR job and the lifecycle of the lab instance are closely interlinked and the hub treats them as one and the same.

Under this mode of operation, the three main functions operate as follows: The close interlink between the lab and the corresponding OAR job, the REST API, and the existence of the `python-grid5000` library used to interact with the API greatly simplifies the code used when instantiating labs on nodes.



4.2.2.4. Grid'5000 metadata bundler

When running experiments on Grid'5000, users generate metadata across multiple services. Said metadata is useful for reproducibility purposes, as well as for scientific dissemination. The g5k-metadata-bundler is a tool designed to retrieve metadata across all the different services and bundle them in a single archive. The bundle only retrieves metadata generated by Grid'5000 services, the collection of data generated by the user experiment is beyond the scope of this application.

More information on this procedure can be found at [https://www.grid5000.fr/w/Grid5000 Metadata Bundler](https://www.grid5000.fr/w/Grid5000_Metadata_Bundler), while an example of the metadata bundler usage can be found below.

Usage

G5k-metadata-bundler is installed on every site frontend in Grid'5000. It can only be executed from the site frontends.

```
g5k-metadata-bundler -s SITE -j JOBID [-o NAME]
-v, --version          Print g5k-metadata-bundler version
-s, --job-site SITE    [MANDATORY] Grid'5000 site from which to extract
-j, --job-id JID       [MANDATORY] Job id of the OAR job to extract
-o, --output NAME      Bundle name to use for the directory/archive
```

Users do **not** need to operate the bundler on the same frontend as the site the jobs was executed on. The bundler download all data pertaining to the queried job and bundle in a archive named `g5k-bundle-SITE - JID .tar.gz` or if an output name has been provided `NAME .tar.gz`. The bundle is provided in as a tar.gz archive which can be manipulated by using the following commands:

• Listing

```
tar -tzf NAME .tar.gz lists all files contained within the bundle
```

Extraction

```
tar -xzf NAME .tar.gz extracts all files to a directory with the same name as the bundle
```

Users operating on older versions of Windows might require third party software to unpack the bundle. (often 7-zip)

Example usage

```
user@fsophia:~$ g5k-metadata-bundler -s nancy -j 3003030
Running g5k-metadata-bundler for job 3003030 at nancy
Downloading https://api.grid5000.fr/stable/sites/nancy/jobs/3003030
Downloading https://api.grid5000.fr/stable/sites/nancy/clusters/graouilly/nodes/graouilly-1?version=7f6b81c2621c6ed3a4fac632f213436813495755
Downloading https://api.grid5000.fr/stable/?version=7f6b81c2621c6ed3a4fac632f213436813495755&deep=true
Downloading https://api.grid5000.fr/stable/sites/nancy/metrics?job_id=3003030&nodes=graouilly-1
Generating README
Compressing bundle
Bundle created at g5k-bundle-nancy-3003030.tar.gz
user@fsophia:~$ ls -lh g5k-bundle-nancy-3003030.tar.gz
-rw-r--r-- 1 user g5k-users 456K Jul 19 09:50 g5k-bundle-nancy-3003030.tar.gz
user@fsophia:~$ tar -tzf g5k-bundle-nancy-3003030.tar.gz
g5k-bundle-nancy-3003030/
g5k-bundle-nancy-3003030/g5k-oarjob-nancy-3003030.json
g5k-bundle-nancy-3003030/README
g5k-bundle-nancy-3003030/g5k-resource-nancy-graouilly-1-7f6b81c2621c6ed3a4fac632f213436813495755.json
g5k-bundle-nancy-3003030/g5k-monitoring-nancy-graouilly-1-3003030.json
g5k-bundle-nancy-3003030/g5k-refapi-7f6b81c2621c6ed3a4fac632f213436813495755.json
```

Figure 6. Grid'5000 metadata bundler usage and example

The bundle contains these different file types:

- `g5k-oarjob-SITE-JID.json`: Job files



Contains the information for a given OAR job JID at Grid'5000 site, such as:

- Submission, start, and end dates
- User and group (group granting access) of the job
- Job types and properties
- Command executed by the job
- List of resources attributed to the job
- OAR events for the job

This information is extracted from the jobs API as:

- *g5k-resource-SITE-NODE-VERSION.json: Resource files*

Contains information about a single NODE, such as:

- Node architecture, bios, ram, and CPU information
- Network, storage, and monitoring devices
- Base configuration information

The bundle will contain one such file for each of the nodes involved in a job.

This information is extracted from the reference API

The VERSION of the information contained in this file will match what it was on the day the job was executed.

- *g5k-refapi-VERSION.json: Reference API files* Contains a full copy of the reference API at VERSIONsThis can be used to look up information about nodes not directly used by the bundled oar jobs.sThe resources files are a subset of this file.s*g5k-monitoring-SITE-NODE-JID.json: Monitoring files* Contains all the monitoring measurements made by [Kwollect](#) for a NODE during the oar job JID.sThe contents will vary depending on how much monitoring was enabled for a given job. See default metrics in [Monitoring Using Kwollect](#).sOften the heaviest files in the bundlesThis information is extracted from [Kwollects](#)
- **README:** *The readme files* contain information pertaining to the execution of the bundler such as:
 - Bundler version
 - Execution date
 - List of warnings and errors that happened during bundling
 - List of files included in the bundle with a short description

Users will also find at the end of every file a small bundler information segment. This segment contains the date at which the file was generated, warnings raised by the file generation and a list of references indicating how this file relates to other files in the bundle.

As the bundler is still in alpha version, comments and feature requests are welcome.

The following is a list of features in the process of review:

- Concerning bundle contents
 - Bundling multiple jobs in a single archive
 - Bundling based on job names
 - Better management of monitoring information when it is too big for download
 - Adding information concerning the standard environment to the bundle
 - Adding image deployments to the bundle



- Adding information concerning the deployed images
- Concerning bundler operation
 - No-compress mode where the bundle is left as a directory containing all files
 - Appending new files to an existing bundle
 - Reduce memory footprint
 - Assess viability of parallel downloads

4.2.3. *Distrinet (elay-based fidelity monitoring of network emulation)*

On the Fed4FIRE testbeds, it is possible to run large-scale networking experiments, but they are restricted to the physical limits (e.g. only up to 11 network cards on nodes), so it's not an infinite scale. Mininet on the other hand is a tool for network simulation but is limited to a single machine. If the power of the testbeds is used with multiple nodes and a distributed version of mininet on this is deployed, that would scale to huge networks.

The objective of Distrinet was to enhance network experiment tools in order to address the orchestration of large-scale experiments on grid and cloud environments and make it easier to automatically reproduce experiments. This section introduces Distrinet-HiFi: a Distrinet plug-in to monitor the fidelity of emulated experiments based on the measurement of packet delays.

4.2.3.1. Prerequisites

The scripts given here use [apssh](#) and [asynciojobs](#) to remotely run parallel commands on a number of nodes. First, it is important to note that is necessary to have a recent version of Python (≥ 3.6) in order to proceed to the installation of the following on the local computer:

```
pip3 install apssh asynciojobs
```

A slice in R2Lab is also required, while the computer must be able to log onto the gateway node. If this is not the case already, it is possible to [register](#) for an account.

If a cluster of computers is preferred to deploy Distrinet-HiFi and/or run the proposed experiments manually, this step can be ignored.

4.2.3.2. Installation

To set up the experiment, the R2Lab nodes must be correctly configured and running the latest stable version of Distrinet. Already available images can be used to set up the testbed, with one master node and one or more worker nodes:

```
rhubarbe load -i u18.04-distrinet_hifi_leader $LEADER_NODE
```

```
rhubarbe load -i u18.04-distrinet_hifi_worker $WORKER_NODE_1 $WORKER_NODE_2 ...
```

The testbed can also be manually installed if the use of R2Lab is not desirable. A recent Linux Kernel is required (the tool has been tested on v4.15.0) on all the nodes, then it is necessary to [install bcc](#) and [download and install Distrinet](#). The final step consists in copying hifi.py to the mininet code directory in the master node (`~/Distrinet/mininet/mininet/`), and the rest of the files to a `~/experiment/` directory previously created in all the nodes of the testbed.



4.2.3.3. Usage

The first step requires to import the Distrinet-HiFi library in the Distrinet script:

```
from mininet.hifi import Monitor
```

and wrap the experiment in the monitoring process:

```
monitor = Monitor(net)  
monitor.start()  
monitor.wait()  
# run the experiment...  
monitor.stop()  
monitor.receiveData()  
monitor.analyse()
```

Before running the experiment, it is necessary to initialise the monitoring agents on each node of the cluster:

```
python3 agent.py --ip=NODE_IP --bastion=LEADER_IP
```

5. Reproducibility Test and Validation

5.1. Reproducibility Tests and Validation Approach

The data repository (CKAN deployment, see the related section below for more details) has been evaluated and tested using several approaches and from various aspects:

- Users' point of view:
 - user experience, ergonomics
 - design options (customization options of the user interface, look-and-feel)
 - intuitiveness (easy interpretability, display of data set sizes)
 - availability
- Technical point of view:
 - User management (ACL, data protection from accidental changes/deletions by other users)
 - Data safety (replication/redundancy)
 - Organization of data (directories, structuring, grouping)
 - Storage details (limits of maximum data set, supported data formats)
- FAIR principles' point of view:
 - Finable
 - Accessible
 - Interoperability
 - Reusability

We note that we did not aim at testing the CKAN server's code at deeper levels (e.g., unit testing, domain testing, integration testing, as a white box), since these are tested by CKAN developers themselves (also see: <https://docs.ckan.org/en/2.10/contributing/test.html>), but the purpose was to

evaluate the solution from SLICES-SC's requirements, goals, and points of view at a higher level (black-box exploratory testing, system testing, smoke tests). Also, we made no or only very limited load and stress tests, since our focus was more on aspects of availability and findability rather than high-throughput, high-performance, real-time data delivery characteristics, although data size testing can be considered as a load testing activity. The GitLab repository (<https://gitlab.distantaccess.com/slices>) was also not tested, as it is maintained by GitLab developers and is already thoroughly tested (given that it is already a very widely used platform).

5.2. Technical validation

5.2.1. User experience

CKAN provides a very clear, intuitive, web-based user interface for the users, which basically requires no manuals to follow in order to use it. There are still manuals available by both CKAN developers (see: <https://docs.ckan.org>) and the ones written specifically by project members in the SLICES-SC project to aid users in data sharing (see: <https://ckan.iotlab.com/dataset/ckan-user-guide>).

User interface (UI) elements (fonts, labels, texts, graphical elements, menu, interactions, colours, etc.) are self-exploratory and easy to overview, read and interpret.

Moreover, CKAN is highly customisable (themes, logos, also metadata fields associated with data/resource elements). The only thing noticed during testing was that the SLICES-SC logo was originally missing from the UI, but it now appears in the current setup, as illustrated below.

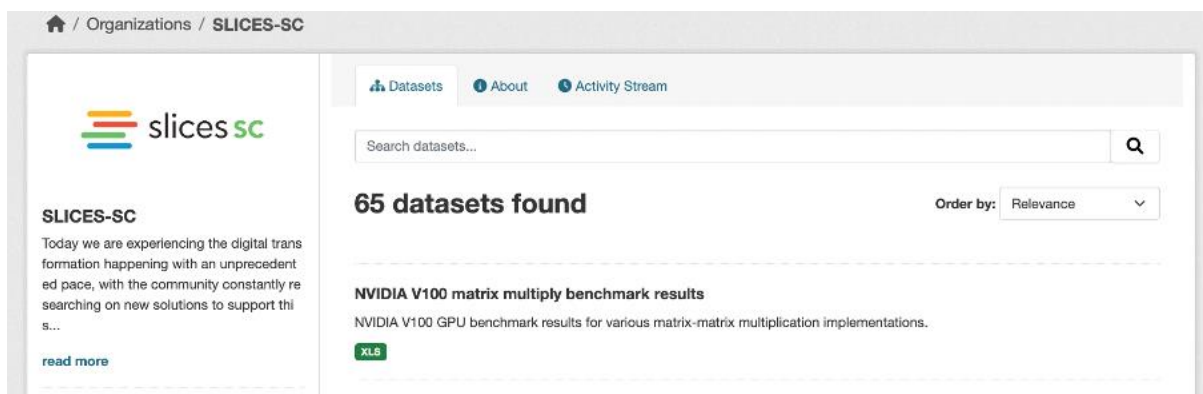


Figure 7. SLICES-SC CKAN server welcome-page

5.3. Technical Aspects

5.3.1. Command-line interface

CKAN also provides a command line interface (CLI) to access and manipulate data. This option is very advantageous when data processing is to be done automatically from programme coand des, shell scripts in a scheduled, orchestrated way in the background, requiring no manual user interventions. This option was not tested in detail in operation, but its presence was verified (see: <https://docs.ckan.org/en/2.10/maintaining/cli.html>).



5.3.2. Links to data

This is also a very useful feature of the CKAN server entailing that for every data element (resource) there is a dedicated direct link from where the data can be downloaded (over the HTTP protocol). This is vital from an automation point of view, especially considering that some other systems always require human interactions to access the data, which makes it impossible to process data in batches automatically. As will also be described later, the links provided pointing to data can be internal (using CKAN's database as storage) or external (specified by the author, uploader), respectively, which allows a great extent of flexibility to place data in remote storage as well (potentially using different technologies, storage backends, replication options, etc.).



Figure 8. URL to data available

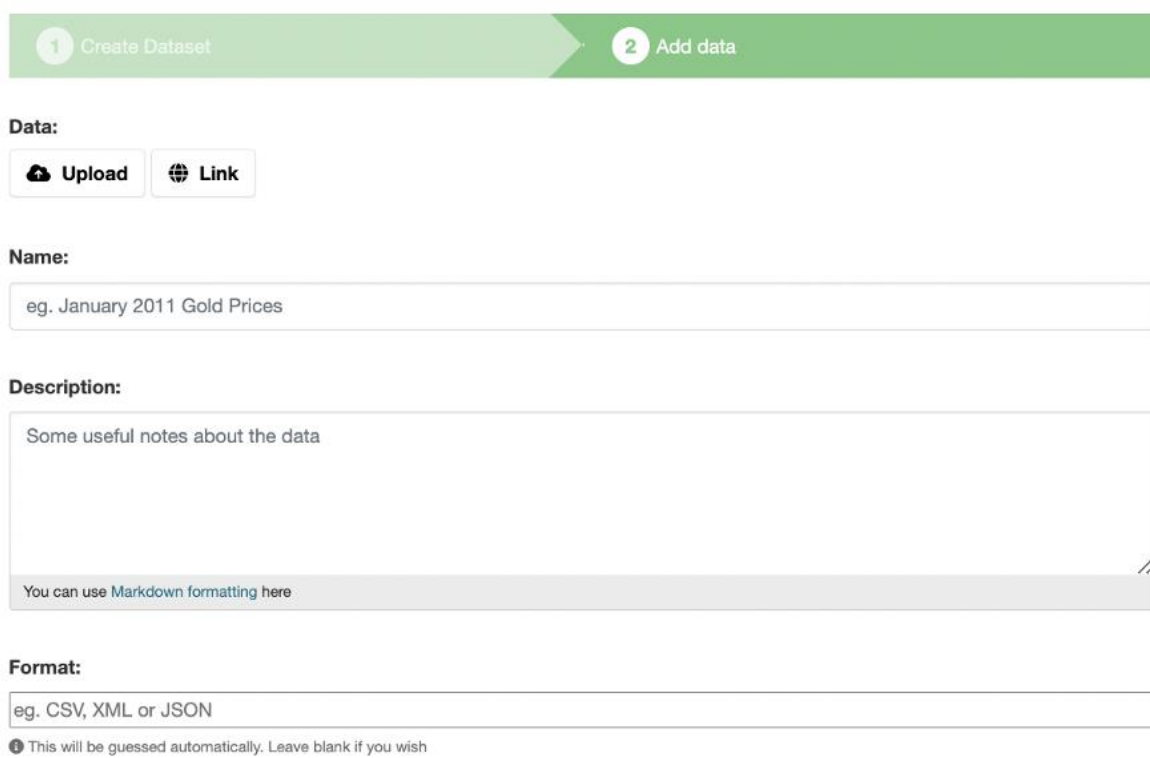


Figure 9. Interface for uploading data or creating a link



5.3.3. ACL and user management

In the initial version of the CKAN server, a single account (username and password) was offered for all the users (all SLICES-SC partners), which is a convenient way to upload and access the data storage but does not prevent users from accidentally modifying, deleting contents that do not belong to other users. This was pointed out during testing and in its subsequent version it has been fixed and improved to offer individual accounts for the different organizations. In this way, data are now protected by accidental changes and are entirely governed by the actual data owners/providers. The use of data is still open publicly for the community without restriction since authentication is required only by updates, uploads, and deletions.

5.3.4. Data size limits

In the first version of the CKAN setup, the maximum file size was set to default 10 MB (confirmed during load testing), which proved to be very low for realistic data. Due to the tests, this setting was soon increased to 100 MB by the operator, which then allowed uploading larger data volumes. Note that this threshold applies to a single entry, the amount of data of a data set composed of any number of entries can way exceed 100 MB in total without any further restrictions.

5.3.5. Data size display

In the tested first version of the data repository, though data entries contained an increased number of meta information, they lacked displaying the file size details, which became visible after the user downloaded the data element and looked at the file size. Since this is a piece of important information that the user should know in most cases prior to the download (eg. there might not be enough free disk space in the target computer), it was proposed to obtain and display this information among the metadata fields. As shown below, data size appears in the current version of the CKAN server.



Additional Information

Field	Value
Data last updated	March 17, 2024
Metadata last updated	March 17, 2024
Created	March 17, 2024
Format	XLS
License	Other (Public Domain)
Has views	False
Id	4a4a12aa-2ea5-4cec-80e4-11caa1e04131
Mimetype	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
Package id	650f3cab-5eae-4f33-bf8d-6a79d52426d9
Position	0
Size	66.8 KiB
State	active
Url type	upload

Figure 10. Metadata fields including size information

5.3.5.1. Availability

The responsiveness of the server (minimum lags at user interactions) and transfer speeds (upload/download rates) were appropriate from the users' point of view (based on manual experiences). There was no deep technical information about the backend infrastructure available, but as it will be described later, CKAN offers an option for replication (redundancy) of the storage of the operator's choice, which avoids potential data loss (and can increase throughput at downloads and uploads). When using external data storage for data elements, this aspect depends on the external storage setup. Also, there was no option to deeply investigate how the scaling procedure happens (in order to increase storage size backing up the database).

5.3.5.2. Federation

It was not planned or required to form federations of CKAN servers (and not addressed in the current project tasks). Nonetheless, *re3data* (<https://www.re3data.org/>) offers a registry for data repositories to search for, where SLICES-SC CKAN is already registered.



5.4. FAIR Aspects

5.4.1. Findability and Accessibility

CKAN offers a simple still powerful search and data “indexing” option via tags, as follows: an arbitrary number of tags can be assigned to data elements of users’ choice. This is the primary structuring mechanism for data sets in CKAN, and though, during the first time, it was difficult to get used to the fact that it is not possible to create directories and subdirectories for structuring data, switching to thinking in labelling. However, this feature proved to be very useful without limiting anything that traditional data organisations made possible.

For example, a directory “a” containing a subdirectory “b” in which a data element “c” is placed can simply be realised by adding tags “a” and “b” to data element “c”. Searching for data elements is possible using tags, so “c” can be found by searching for tags “a” and “b”, but also by searching for “a” or “b”. This associative structuring is thus very flexible and still allows very specific searches (with AND relation among all tags – as happens in traditional directory trees). The search options provided by CKAN is very rich (and well documented): advanced users can write advanced search expressions to find data and filter for specific metadata, e.g. data published by a specific author (“author:”), title matching a specific pattern (“title:Raw data~”), created in a specific time period (“+2022 -2023”). Though tagging is a very efficient and flexible way to structure data, there is also an option to create Groups of data sets.

As noted during testing, Digital Object Identifiers (DOIs) are not automatically generated individually for newly uploaded datasets, however, as described in subsequent sections, a DOI is available for the CKAN server (<http://doi.org/10.17616/R36S8M>). That being said, it is still possible to create DOIs for data sets individually using any DOI registry of user’s choice and indicate it among the metadata fields of the corresponding data element.

5.4.2. Interoperability, Reusability, and Metadata

CKAN provides a flexible way of how data must be and can be associated with metadata, which can even be extended with custom metadata fields and values. There are several mandatory metadata fields (most of them are automatically filled, thus, hidden at the upload form) such as upload/creation/modification date, id, owner, size information, and other optional/configurable fields such as mime-type, license, format, as illustrated below.

Custom fields make it possible to extend existing metadata and add further metadata to data resources (e.g., DOI, schema URLs, further data semantics description information) on demand.



Description:
eg. Some useful notes about the data
You can use Markdown formatting here

Tags:
eg. economy, mental health, government

License:
Please select the license
License definitions and additional information can be found at opendefinition.org

Organization:
SLICES-SC

Visibility:
Private

State:
Active

Source:
<http://example.com/dataset.json>

Version:
1.0

Author:
Joe Bloggs

Author Email:
joe@example.com

Maintainer:
Joe Bloggs

Maintainer Email:
joe@example.com

Custom Field:
Key: Value:

Figure 11. Metadata fields in CKAN

6. Regulatory and Ethical Compliance Updates

6.1. Regulatory Compliance Updates

Ever since the first iteration of the Data Management Plan, a number of changes have taken place with regard to the regulatory and legislative framework both at the European Union level, as well as



on a national level. Such examples may include the publication of the final text for the Artificial Intelligence (AI) Act, the Data Act, as well as the evolution of the cybersecurity framework.

An in-depth analysis of the regulatory evolution is provided in the SLICES ELES Report (Deliverable 7.2), which thoroughly analyses both the applicable legal and ethical framework, as well as its relevance for SLICES-SC and the SLICES infrastructure. As such, SLICES-SC has closely monitored said developments and has adapted its compliance activities and policies accordingly.

In this regard, and in order to ensure the compliance of SLICES-SC and the SLICES infrastructure as a whole, compliance is viewed not as a one-time activity, but as an ongoing and continuous exercise. As a result, and as will be further analysed in the following sections, compliance within SLICES is ensured through various mechanisms, including the establishment of the Data Protection Office.

6.2. Ethical Compliance Updates

The ethical considerations for SLICES have been outlined in the SLICES ELES Report Deliverable 7.2). The report has identified, in particular, security, data privacy, compliance and bias in AI, sustainability, fairness, transparency, integrity as the leading ethical considerations for the SLICES activities. In order to establish whether these issues present any threat to the development of the ERIC, the document analysed various pieces of legislation, including derived from national legislation of the most relevant EU Member States, and has mapped SLICES's compliance with them.

7. Data Management Plan

7.1. Data Summary

In the course of the SLICES activities, a variety of datasets have been produced and reused, as further examined in the upcoming sections. In accordance with the ethical and legal requirements, as examined in D7.2, as well as the FAIR data principles, SLICES aims to handle all data processing and data generation in a manner that complies with data protection and privacy obligations. When appropriate, SLICES aims to ensure interoperability and provide high-quality datasets in an accessible manner.

Overall, SLICES makes use of the data in a compliant and beneficial to society manner. The data used and created during the project is of vital importance not only for achieving the project's goals but also for the continuation of its activities. It will aid the evolution of the research community in Europe and will contribute to the European policies on the transition to a technological European economy, as well as policies promoting sustainability, energy efficiency, and education, among others.

In order to ensure the partners' alignment with data-related requirements, a consultation using questionnaires has been concluded, along with bilateral and multilateral discussions with partners. The results of these discussions and activities have highlighted the partners' related practices and shall be further analysed below.

In particular, the present and following sections delve deeper into the data-related activities of all partners, focusing more on those who have reportedly actively been involved in the generation and reuse of datasets, namely: MI, UTH, IMDEA, IMEC, SZTAKI, PSNC, INRIA



The questionnaire, presented in Annex consists of five parts:

- **Part 1 - Data processing activities:** gives an overview of the type of data used for the project as well as the means for its collection.
- **Part 2 - Data management:** This part contains descriptive and in-depth questions on the datasets if any has been used.
- **Part 3 - FAIR data:** The questions give an overview of measures taken by partners to ensure that datasets (if present) are in accordance with the FAIR principles.
- **Part 4 – IPR rights:** This part presents the generation of any IPR during the project.
- **Part 5 – Ethics and personal data protection:** The last part inquiries about the ethical and legal aspects of the data collection, storage, processing and purpose of use.

7.1.1. Types of Data

The SLICES-SC project allows the researchers to do experiments related to ICT. In this context, different sorts of data are generated, included source code and experiment results. Deliverables, surveys, material for dissemination and communication are also generated during the full duration of the project.

Several types of data were identified in the data management plan:

- Observational data
- Experimental data
- Simulation data
- Derived data
- Metadata

7.1.2. Formats of Data

The analysis of possible data formats was done in the data management plan and the outcomes of this analysis were:

- Open file formats
- Uncompressed files or compressed files with an open format
- Unencrypted files or encrypted files provided with appropriate mechanisms to decrypt them
- Commonly used formats such CSV or interoperable formats like JSON, XML and YAML
- Formats developed and maintained by open standards organisations

7.1.3. Purpose of data generation or reuse

The main objective of the SLICES-SC project is the creation of an initial online portal to let researchers doing experiments in the SLICES-SC testbeds. The SLICES-SC partners are investigating and analysing the possible solutions to implement the components used to realise the project objectives. In this context, the documents and data are generated to report the results of the studies made in the project.



7.1.4. Data Size

The size of the generated results of each experiment is variable and depends on the goals of the experiment.

The preliminary estimations done in the data management plan mentioned the following numbers:

- 5'000 users
- 50GB of data per user on the national nodes
- 1TB of data per user on the cloud
- 0.25PB-1PB of data storage on the datacentres hosting the nodes
- 5PB of data storage on the central datacentre which is cloud-based

7.1.5. Origin of generated or reused data

The data are generated and reused by the project partners and the researchers using the first services of the SLICES Research Infrastructure. The researchers using the SLICES-SC portal and other related services are the main users of the data provided in the context of SLICES-SC.

7.1.6. User Groups

Four user groups were examined in the DMP:

- Research User Group
- Industry User Group
- Research/Industry Support User Group
- External Partners User Group

7.1.7. Dataset License Types

The conclusions of the analysis of the different dataset license types made in the SLICES-SC data management plan were:

- The end-users should be able to choose a suitable license for their data.
- The license types should be available in the metadata.
- If a suitable license is not available, the Public Domain license should be the default license type.
- The applications or services retrieving the metadata should display the information related to the dataset license types.

7.1.8. Interaction with Other Infrastructures/Systems

Concerning the interaction with other infrastructures or systems, several requirements were elaborated in the data management plan:

- Services exposing information should be operational. For instance, metadata discovery service.
- Available resources should provide information to licensing, terms of utilisation, etc.
- Data should be completed by metadata to guarantee the discovery.



- Of course, data should be FAIR (Findable, Accessible, Interoperable and Reusable).
- Compliance with national and international regulations through appropriate procedures and processes.

7.2. Categories of data

SLICES involves the analysis and creation of multiple types of datasets. The data collected from other projects, namely SLICES-DS and SLICES-PP, has been reused also by the SLICES-SC project. These datasets are meant to support the goals of the project and fulfil particular functions, such as dissemination and organization, future references, contributing to the open access to facilitate exchange between RIs and other stakeholders, reproducibility. More specifically, the following categories of data have been identified:

- **Non-personal data:** The categories of non-personal data that are involved in the SLICES project are:
 - Open data generated through experiments on a CKAN server, which disseminates the publicly available data produced during the SLICES-SC trials.
 - Source code for SLICES-SC experiments, which is carried out within the framework of the SLICES-SC project and is accessible on a GitLab server.
 - Data on LTE Networks and testbeds is present in the form of statistics and real-time data.
 - Open data from weather readings
- **Personal data** is collected in four ways, namely:
 - Directly from data subjects who belong to the research team.
 - Directly from data subjects outside the research team: In the case of one of the datasets, the personal data is collected directly from applications to Open Call process; Another was created through registrations of participants in SLICES plenary meetings and training events.

Personal data collected involve emails, names, phone numbers, affiliations, IP addresses, while no special categories of data are being processed. When personal data is processed, a data protection by design and by default as per the GDPR was adopted and implemented. Moreover, strict policies and rules on how the data is handled according to data protection frameworks have been put in place. Said policies also extend to the processing of non-personal data, which is intended to be as open as possible, pursuant to the FAIR principles. Where applicable, minimisation, pseudonymisation, and anonymisation techniques have been deployed, as well as the erasure of the data and further appropriate technical and organisational measures.

The following Table provides an overview of the established datasets by each partner.

Partner	Name of Dataset	Type of data
MI	Open data generated through experiments	Non-personal data
	Source code for SLICES-SC experiments.	Non-personal data
UTH	UE NETWORK TRAFFIC TIME-SERIES (APPLICATIONS, THROUGHPUT, LATENCY, CQI) IN LTE/5G NETWORKS	Non-personal data





	UE STATISTICS TIME-SERIES (CQI) IN LTE NETWORKS	Non-personal data
	UTH Agricultural testbed dataset	Non-personal data
	UTH Energy measurements testbed	Non-personal data
	NITOS testbed usage/user access	Personal data
IMDEA	SLICES-SC events participants	Personal data
SZTAKI	Open Call applicants' data personal	Personal data
PSNC	Data with readings from the air quality and meteorological station from May 2023 in Poznan, Poland	Non-personal data
Imec	Account information of users through the Slices portal (https://portal.slices-sc.eu) and log information of users using our infrastructures	Personal data

Table 1. Categories of data and datasets

Furthermore, all partners were requested to complete the following table in order to ascertain whether and what data is processed by the consortium members.

Name of the used dataset(s)	
Short description of the dataset(s)	
If the dataset includes personal data, please specify the type of personal data.	
Purpose for which you use/ process the dataset(s)	
Format(s) of dataset(s)	
Where will you store the dataset(s)?	
What is the main source of the dataset(s)?	
Who owns the dataset(s)?	
Origin of the dataset	
Are there any restrictions for the use of the datasets?	
Who has access to the datasets?	
How long will you keep the datasets?	
Under which licence did you obtain access to the datasets?	
Additional comments	

Table 2. Dataset explanation table

Based on this, the following sections provide information on the datasets generated and/or utilised by partners MI, UHC, IMDEA, SZTAKI, PSNC, Imec and INRIA.





Name of the used dataset(s)	Open data generated through experiments
Short description of the dataset(s)	Mandat International is hosting and managing the CKAN server which publishes the open data generated in the context of the SLICES-SC experiments.
If the dataset includes personal data, please specify the type of personal data.	The datasets are not included any personal data.
Purpose for which you use/ process the dataset(s)	Publication of datasets for reuse.
Format(s) of dataset(s)	Compressed files (ZIP, RAR, 7z, TAR), JSON, XLS/XLSX, CSV, PDF, TXT.
Where will you store the dataset(s)?	In Switzerland at https://ckan.iotlab.com/organization/slices-sc
What is the main source of the dataset(s)?	The project partners.
Who owns the dataset(s)?	The project partners.
Origin of the dataset	The project partners.
Are there any restrictions for the use of the datasets?	No. Each dataset has a license.
Who has access to the datasets?	The projects partners, the researchers involved in ICT R&D activities.
How long will you keep the datasets?	As long as it is needed.
Under which licence did you obtain access to the datasets?	Each dataset has a license. The main license is Creative Commons Attribution, followed by Creative Commons Attribution Share-Alike. There are also public domain and open licenses.
Additional comments	

Table 3. MI Dataset(s)

Name of the used dataset(s)	Source code for SLICES-SC experiments.
Short description of the dataset(s)	A GitLab server is available to store the source code of the experiments done in the context of the SLICES-SC project.





If the dataset includes personal data, please specify the type of personal data.	No personal data are stored in the GitLab server.
Purpose for which you use/ process the dataset(s)	Storage of the source code for future reuse.
Format(s) of dataset(s)	HCL, Jinja, Shell, Jupyter Notebook, Python, Lua, Java, Scala.
Where will you store the dataset(s)?	In Switzerland at https://gitlab.distantaccess.com/slices
What is the main source of the dataset(s)?	The project partners.
Who owns the dataset(s)?	The project partners.
Origin of the dataset	The project partners.
Are there any restrictions for the use of the datasets?	No.
Who has access to the datasets?	The project partners, the researchers involved in R&D activities in the domain of ICT.
How long will you keep the datasets?	As long as it is needed.
Under which licence did you obtain access to the datasets?	Currently, no license was defined for the source code.
Additional comments	

Table 44. MI Dataset(s)

Name of the used dataset(s)	UE NETWORK TRAFFIC TIME-SERIES (APPLICATIONS, THROUGHPUT, LATENCY, CQI) IN LTE/5G NETWORKS
Short description of the dataset(s)	This dataset includes real-world time-series statistics from network traffic on real commercial LTE networks in University of Thessaly (UTH), Greece. The purpose of this dataset is to capture the QoS/QoE of three COTS UEs interacting with three edge applications.
If the dataset includes personal data, please specify the type of personal data.	N/A





Purpose for which you use/ process the dataset(s)	Research results, publication of research results
Format(s) of dataset(s)	.csv, .txt, .zip
Where will you store the dataset(s)?	SLICES-SC CKan server, IEEE Dataport
What is the main source of the dataset(s)?	UTH NITOS testbed, all the measurements have been collected in the testbed
Who owns the dataset(s)?	UTH personnel involved in SLICES
Origin of the dataset	NITOS testbed
Are there any restrictions for the use of the datasets?	No
Who has access to the datasets?	Open Dataset, Open-Source code
How long will you keep the datasets?	5 years
Under which licence did you obtain access to the datasets?	Self-generated datasets
Additional comments	

Table 5. UTH Dataset(s)

Name of the used dataset(s)	UE STATISTICS TIME-SERIES (CQI) IN LTE NETWORKS
Short description of the dataset(s)	This dataset includes real-world Channel Quality Indicator (CQI) values from UEs connected to real commercial LTE networks in Greece. Channel Quality Indicator (CQI) is a metric posted by the UEs to the base station (BS). It is linked with the allocation of the UE's modulation and coding schemes and ranges from 0 to 15 in values.
If the dataset includes personal data, please specify the type of personal data.	N/A
Purpose for which you use/ process the dataset(s)	Research results, publication of research results
Format(s) of dataset(s)	.csv, .txt, .zip
Where will you store the dataset(s)?	SLICES-SC CKan server, IEEE Dataport





What is the main source of the dataset(s)?	UTH NITOS testbed, all the measurements have been collected in the testbed
Who owns the dataset(s)?	UTH personnel involved in SLICES
Origin of the dataset	NITOS testbed
Are there any restrictions for the use of the datasets?	No
Who has access to the datasets?	Open Dataset, Open-Source code
How long will you keep the datasets?	5 years
Under which licence did you obtain access to the datasets?	Self-generated datasets
Additional comments	

Table 6. UTH Dataset(s)

Name of the used dataset(s)	UTH Agricultural testbed dataset
Short description of the dataset(s)	This dataset includes real-world metrics collected from multiple agricultural nodes, reporting environmental measurements, and channel quality measurements (zigBee, LoRAWAN, 5G).
If the dataset includes personal data, please specify the type of personal data.	N/A
Purpose for which you use/ process the dataset(s)	Research results, publication of research results
Format(s) of dataset(s)	.csv, .txt, .zip
Where will you store the dataset(s)?	NITOS testbed Self-hosted
What is the main source of the dataset(s)?	UTH Agricultural NITOS testbed, all the measurements have been collected in the testbed
Who owns the dataset(s)?	UTH Team
Origin of the dataset	NITOS testbed
Are there any restrictions for the use of the datasets?	No
Who has access to the datasets?	Open Dataset, Open-Source code





How long will you keep the datasets?	5 years
Under which licence did you obtain access to the datasets?	Self-generated datasets
Additional comments	

Table 7. UTH Dataset(s)

Name of the used dataset(s)	UTH Energy measurements testbed
Short description of the dataset(s)	This dataset includes real-world metrics collected from over 50 smart homes, reporting back real-time energy consumption, energy monitoring, remote actuation for smart devices, and link status for smart devices
If the dataset includes personal data, please specify the type of personal data.	N/A
Purpose for which you use/ process the dataset(s)	Research results, publication of research results
Format(s) of dataset(s)	.csv, .txt, .zip
Where will you store the dataset(s)?	NITOS testbed Self-hosted
What is the main source of the dataset(s)?	UTH Smart-Home testbed, all the measurements have been collected in the testbed
Who owns the dataset(s)?	UTH Team
Origin of the dataset	NITOS testbed
Are there any restrictions for the use of the datasets?	No
Who has access to the datasets?	Open Dataset, Open-Source code
How long will you keep the datasets?	5 years
Under which licence did you obtain access to the datasets?	Self-generated datasets
Additional comments	

Table 8. UTH Dataset(s)





Name of the used dataset(s)	NITOS testbed usage/user access
Short description of the dataset(s)	This dataset includes information on experimenters accessing the testbed and using the tested services
If the dataset includes personal data, please specify the type of personal data.	Users are identified through email addresses
Purpose for which you use/ process the dataset(s)	Testbed access
Format(s) of dataset(s)	.csv, .txt
Where will you store the dataset(s)?	NITOS testbed Self-hosted
What is the main source of the dataset(s)?	Testbed Usage
Who owns the dataset(s)?	UTH Team
Origin of the dataset	NITOS testbed
Are there any restrictions for the use of the datasets?	Yes, data needs to be anonymized first
Who has access to the datasets?	NITOS testbed administrators
How long will you keep the datasets?	5 years
Under which licence did you obtain access to the datasets?	Self-generated datasets
Additional comments	

Table 9. UTH Dataset(s)

Name of the used dataset(s)	SLICES-SC events participants
Short description of the dataset(s)	This dataset contains the name, affiliation and email of the participants of the Spanish National Roadshow, plenary meetings, training events
If the dataset includes personal data, please specify the type of personal data.	Name, affiliation, email and phone number
Purpose for which you use/ process the dataset(s)	Dissemination of information related to a particular event and potential announcement of future events related to SLICES-RI and SLICES-Spain.





Format(s) of dataset(s)	Excel
Where will you store the dataset(s)?	Intranet
What is the main source of the dataset(s)?	Registration form of the event
Who owns the dataset(s)?	IMDEA Networks
Origin of the dataset	Organization of the SLICES-SC national roadshow event, local events, plenary meetings, training events.
Are there any restrictions for the use of the datasets?	Restricted to members of the SLICES-SC consortium
Who has access to the datasets?	Restricted to members of the SLICES-SC consortium
How long will you keep the datasets?	5 years after the close of the SLICES-SC project
Under which licence did you obtain access to the datasets?	No license
Additional comments	

Table 10. IMDEA Dataset(s)

Name of the used dataset(s)	Open Call applicants' data
Short description of the dataset(s)	Potential experimenters apply for SLICES-RI access via submitting the proposal template .
If the dataset includes personal data, please specify the type of personal data.	name, email address, affiliation
Purpose for which you use/ process the dataset(s)	To organize the Open Call process.
Format(s) of dataset(s)	Excel file
Where will you store the dataset(s)?	DROPSU
What is the main source of the dataset(s)?	Application sheets received in either .pdf of MS Word format.
Who owns the dataset(s)?	SLICES project
Origin of the dataset	Directly from the applications.





Are there any restrictions for the use of the datasets?	No
Who has access to the datasets?	SLICES partners, who have access to DROPSU
How long will you keep the datasets?	As long as DROPSU is active. Personal folder on computer: 5 years after the project ends.
Under which licence did you obtain access to the datasets?	Not applicable
Additional comments	Not applicable

Table 11. SZTAKI Dataset(s)

Name of the used dataset(s)	Data with readings from the air quality and meteorological station from May 2023 in Poznan, Poland
Short description of the dataset(s)	Data with readings from the air quality and meteorological station from May 2023 in Poznan.
If the dataset includes personal data, please specify the type of personal data.	N/A
Purpose for which you use/ process the dataset(s)	Collect data from air quality and meteorological station
Format(s) of dataset(s)	XKS
Where will you store the dataset(s)?	Zenodo
What is the main source of the dataset(s)?	air quality and meteorological station
Who owns the dataset(s)?	PSNC
Origin of the dataset	PSNC
Are there any restrictions for the use of the datasets?	No
Who has access to the datasets?	It's open
How long will you keep the datasets?	No limitations
Under which licence did you obtain access to the datasets?	N/A





Additional comments	-
----------------------------	---

Table 12. PSNC Dataset(s)

Name of the used dataset(s)	Account information of users through the Slices portal (https://portal.slices-sc.eu) and log information of users using our infrastructures
Short description of the dataset(s)	Account information of users through the Slices portal (https://portal.slices-sc.eu) and log information of users using our infrastructures
If the dataset includes personal data, please specify the type of personal data.	Username, password, e-mail. First name, last name, student/academic researcher/industrial researcher, company/institution, city, country (https://portal.slices-sc.eu/signup)
Purpose for which you use/ process the dataset(s)	To be able to track who is using the infrastructure
Format(s) of dataset(s)	Database for the account portal
Where will you store the dataset(s)?	Gent, Belgium
What is the main source of the dataset(s)?	https://portal.slices-sc.eu/signup
Who owns the dataset(s)?	Imec
Origin of the dataset	People creating accounts
Are there any restrictions for the use of the datasets?	Only usage is to track who is using the infrastructures
Who has access to the datasets?	Imec system administrators
How long will you keep the datasets?	As long as the Slices portal needs to be online
Under which licence did you obtain access to the datasets?	N/A
Additional comments	

Table 13. Imec Dataset(s)





7.3. Overview of Personal Data Collection, Processing, and Safeguards

According to the GDPR, personal data must be protected on several fronts:

1. **Through the obligations and rights of the partners as controllers or processors.** From a regulatory standpoint, each partner may take on multiple roles and duties contingent on the activities conducted during the research project. Part 5 of the questionnaire on ethics and personal data protection has been created to give a better overview of the roles of each partner when dealing with data. If it is determined that any of the partners gave an affirmative response to the processing of personal data, they will need to follow the GDPR, as well as other relevant legislation to ensure the safety of the processing.
2. **Through the appointment of a Data Protection Officer (DPO).** The project's DPO oversees the project's adherence to all applicable regulatory requirements, both present and future, and manages the data protection strategy at the project level.
3. **Through the performance of a Data Protection Impact Assessment (DPIA).** Pursuant to Article 35 of the GDPR, a DPIA is conducted where: *a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data.*² Thus, if any of the processed datasets present a high risk to the users, the DPO shall perform a DPIA.
4. **By adhering to fundamental principles for data processing, including lawfulness fairness, transparency, purpose and storage limitation, data minimisation, accuracy, integrity, confidentiality and accountability.**
5. **By processing data lawfully, in accordance with one of the GDPR lawful bases.**
6. **By safeguarding data subjects' rights, interests, and freedoms and facilitating the exercise of said rights.**

As already highlighted above, MI, UTH, IMDEA, SZTAKI, Imec and INRIA deal with personal data. Additional details on the goals of the data collection and processing are given in the tables below, along with a summary of the security measures put in place by each of the pertinent partners.

MI

<i>For what purpose(s) did you collect the aforementioned personal data?</i>	User's email, name and password are collected to limit the access to the edition of the datasets and source code.
<i>Did you process the generated data for any further purposes than the ones it was originally collected for?</i>	No
<i>If you answered yes to the previous question, then please describe the purpose of this additional processing:</i>	N/A

² General Data Protection Regulation 2016, Article 35(1)



<i>How did you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?</i>	By email
<i>How did you plan to collect and document the consent of the data subjects whose personal data will be processed by you?</i>	Each new user of GitLab or CKAN must request a new account for GitLab or CKAN by email with the following required information: first name, last name and email address
<i>How and where did you store the data?</i>	In Switzerland, in Geneva, in our servers (GitLab and CKAN).
<i>For how long did you keep the data?</i>	As long as it is necessary. A user can request by email the deletion of his account and his demand will be handled almost immediately.
<i>Describe the measures in place to anonymize and/or pseudonymise the personal data whenever possible?</i>	We rely on the implementation made by the developers of GitLab and CKAN.
<i>Did you (plan to) share personal data with other partners within the project?</i>	Only to the data subjects themselves.
<i>What risk do you foresee for the individuals (data subjects) whose data will be processed by future instances of the SLICES project in the future?</i>	As the project is using different identity management systems, for instance those implemented by GitLab and CKAN, the attack surface is relatively large.
<i>What would be your suggestions to minimise the risks for the data subjects?</i>	To use only one identity/user manager across all SLICES-RI. This topic is under discussion in the context of the SLICES-PP project.

Table 14. MI's response to personal data collection, processing, and safeguards

UTH

<i>For what purpose(s) did you collect the aforementioned personal data?</i>	For managing user access to the testbed
<i>Did you process the generated data for any further purposes than the ones it was originally collected for?</i>	No
<i>If you answered yes to the previous question, then please describe the purpose of this additional processing:</i>	N/A



<i>How did you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?</i>	NITOS testbed terms of use
<i>How did you plan to collect and document the consent of the data subjects whose personal data will be processed by you?</i>	Users must accept the terms of use before accessing the testbed
<i>How and where did you store the data?</i>	NITOS testbed infrastructure
<i>For how long did you keep the data?</i>	5 years
<i>Describe the measures in place to anonymize and/or pseudonymise the personal data whenever possible?</i>	Generalization and data encryption are the two most common measures applied.
<i>Did you (plan to) share personal data/ transfer it/ make it publicly accessible/ perform data protection impact assessment?</i>	No
<i>What risk do you foresee for the individuals (data subjects) whose data will be processed by future instances of the SLICES project in the future?</i>	No particular risk, the data collected regards only the operation of the testbed
<i>What would be your suggestions to minimise the risks for the data subjects?</i>	No risks to be expected

Table 15. UTH's response to personal data collection, processing, and safeguards

IMDEA

<i>For what purpose(s) did you collect the aforementioned personal data?</i>	The main purpose is for managing the organization of project's events and for dissemination future events between potential participants.
<i>Did you process the generated data for any further purposes than the ones it was originally collected for?</i>	Yes
<i>If you answered yes to the previous question, then please describe the purpose of this additional processing:</i>	The data is processed for reporting dissemination activities of the project.
<i>How did you inform the individuals (the data subjects) about the purpose of the data</i>	The individuals are informed during the event registration process.



<i>processing of their personal data in the project?</i>	
<i>How did you plan to collect and document the consent of the data subjects whose personal data will be processed by you?</i>	The specific consents have been collected at the beginning of participating on each particular event.
<i>How and where did you store the data?</i>	Admin intranet.
<i>For how long did you keep the data?</i>	5 years after the official closure of the SLICES-SC project.
<i>Describe the measures in place to anonymize and/or pseudonymise the personal data whenever possible?</i>	In the case of datasets related to SLICES-SC events, it would be interesting the institution and business sector, instead of the exact personal details.
<i>Did you (plan to) share personal data with other partners within the project?</i>	Yes
<i>What risk do you foresee for the individuals (data subjects) whose data will be processed by future instances of the SLICES project in the future?</i>	The main risk is the no legitimate use of this personal data for external organizations that could use it for designing and generating cyber-attacks to the infrastructure.
<i>What would be your suggestions to minimise the risks for the data subjects?</i>	The main recommendation will be to define a global Data Policy that covers all the activities within SLICES-RI, not only at national level, but also at whole research infrastructure level.

Table 16. IMDEA's response to personal data collection, processing, and safeguards

SZTAKI

<i>For what purpose(s) did you collect the aforementioned personal data?</i>	This data is needed to organize and coordinate the Open Call process.
<i>Did you process the generated data for any further purposes than the ones it was originally collected for?</i>	No
<i>If you answered yes to the previous question, then please describe the purpose of this additional processing:</i>	N/A
<i>How did you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?</i>	The rules of the Open Call Application are detailed here .



<i>How did you plan to collect and document the consent of the data subjects whose personal data will be processed by you?</i>	They indicate their consent when submitting their application.
<i>How and where did you store the data?</i>	DROP-SU Personal folder on prem
<i>For how long did you keep the data?</i>	DROP-SU: according to DROP-SU rules Personal folder: 5yrs after the end of the project
<i>Describe the measures in place to anonymize and/or pseudonymise the personal data whenever possible?</i>	N/A
<i>Did you (plan to) share personal data/ transfer it/ make it publicly accessible/ perform data protection impact assessment?</i>	No
<i>What risk do you foresee for the individuals (data subjects) whose data will be processed by future instances of the SLICES project in the future?</i>	No risks are foreseen
<i>What would be your suggestions to minimise the risks for the data subjects?</i>	No suggestions

Table 17. SZTAKI's response to personal data collection, processing, and safeguards

INRIA

<i>For what purpose(s) did you collect the aforementioned personal data?</i>	Email addresses and IP addresses of the users: to keep a record of the users and for statistics
<i>Did you process the generated data for any further purposes than the ones it was originally collected for?</i>	No
<i>If you answered yes to the previous question, then please describe the purpose of this additional processing:</i>	N/A
<i>How did you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?</i>	When asking for an account to access the system, they will be asked to read and validate a consent form



<i>How did you plan to collect and document the consent of the data subjects whose personal data will be processed by you?</i>	Automatically in the user database
<i>How and where did you store the data?</i>	In protected storage (NAS) at INRIA, under the management of INRIA IT team
<i>For how long did you keep the data?</i>	As long as the account is active then one year after.
<i>Describe the measures in place to anonymize and/or pseudonymise the personal data whenever possible?</i>	We need to keep nominative information to detect any misuse of the platform.
<i>Did you (plan to) share personal data/ transfer it/ make it publicly accessible/ perform data protection impact assessment?</i>	No
<i>What risk do you foresee for the individuals (data subjects) whose data will be processed by future instances of the SLICES project in the future?</i>	Very low risk: they use the platform and we keep track of the users.
<i>What would be your suggestions to minimise the risks for the data subjects?</i>	No suggestions

Table 18. INRIA's response to personal data collection, processing, and safeguards

Imec

<i>For what purpose(s) did you collect the aforementioned personal data?</i>	For running the SLICES-SC system.
<i>Did you process the generated data for any further purposes than the ones it was originally collected for?</i>	No
<i>If you answered yes to the previous question, then please describe the purpose of this additional processing:</i>	N/A
<i>How did you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?</i>	During sign-up process for account.
<i>How did you plan to collect and document the consent of the data subjects whose personal data will be processed by you?</i>	During sign-up process for account.



<i>How and where did you store the data?</i>	In a database in a datacentre of Imec at Gent, Belgium.
<i>For how long did you keep the data?</i>	As long as the project is running.
<i>Describe the measures in place to anonymize and/or pseudonymise the personal data whenever possible?</i>	We don't anonymize this data.
<i>Did you (plan to) share personal data with other partners within the project?</i>	Only if needed when the partner's infrastructure is used by persons.
<i>What risk do you foresee for the individuals (data subjects) whose data will be processed by future instances of the SLICES project in the future?</i>	No risk, but people might need to create new accounts when the ERIC has been setup with a new user portal.
<i>What would be your suggestions to minimise the risks for the data subjects?</i>	Create new accounts for the operations phase. And expire those accounts. (now we only expire projects so users can not use the infrastructure anymore, but the account information is being kept).

Table 19. Imec's response to personal data collection, processing, and safeguards

7.4. Data Security

Art. 32(1) GDPR states that in order to ensure a level of security appropriate to the risk, the state of the art, the costs of implementation, and the nature, scope, context, and purposes of processing, the data controller and the data processor must implement the necessary technical and organizational measures (TOMs). The article outlines the following measures as granting appropriate level of protection:

1. the **pseudonymisation and encryption** of personal data;
2. the ability to ensure the **ongoing confidentiality, integrity, availability and resilience of processing systems and services**;
3. the ability to **restore the availability and access to personal data in a timely manner** in the event of a physical or technical incident;
4. a process for **regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing**.³

³ General Data Protection Regulation 2016, Article 32(1)



In addition to the above requirement and the standards/legislation identified and analysed in D7.2, partners have designed and implemented adequate frameworks so that data is secure throughout its collection, processing, retention, and transfer to third parties and/or countries.

Most notably, the SLICES partners have implemented appropriate measures across all stages of the project to ensure that the data is safely stored in trusted repositories for long-term preservation and curation. Overall, access to the data stored through the SLICES infrastructure requires HTTPS which is using TLS for the encryption of the data packets.

Furthermore, access to personal data is limited to authenticated and authorised users. Access to it can be obtained only upon request and approval of the request. The data itself is encrypted or anonymized, generalized or protected by GitLab and CKAN.

Below is an overview of the TOMs adopted by partners to ensure the security of the data.

Partner	What technical and organizational measures (TOMs) are in place to protect and secure the personal data?
MI	The personal data are encrypted and their access is limited to few people managing the GitLab and CKAN servers.
UTH	Access control, data encryption, data anonymization when necessary, logging and monitoring, policies and procedures as well as training.
IMDEA	The access to the datasets is restricted and any request for access should be evaluated from the admin team that manage the SLICES-SC related events datasets.
SZTAKI	We have an institutional Privacy Policy and an appointed DPO
INRIA	Data will be stored in computer room with protected access and unreachable from Internet.
Imec	Limited access to the servers + server security (firewalling)

Table 20. Technical and organisational measures

8. FAIR Data Access Policy

8.1. FAIR Principles Overview

The FAIR principles stand for: findable, accessible, interoperable, and reusable. They are developed to guide good data management practices for better data re-usability, particularly in scientific research, but are also broadly applicable across various domains. They are designed to also ease the access to data of the research community in a sustainable manner.

The FAIR principles allow better transparency and knowledge in the research process. Accessibility also promotes collaboration and enables to locate and use datasets more efficiently, saving time and effort. FAIR data is structured and formatted so that it facilitates interoperability, integration with other datasets, and use of various tools and platforms, enabling more comprehensive analyses. Also, it fosters better data quality and integrity, as well as providing detailed metadata and documentation.



Importantly, and as a result, FAIR enable other researchers to better understand and validate findings, enhance trustworthiness, credibility and reproducibility of research, which is essential for advancing scientific knowledge.

SLICES-SC applies the FAIR principles as presented in following sections.

8.2. Findability

Data should be easy to find for both humans and computer programmes. This involves creating and assigning unique identifiers to datasets, providing metadata (descriptive information about the data), and making the data discoverable through search engines or data repositories.

The data should be findable if the experimenter will reuse them later. To achieve this objective, several requirements should be taken into account:

1. **A globally unique and persistent identifier should be assigned** to the data and related metadata. SLICES-SC will use a DOI when the data are uploaded.
2. **Rich standardised metadata describe the data.** In SLICES-SC, the metadata will use the enhanced version of the DublinCore format.
3. **The metadata contain the identifier of the data that the metadata refer to.** The metadata in SLICES-SC will contain the DOI assigned to the data.
4. **Metadata are stored and indexed.** In SLICES-SC, a resource discovery service will permit to search the data thanks to queries containing keywords or metadata.

Data findability is ensured by the partners as summarised in the following table:

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met. A) Findable
MI	The data and related metadata have unique URLs and are searchable.
UTH	The datasets 1-4 that are reported are findable by storing them accordingly in repositories (e.g. CKAN, NITOS repository). The data in both repositories are annotated with respective metadata, describing content, context, and provenance. Each dataset has a unique ID (e.g., DOIs, URNs) allowing them to be reliably found and referenced over time.
IMDEA	Only internally in the repositories of the organisation and SLICES-SC project.
SZTAKI	N/A
PSNC	Data is published in Zenodo.

Table 21. Data findability according to partner responses



8.3. Accessibility

Once the data is found, it should be easily accessible, either through direct access or mediated access, with clear terms of use. This also means that data should be available in a format that can be understood, known by both humans and machines, and barriers to access such as authentication requirements should be minimised.

This principle ensures that the means to access the data is correctly described. It can include of course authentication and authorization. Several elements will be put in action in SLICES-SC:

1. **A standardised communication protocol is used to retrieve the data or the metadata through their identifier.** Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) will be employed in SLICES-SC. REST APIs with payloads using JSON, XML and YAML formats will be also supported.
2. **The standardised communication protocol is free, open and implementable by everybody.** The mentioned OAI-PMH protocol answers to this requirement. By definition, the REST APIs are also free, open and universally implementable.
3. **The communication protocol offers the authentication and the authorisation procedures, if necessary.** In SLICES-SC, the access to non-public data will be protected by authentication and authorization. On the other hand, the metadata will be completely open and so, no authentication and authorisation are required.
4. **The metadata remains available, even if the data are removed.** All the metadata will be stored in a dedicated data store in the SLICES infrastructure during the lifetime of the infrastructure and the project. So, the lifetime of the metadata will depend on the lifetime of the host data repository.

According to the questionnaires, data accessibility is ensured by the partners as described below:

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met. B) Accessible
MI	The data and the associated metadata are using standardized communication protocols to retrieve them. Furthermore, an authentication and authorization procedure are in place to limit the access to the edition of the datasets and the related metadata.
UTH	Datasets 1-4 are provided using an Open Access policy. CC licenses are applied on the datasets so as the terms under which they can be accessed and used are clear.
IMDEA	Only internally in the repositories of the organization and SLICES-SC project.
SZTAKE	N/A
PSNC	It is open to everyone.

Table 22. Data accessibility according to partner responses



8.4. Interoperability

Data should be structured in a way that allows it to be integrated with other data, processing algorithms, platforms and applications as much as possible. This can be achieved using standardised data formats, vocabularies, and ontologies, as well as following common data exchange protocols.

In consequence, the interoperability and the integration of the data are two important points to follow. In the context of SLICES-SC, the interoperability will be facilitated through the following topics:

1. **Knowledge representation is realized through a formal, accessible, shared and broadly applicable language.** As mentioned earlier, the enhanced version of the DublinCore format, which is standardized as ISO 15836, ANSI/NISO Z39.85 and IETF RFC 5013, will be used for the metadata. Concerning the REST APIs, JSON, XML and YAML will be the reference formats. It is not excluded to translate or map them in other formats.
2. **The data and the metadata should use vocabularies following the FAIR principles.** In SLICES-SC, the chosen vocabularies are ISO 3166 (country codes), ISO 639-3 (language codes), ISO 8601-1 (date and time representation), DCMI-Period and DCMI-Point. Other standards compliant with FAIR principles could be added in the future.
3. **Data and metadata should include qualified references to other data and metadata.** In SLICES, a metadata property named "Relation" will be used to convey the references.

Interoperability is further ensured by the partners as described in the table below:

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met.
	C) Interoperable
MI	The datasets are using well-known and interoperable data formats.
UTH	Datasets 1-4 use a standardised format for their metadata, so as they are interoperable. Other systems (e.g. the DMI of SLICES-RI) can understand data stored and annotated within SLICES-SC. Other APIs (e.g. REST) can be provided with the UTH data upon demand, allowing them to be used to other systems as well.
IMDEA	N/A
SZTAKI	N/A
PSNC	Data in the XLS file, which can be transformed to any required format.

Table 23. Data interoperability according to partner responses

8.5. Reusability

Data should be well-described and well-documented. This includes providing clear information about the data's provenance, semantics, how it was collected and processed, and any terms or conditions for its reuse if any. Additionally, data should be preserved and maintained over time to ensure its ongoing, long-term usability. As per the GDPR, data reuse is allowed under strict conditions.





The re-usability of the data is a key element in the FAIR principles. Several requirements should be respected:

1. **Data and metadata are described with accurate and relevant attributes.** In the context of SLICES-SC, the use of the enhanced version of the DublinCore format will ensure satisfactory description and efficient discovery.
2. **The data and the metadata are published with a clear and accessible data usage license.** A list of licenses will be provided by SLICES-SC to the researchers when they build their metadata. The Public Domain license will be the default license.
3. **The origin of the data and metadata is provided accurately.** In SLICES-SC, the experimenters should provide any changes concerning the data ownership. This is an important element to ensure the authenticity and the integrity of the published data.

The data and metadata follow the standards relevant to the research community. This requirement is met in SLICES by the DublinCore format which is already standardized as ISO 15836, ANSI/NISO Z39.85 and IETF RFC 5013.

Data Reusability will be facilitated by the partners by taking the measures summarised in the following table:

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met. D) Reusable
MI	Metadata encompass clear attributes to facilitate the reuse of the different datasets.
UTH	Datasets 1-4 are reusable, as they are sufficiently documented, describing the data, their collection methods, any processing steps, and potential use cases. The licenses under which the data are published are also available along with each dataset, allowing users to understand what they can do with the data, and what not.
IMDEA	The main reuse of these particular datasets is related to the dissemination and organisation of future events at SLICES national nodes.
SZTAKI	N/A
PSNC	Anyone can use the dataset for further processing

Table 24. Reusable data according to partner responses

8.6. Open Data Repositories

8.6.1. Open CKAN Server

A CKAN server hosting the open data generated and used in the SLICES-SC project was put in place at the beginning of the project. The link to the CKAN server is <http://ckan.iotlab.com/organization/slices->



[sc](#) and is listed in the re3data.org website which is a registry of research data repositories. The Digital Object Identifier (DOI) of the CKAN server is the following: <http://doi.org/10.17616/R36S8M>.

The following schema shows the different interactions with other WPs:

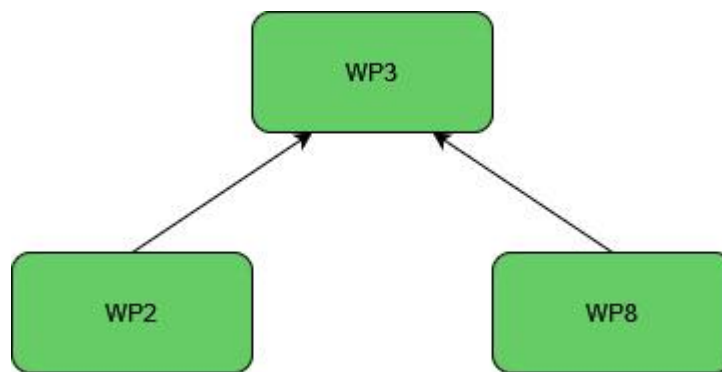


Figure 12. Interactions with other WPs

The CKAN open data server was put in place in the context of WP3, in particular, the Task T3.3 “Data Protection Office and Open data server”. Typically, the CKAN open data server is used by WP2, Task T2.3 named “Support of reproducibility methods for transnational access”, notably to store the results and the data used during the experiments. WP8 is also using the CKAN server for the implementation of the transnational and virtual accesses for the SLICES Research Infrastructure.


The open data can be directly stored on the CKAN open data server, but the CKAN server can also host the references to the data. It means that the effective storage of the data is done on the partners’ premises or on other open data repositories, such as Zenodo, but a reference to these datasets is in fact stored on the CKAN server. Here can be found an example of an open dataset referenced on the SLICES-SC CKAN server:

Home / Organizations / SLICES-SC / 5G QoS Measurements

5G QoS Measurements

Followers
0

Organization



SLICES-SC

Today we are experiencing the digital transformation happening with an unprecedented pace, with the community constantly researching on new solutions to support this... [read more](#)

Social

[Google+](#)

[Twitter](#)

[Facebook](#)

License

[Creative Commons Attribution](#)

Dataset Groups Activity Stream Related

5G QoS Measurements

Measurement data of a low-latency network functions. The measurement consists of a simple forwarder based on DPDK and a more complex intrusion prevention system (Snort 3). The dataset contains hardware-timestamped packet traces measured on the ingress and egress ports of the investigated device under test.

The data was published in two papers at [NOMS 2020](#) and in the [IEEE Comms. Magazine](#).

All data you find in this repository are [compressed pcaps](#).

The scripts that were used to create the data can be found in our [code repository](#).

Data and Resources







	latencies-pre-rate10000-snort-norules.pcap.zst Ingress for simple virtualized Linux forwarder at a data rate of 10 kpkt/s	Explore
	latencies-post-rate10000-snort-norules.pcap.zst Egress for simple virtualized Linux forwarder at a data rate of 10 kpkt/s	Explore
	latencies-pre-rate10000-dpdk-l2fwd.pcap.zst Ingress for non-virtualized DPDK forwarder at a data rate of 10 kpkt/s	Explore
	latencies-post-rate10000-dpdk-l2fwd.pcap.zst Egress for non-virtualized DPDK forwarder at a data rate of 10 kpkt/s	Explore
	latencies-pre-rate20000-dpdk-l2fwd.pcap.zst Ingress for non-virtualized DPDK forwarder at a data rate of 20 kpkt/s	Explore
	latencies-post-rate20000-dpdk-l2fwd.pcap.zst Egress for non-virtualized DPDK forwarder at a data rate of 20 kpkt/s	Explore

Figure 13. Dataset for 5G experiments

This dataset is composed of different resources, namely different files. By clicking on a file name, the user is automatically redirected to a Web page describing the individual file as shown below:



Home / Organizations / SLICES-SC / 5G QoS Measurements / latencies-pre-rate10000-snort-...

latencies-pre-rate10000-snort-norules.pcap.zst

[Go to resource](#)

URL: <https://github.com/gallenmu/low-latency/blob/master/measurements/motivation/pcap/latencies-pre-rate10000-snort-norules.pcap.zst>

Ingress for simple virtualized Linux forwarder at a data rate of 10 kpkt/s

There are no views created for this resource yet.

Resources

- latencies-pre-rate10000-s...
- latencies-post-rate10000-...
- latencies-pre-rate10000-d...
- latencies-post-rate10000-...
- latencies-pre-rate20000-d...
- latencies-post-rate20000-...
- latencies-pre-rate30000-d...
- latencies-post-rate30000-...
- latencies-pre-rate40000-d...
- latencies-post-rate40000-...
- latencies-pre-rate50000-d...
- latencies-post-rate50000-...
- latencies-pre-rate60000-d...

Additional Information

Field	Value
Last updated	January 4, 2023
Created	January 4, 2023
Format	unknown
License	Creative Commons Attribution
created	18 days ago
id	21df36a3-7e94-4626-bd42-b6fe85b70c8a
package id	fc1235b0-64ba-440a-a62d-219c838d5296
revision id	89dff62-b345-4e81-a0ee-5f76f9d338ce
state	active

[Hide](#)

Figure 14. Example of a file stored on GitHub

Below the file name, the URL permits to access the file directly, independently if the file is stored on the CKAN server or in another external location. In this example, the file is stored on GitHub and can be reached through the URL or alternatively by clicking on the button named "Go to resource". This solution avoids duplicating the storage of the resources in different locations and thus, reduces the necessary disk space on the CKAN server. It improves also the scalability in case of large datasets, in particular for live data with a high frequency. So, in the end, it is the best solution to store the references, so the URLs to the different resources, and the required metadata.

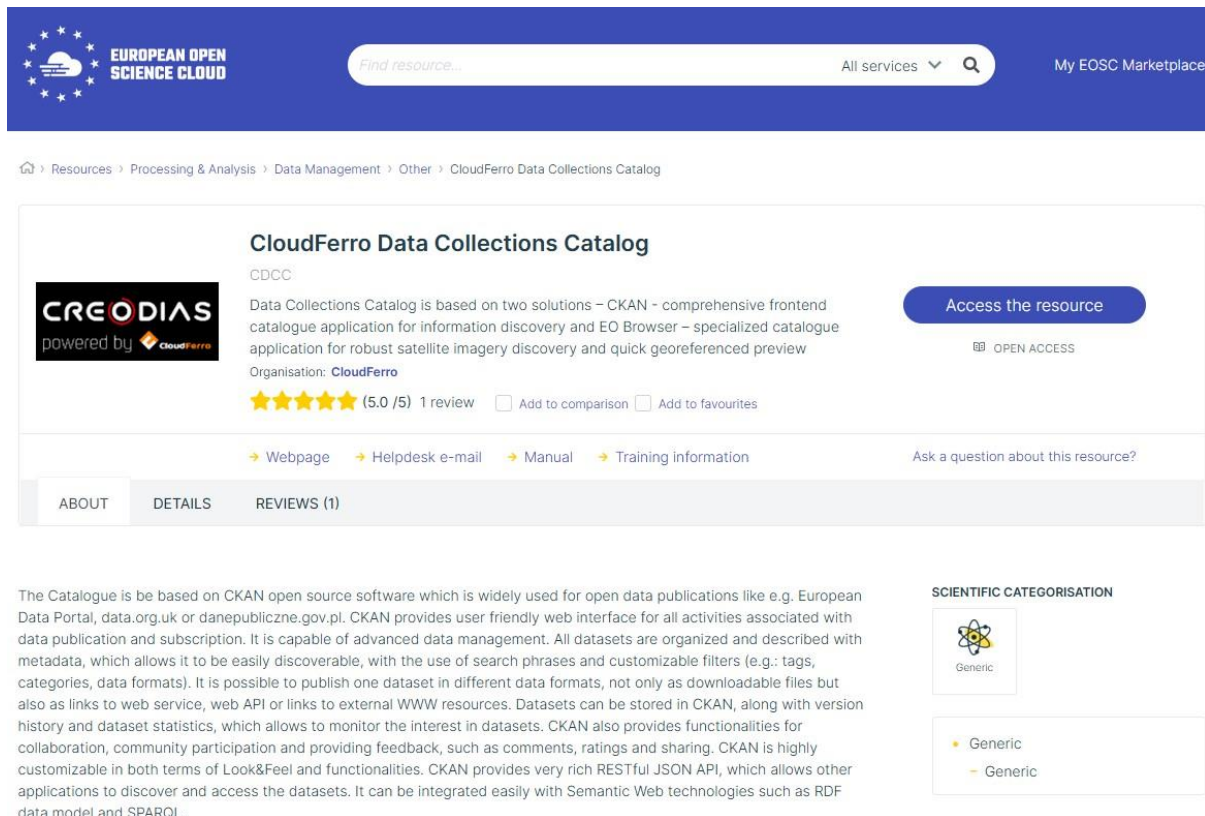
It is also possible to increase the scalability and reliability of the CKAN server by creating a CKAN cluster. Technically, it means that the different elements of the CKAN server are duplicated; these components are distributed in the front end and the back end. The CKAN front end corresponds to a Web Server Gateway Interface (WSGI) and can be run on multiple Web servers like Apache and NGINX. A load balancer is put in place in front of the Web servers (Apache or NGINX) through the Web servers' configuration files. The back end is composed of the PostgreSQL database and the Solr search platform; these components can be easily replicated via their respective configuration files. More information on how to build a CKAN cluster can be found on the CKAN Wiki:

<https://github.com/ckan/ckan/wiki/CKAN-High-Availability>. A CKAN cluster avoids a single point of failure.

In terms of security, CKAN offers some features that are summarised in the CKAN software official documentation: <https://ckan.org/features/security>. First of all, the technical team of CKAN is using different tools such as Dependabot (<https://docs.github.com/en/code-security/dependabot>) and CodeQL (<https://codeql.github.com/>) to automatise some development and maintenance tasks of the CKAN software and to detect potential security vulnerabilities in the source code. It is also possible to send discovered security vulnerabilities to the CKAN technical team through emails at security@ckan.org. These security vulnerabilities and other issues are discussed in dedicated developer meetings twice per week. Furthermore, a plugin is available at <https://github.com/data-govt-nz/ckanext-security> to improve the security of the CKAN software through different features: protection against brute force, two-factor authentication, blacklist of certain types of files, etc.

The CKAN open data server can be published in the EOSC catalogue as a service for data management. An example is CloudFerro Data Collections Catalog (<https://marketplace.eosc-portal.eu/services/cloudferro-data-collections-catalog>).

The following image displays this service published in the EOSC marketplace:



The screenshot shows the EOSC Marketplace interface. At the top, there is a search bar and navigation links. The main content area displays the service 'CloudFerro Data Collections Catalog' (CDCC). It includes a logo for 'CREODIAS powered by CloudFerro', a description of the service, and a rating of 5.0/5 with 1 review. There is a button to 'Access the resource' and a link to 'OPEN ACCESS'. Below the main content, there are tabs for 'ABOUT', 'DETAILS', and 'REVIEWS (1)'. A 'SCIENTIFIC CATEGORISATION' section is visible on the right, showing 'Generic' as a category.

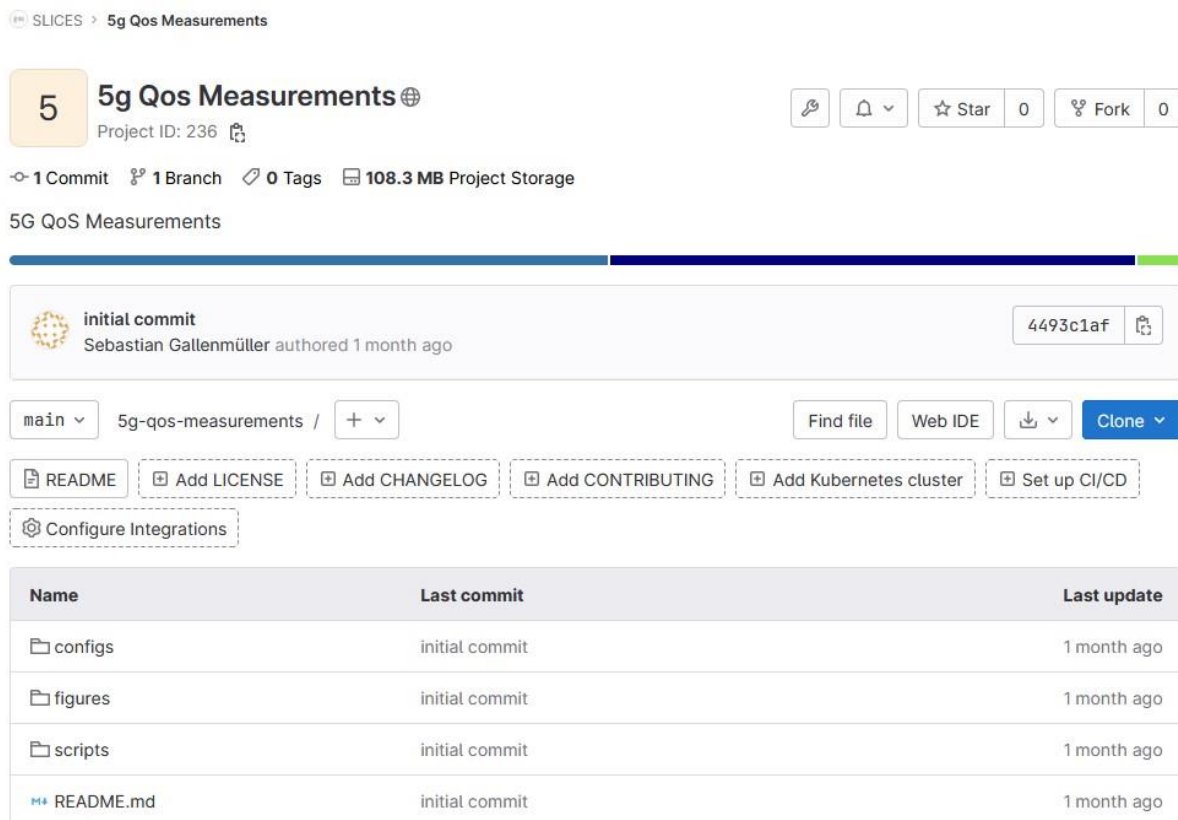
Figure 15. CKAN available through EOSC

8.6.2. GitLab

A source code repository was created on a GitLab server hosted and managed by the Swiss SLICES node. This repository permits storing the source code used by the researchers when building their experiments, the related technical documentation such as README.md files or Wikis and other useful files. The URL to the source code repository dedicated to SLICES on the GitLab installed in Europe is <https://gitlab.distantaccess.com/slices>.

GitLab, such as GitHub, uses the Git distributed version control system and permits storing and tracking any changes on the source code in the context of software development. In the SLICES-SC project, a dedicated group named “SLICES” was put in place.

The following picture shows an example of a project made for an experiment done in the SLICES research infrastructure:



The screenshot shows a GitLab repository page for a project named "5g QoS Measurements". The page includes a header with the project name, a "Project ID: 236", and statistics for commits (1), branches (1), tags (0), and project storage (108.3 MB). Below the header, there is a commit history table showing the "initial commit" by Sebastian Gallenmüller, authored 1 month ago, with a commit hash of 4493c1af. The table lists files in the repository: configs, figures, scripts, and README.md, all with their last commit and update dates.

Name	Last commit	Last update
configs	initial commit	1 month ago
figures	initial commit	1 month ago
scripts	initial commit	1 month ago
README.md	initial commit	1 month ago

Figure 16. Experiment done on SLICES-SC testbeds and published on GitLab

For instance, in this project stored on GitLab, there are several configuration files containing the network configuration of the virtual machines used during the experiments and the parameters used by the Snort software. The project includes scripts written in Python and bash scripts. A README.md file provides the basic information about the experiment. Publishing such projects on GitLab permits the reproducibility of experiments. Other features of GitLab can be also used to realise experiments in a quick way; for instance, CI/CD (continuous integration/continuous deployment) pipelines,



reporting of issues, Wikis are available to ease the life of the developers and researchers doing experiments with the SLICES research infrastructure.

8.6.3. Dataverse

Although CKAN is the recommended repository solution for storing data sets and related metadata in the SLICES-SC project, as described in section 11.2, we investigated other options with similar objectives having slightly different characteristics, advantages and disadvantages for the case new demands arise later with different requirements.

Dataverse is an open-source research data repository software (<https://dataverse.org/>) developed by the Institute for Quantitative Social Science (IQSS) at Harvard University, which is a web application aims to aid sharing, preserving, citing, exploring and analysing research data to support open science. Dataverses (similarly to directories) can contain further, nested dataverses as well as datasets (files, archives), each associated with descriptive metadata. It has more than 100 installations worldwide and widely used by academic institutions, research centres and government organisations.

Dataverse offers the following key features:

- **Data repository:** where data (raw or processed data, documents, any other supplementary material) can be uploaded to and shared with others
- **Metadata:** associated with datasets containing detailed information about data such as the creator, description, subject, etc.
- **Data citation:** each dataset uploaded gets a persistent identifier (typically, DOI)
- **Access control:** dataset owners control access to their data (but data can also be made public)
- **Versioning:** support to allow, record and track changes of datasets
- **Integration:** can be integrated with other platforms or tools
- **Compliance:** Dataverse helps in enforcing data management practices fulfilling institutional (or publisher) requirements or other standards

From the above list, we can see that Dataverse also serves properly **FAIR** principles and offers similar options as CKAN does. Technically, the **storage** solution backing up Dataverse where datasets are physically stored is configurable, and there are several options to choose from or integrate with, including local file system (servers), network attached storage (NAS), S3 (Amazon storage service), Swift (OpenStack's object storage), Azure blob storage, and even Globus.

Dataverse offers integration, and **federation** options at different levels to connect different Dataverse installations together (e.g., Harvard Dataverse Network provides federated access to datasets in different Dataverse deployments using OAI-PMH protocol). At the top level, metadata "harvesting" can collect and make metadata and datasets searchable, and findable from a single interface even if the data are physically located in different Dataverse repositories. In these cases, metadata format is often synchronised across the different Dataverse installations, using a common, uniform metadata schema (e.g., Dublin Core) to ensure better, seamless interoperability, and consistency (and to allow using semantic query languages such as RDF and SPARQL). Similarly, datasets can be made accessible through a single API, transparently where the particular dataset is actually hosted.

Dataverse provides strong support for using **custom metadata** fields and structures to adapt to domain-specific requirements, institutional practices, discipline-specific standards. In Dataverse, metadata are organized into blocks, where each block contains a set of metadata fields. Users can create new blocks and add new fields (or modify existing ones). Each field has a name, type (e.g., text,



data, selection from a vocabulary), labels, tooltips, properties (e.g., required) and validation rules. A block of metadata fields can be defined in a JSON configuration file, which can then simply be registered in the Dataverse installation and associated with a metadata template. Metadata can also be integrated with ontologies (enabling relationships between metadata fields and blocks). All these operations can be done in the GUI or via API.

Research object (RO) is a concept designed to package, describe, share and reuse (complex) research data. RO aims not merely to contain the raw data but also to describe the context and the processes involved in the given research to ensure interpretation and reproducibility.

The key components of ROs are as follows:

- **Data:** the research raw data obtained, generated, and observed.
- **Metadata:** detailed information describing the data (how it was collected, processed, in what environment)
- **Workflows:** methods and workflows used to generate, process or analyse the data, which can be descriptions, software or workflow descriptors.
- **Provenance:** information about the origin of the data or its history.
- **Supporting documents:** additional documentation related to the data (such as publications, presentations, and other outputs)
- **Links:** connection with other data resources (datasets, other ROs)

The aim of ROs is thus to help in sharing, comprehension, interpretation, reusing research data but also to ensure the reproducibility of experiments, data analytics processes and workflows.

RO-Crate (<https://www.researchobject.org/ro-crate/>) is a specific implementation of the RO concept. RO-Crate uses state-of-the-art technologies and standards to share research objects in an interoperable way. It uses JSON-LD (JavaScript Object Notation for Linked Data) to structure and describe metadata. It allows easy integration with other web technologies and systems. RO-Crate contains a specific directory (“crate”) for the primary data file and a special metadata file (“ro-crate-metadata.json”), which describes the contents of the package (data files, provenance, context, relationships). RO-Crate can be extended to accommodate other domain-specific requirements (additional metadata, data schemas, etc.).

RO-Crate thus aids reproducibility by packaging data, methods and metadata together into a single package. It enables reusing data and analytics workflows, methods and the interpretation of research data; and due to standard data formats, it allows better interoperability between different systems.

We **tested** the Dataverse API on a local Dataverse installation with a focus on the end-user functions (<https://guides.dataverse.org/en/latest/api/native-api.html>) to ensure that the same functionality is available **programmatically** via the GUI by human interactions. The following native API function-groups have been tested: Dataverse Collections, Datasets, Files, Users Token Management, Explicit Groups, Infos, Notifications. In addition, the Search API, the Data Access API and the Dataset Semantic Metadata API have been investigated too. (We omitted administration-related functionality requiring special permissions.) As a result, we found that most of the functions/API services worked correspondingly to the documentation, though there were some minor deviations from the expected outcomes, and we got imprecise (unclear) error messages in some cases.

In summary, both CKAN and Dataverse aim to allow data management and sharing having similar objectives compliant with FAIR principles, but CKAN is perhaps slightly more suitable for **open data** platforms due to its strong emphasis on metadata management, public accessibility, making datasets



easily discoverable, compliance with open-data standards. Dataverse can though be better tailored for specific research communities, domain-specific metadata, access control, data preservation and reproducibility in the long term.

8.6.4. EOSC Integration

EOSC permits publishing open data related to any kind of experiments. EOSC facilitates the reproducibility of experiments by publicly providing information on the datasets used during the experiments, the results from the experiments and all useful documentation on the realised experiments.

EOSC encompasses a specific category dedicated to data storage, including for archiving purposes.

In terms of data protection, the FAQ of the EOSC portal (<https://eosc-portal.eu/about/faqs>) mentions:

“Can I use EOSC Portal services to store and/or to process sensitive data (e.g. medical data)?

Generally, not. Most of the EOSC Portal storage services do not use encryption at rest, and do not possess specific certifications to handle sensitive data. However, there are few providers that are specifically designed for collecting, storing and processing sensitive data - for example ePouta and TSD.”

The EOSC integration was investigated in the context of the SLICES-DS project, in particular in the WP4 deliverables which describe the integration of the SLICES research infrastructure and related services into EOSC. The APIs (Application Programming Interfaces) offered by EOSC were examined, in particular the EOSC API dedicated to the open data providers. Indeed, it is possible to generate from the Swagger file published by EOSC the client in different programming languages (Java, Python). Examples of how to use this EOSC API were also provided through the deliverables written in the context of the SLICES-DS project; SLICES-SC is leveraging on the results of EOSC integration presented by the SLICES-DS partners.

8.7. Analysis for the open data storage

This section presents an analysis concerning the models used for the storage of open data. Indeed, a general model for open data sharing should be defined. Based on the current state of the art, an analysis was done, essentially by comparing the distributed model and the centralised model.

First of all, the centralised model for the open data storage means that a single location in the global network is used for the storage of all the open data. All the users must be able to access the open data storage; two important requirements appear, namely the scalability in terms of the number of users and a large bandwidth for the remote connections to the data storage. As located at a single point, it is easier to have a complete view of the saved open data. The management and maintenance are facilitated due to the single location of the data storage. Indeed, it requires only one server to be regularly updated and backed up. The main drawbacks are the single point of failure of this centralised model, the time efficiency and finally, the productivity in case of high traffic and insufficient resources available on the server dedicated to the open data storage.

In the distributed model for the open data storage, different locations on the global network are used to store the open data generated by the research infrastructure. This model permits the reduction of the resources necessary to properly run each local server involved in the open data storage. In the



same way, there are fewer interferences between users when they are accessing the same datasets. In case of a failure, it is possible to retrieve the open data in another location. The storage size stays the same as the centralised model per location, as all the open data are anyway replicated, but the distributed model adds redundancy. The maintenance must be done in each location where an open data server is installed; this means that more IT people are working on the tasks associated to the maintenance, the updates and the backups.

The comparison between the centralised model and the distributed model shows several advantages for the distributed model of open data storage. The scalability is better with the distributed model because the number of handled requests, including those coming through the open data server API, is higher. Indeed, fewer resources are needed per location/node to handle properly and timely the requests. The second point is linked to the security and the continuity of service: as there is no single point of failure in the distributed model, the availability of the open data is ensured. But the maintenance is more demanding in the distributed model of open data storage; this point requires more investments to cover the maintenance costs, in particular for the duplication of hardware and software and the necessary manpower. The user experience is also better in the distributed model as the concurrence between users is reduced by the multiplication of available servers dedicated to open data. A negative possible point is regulatory compliance when the open data servers are located in different countries, in particular, if one or several open data servers are located outside the European Union. In this case, particular care should be given to the transfer of data to third countries.

In the context of the SLICES Research Infrastructure, different locations dedicated to the open data storage are already available, essentially due to the existence of testbeds and related open data policies implemented in previous projects such as Fed4FIRE+. The testbeds are the primary source of open data. The results of experiments and the related documents can be stored on dedicated open data servers such as CKAN or Zenodo. For instance, the Fed4FIRE+ project has extensively used Zenodo to store the outcomes of the Fed4FIRE+ open calls. The European Open Science Cloud (EOSC) is now the recommended cloud to distribute the open data generated by the SLICES Research Infrastructure. The SLICES-DS deliverable D4.5 “SLICES infrastructure and services integration with EOSC, Open Science and FAIR: Recommendations and design patterns (final report)”⁴ mentions two other options: OneData and EGI DataHub. OneData⁵ allows the sharing of experiment data, simulations and publications at large scale. Indeed, the design of OneData was made for experiments requiring a large amount of datasets and high-performance computing (HPC). EGI DataHub⁶ is a service provided by the EGI (European Grid Infrastructure) community to publish datasets through a central portal. EGI DataHub is using OneData (<https://onedata.org/#/home>) as the underlying technology and is already available on the EOSC marketplace. Concerning regulatory compliance, no above-mentioned solutions are certified and so, it is clearly recommended to not store any personal or sensitive data in these open data repositories.

During the analysis, a comparison between the storage of full datasets and the storage of references and metadata was also undertaken. The results are the recommendation to store only the references accompanied by the metadata in the open data server if the data are already stored in other locations such as the testbeds, Zenodo, GitHub or GitLab. The SLICES-SC project recommends keeping the data

⁴ SLICES-DS deliverable D4.5, http://www.slices-ds.eu/wp-content/uploads/2022/12/SLICES-DS_D4.5_approval_disclaimer.pdf [Last accessed 17 July 2024]

⁵ OneData website, <https://onedata.org/#/home> [Last accessed 17 July 2024]

⁶ EGI DataHub website, <https://www.egi.eu/service/datahub/> [Last accessed 17 July 2024]



stored in the testbeds and inserting into the open data repositories only the references and the metadata of the datasets. It will reduce the size of necessary disk space on the open data storage instances.

In parallel with the work done in SLICES-SC concerning data management, SLICES-PP is studying and organising the concrete data management in the WP8: these activities encompass the organisational and technical measures linked to the data management, including the storage of the data and metadata in the nodes of the SLICES research infrastructure.

9. Intellectual Property Rights Management

SLICES prioritizes open access to its datasets, products, and/or solutions, as defined by the project, in accordance with the principles of Open Science and FAIR data. It aims to produce outcomes that are open, standard-based, and will, to the greatest extent possible, prevent vendor lock-in while also being in balance with the partners' intellectual property rights (IPR) generated within the project. IPR considerations were introduced in the Consortium Agreement and have been present throughout the exploitation discussions for this and associated projects, so as to maximize impact of the SLICES-RI and associated activities/solutions.

10. Allocation of Resources

With regards to the allocation of resources, there is no specific budget for making the data FAIR in the SLICES-SC project. The work done related to the data management is realised through the resources dedicated in the different Work Packages of the project. It is expected that a SLICES Data Manager will take care of the data management across the whole SLICES Research Infrastructure. It has been discussed and agreed that some components and services available in the SLICES-SC project, such as the CKAN open server and the GitLab source code repository, will be still online and used during the pre-operation of the SLICES Research Infrastructure. This project phase is managed in the SLICES-PP project.

11. Data Protection Office and Organisation

11.1. Coordination with SLICES-DS and SLICES-PP

As the SLICES infrastructure envisions a seamless transition among the different projects and into a fully operational research infrastructure, complementarity and coordination are important in all aspects of the SLICES ecosystem, including in regard to data protection. Throughout the project, consortium partners continuously expressed their dedication and willingness to foster a high level of compliance with relevant data protection regulations and to create an overall environment where the safeguarding of the rights of data subjects is a priority.

In this context, the various SLICES research projects are intrinsically connected and have contributed to the design and establishment of the SLICES Research Infrastructure in a seamless manner. Over the course of the SLICES research, there have been many learning opportunities along the way, which are being integrated into the infrastructure ultimately benefiting the end-users.



Most notably, SLICES-DS constitutes the design phase of all the envisioned SLICES activities, ensuring adequate planning of the activities and requirements. As such, SLICES-DS has incorporated the data protection security by design approach into the SLICES activities. In turn, SLICES-SC has focused primarily on the establishment of the research community that would benefit from the SLICES infrastructure and, thus, has further analysed data protection requirements from various stakeholders' perspectives including the end-users.

The knowledge acquired during these two initial research projects is being carried onwards to the SLICES-PP project, encompassing the preparatory phase that will validate the requirements to engage in the implementation phase of the RI lifecycle. Through SLICES-PP, the technical architecture will be finalised and the final policies and decision processes for the governance of SLICES-RI will be put in place, along with the business model, the required human resource capacities and training programmes.

An example of the knowledge generated and transferred is reflected in the establishment of the Data Protection Coordination Committee (DPCC), which fulfilled crucial tasks in regard to the coordination of data protection activities within the project, playing a crucial role, among others, in the communication between and coordination of partners' individual Data Protection and Legal Officers, ensuring alignment at a partner and project level. As a result, the need for a dedicated body, which is in charge of data protection and compliance in regard to the SLICES RI has been made even more evident.

In order to address the aforementioned need, it was decided that a permanent body would be established to take over the tasks of the DPCC and ensure partners' alignment within the SLICES infrastructure. This body is envisaged to centralise and enhance the monitoring of data protection activities throughout the SLICES infrastructure as a whole. Thus, the establishment of this body, as will be further analysed below, will contribute to the successful navigation of the project through the sphere of data protection, privacy, and data management.

11.2. Data Protection Coordination Committee

Given this complex environment, the Data Protection Coordination Committee's role was to ensure compliance with legal and ethical requirements, in order to ensure a holistic approach to compliance. This has been possible thanks to the contribution of partners' DPOs, Legal and Ethics Officers. The DPCC has been organised as follows:

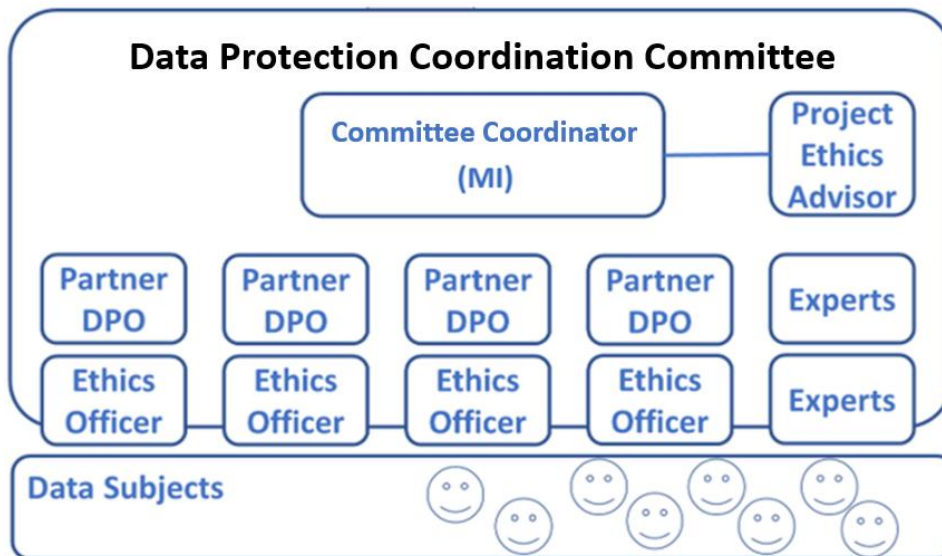


Figure 17. Data Protection and Coordination Committee

As illustrated in the figure above, Mandat International has been appointed as the project’s Data Protection Officer and, thus, the Coordinator for the Data Protection Coordination Committee. The partners’ DPOs and legal experts were also involved in this procedure in order to ensure an additional layer of accountability and legal/ethical knowledge-sharing, strengthening the project’s integrity overall.

The above-described structure of the Data Protection Coordination Committee guarantees a comprehensive oversight and coordination effort, promoting cooperation and a collaborative approach to legal and ethics compliance by exchanging best practices. However, this structure is also beneficial for end-users of the SLICES infrastructure.

It is formally recommended that a similar structure is followed in all future stages of the SLICES-RI and associated ERIC activities, as the establishment of independent oversight roles is increasingly requested in the context of recently approved regulations even beyond the GDPR (e.g. AI Act). Such an approach would enable to solidify the ethics, security and privacy-by design approach of SLICES while maximizing end-user trust in its solutions and stakeholders.

11.3. Data Protection Office

The central body taking over the DPCC tasks, as described above, was defined as the Data Protection Office with the aim to strengthen the coordination of data protection of the SLICES infrastructure and amplify the trust of data subjects. The website of the office, which assumes the role of the main point of contact regarding data protection for both partners and data subjects, can be accessed via this link: <https://slices-forum.eu/dpo/>. The following Figure presents its home page, providing an overview of the activities that can be performed through the website.



SLICES Data Protection Office



Figure 18. Home page of the Data Protection Office website

As also validated through its website, the Data Protection Office's main tasks can be divided into two points of focus, namely: The coordination of data protection actions within the project, including coordination with the partners' Data Protection Officers, and the provision of a central mechanism for data subjects to exercise their data protection rights.

In regard to the first category of activities, the Data Protection Office fulfils several tasks of increased significance. As such, it is the central point of contact for partners, in case they have to update any information in regard to personal data and in case their approach towards the collection of data changes. The latter includes, for example, if partners are collecting personal data of a different nature than originally foreseen in the initial data management questionnaires or if their method and conditions of data collection, storage or access change.

It is worth mentioning that the Data Protection Office has stored the original descriptions of partners' data-related activities and is in charge of updating them according to updated information received. In turn, partners have the obligation to inform without delay the Data Protection Office upon any changes in their data-related activities or persons in charge, so as to keep the information updated at all times.

These evolutions and updates can be easily and comprehensively reported by the partners via the online form for SLICES partners to notify changes to their data protection activities on the website of the SLICES Data Protection Office. Through the form, partners can indicate which kind of information update they would like to provide, as demonstrated in the figure below.



CONTACT PERSON

Name: *

Email address: *

Phone number: *

INFORMATION UPDATE

Please indicate the information you wish to update: *

- I wish to update the information of my organisation's Data Protection Officer (1).
- I wish to update the information on the categories of data (2) my organisation is collecting or processing in the context of SLICES.
- I wish to update the information on the methods and purposes (3) of data collection or processing of my organisation in the context of SLICES.
- I wish to update the information on the data storage and data retention (4) of my organisation in the context of SLICES.
- I wish to update the information on the Technical and Organisational Measures (5) my organisation is implementing in the context of SLICES.
- I wish to update the information on the data-sharing activities (6) my organisation is performing in the context of SLICES.
- Other

Next

Figure 19. Online Form to update SLICES Partners' Data Protection Activities

Assigning the coordination and collection of such changes and updates to a central body not only ensures that the partners are continuously complying with their obligations but it also provides an additional oversight and accountability mechanism.

Upon receipt of partners' communication through the form, the Data Protection Office updates the record of partners' data-related activities with the new information. Where required, the Data Protection Office may contact the partner in question to request clarifications or raise potential issues with them so as to address them as soon as possible.

Additionally, the Data Protection Office forms the primary point of contact for data subjects to exercise their data subject rights and, thus, acts as a bridge between SLICES partners and data subjects. Contact with the Data Protection Office for this purpose can easily be established through the relevant online form to exercise data subjects' rights on the official website of the Data Protection Office. As such, data subjects can indicate which kind of right they wish to exercise with regards to their data and the details of the relevant request, as demonstrated below.



Contact Information:

Name *

Email * Phone number

Affiliation (if applicable):

Address: *

Dropdown

REQUEST/Exercise of Data Subject rights:

Please indicate which type of personal data is involved in your request: *

personal data (any information relating to identified or identifiable individuals, including for instance email or IP addresses)

special categories of data (personal data revealing sensitive information such as sexual orientation, racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, as well as any health, genetic or biometric data related to the data subjects)

How was the data that is the cause for your complaint obtained from you? *

directly from me, with my consent

directly from me, with another legal basis

not directly from me, without my knowledge

not directly from me, with my knowledge

Please indicate the general nature of your request and the rights that you wish to exercise:

I wish to be informed about the processing activities performed using my personal data (Right to be informed)

I wish to access to my personal data (Right to access)

My personal data is inaccurate (Right to rectification)

I wish to restrict the processing of my personal data (Right to restrict processing)

I wish to transfer my data to a different data controller (Right to data portability)

I wish to object to the processing of my personal data (Right to object)

I do not want to be subjected to automated decision-making or profiling (Rights related to automated decision-making and profiling)

I want to withdraw the consent that I have previously given to the processing of my personal data (Right to withdraw consent)

Other

Please provide any further information related to your request below.

Submit

Figure 20. Online Form for Data Subjects

Upon receiving a data subject’s request to exercise one of their rights, the Data Protection Office will forward the request to the partner concerned so as to ensure a timely response within a month since the receipt, unless there is a valid and documented justification for any delay. The Data Protection Office maintains a record of received data subjects’ communication, including information on the date





and time received, the partner or partners involved, as well as with the follow-up actions and justifications for any inaction, where applicable.

The SLICES Data Protection Office shall act as a point of contact for Supervisory Authorities. Where requested, it shall provide the competent Supervisory Authorities with all requested information about the data collected and processed through the SLICES infrastructure, tools, and solutions. Similarly, where applicable, it shall duly notify the competent Supervisory Authorities in case of a data breach, as well as the mitigation measures adopted.

Through the SLICES Data Protection Office website, it is also possible to find more information on the structure, organisation and governing rules of the Office, as evident below.

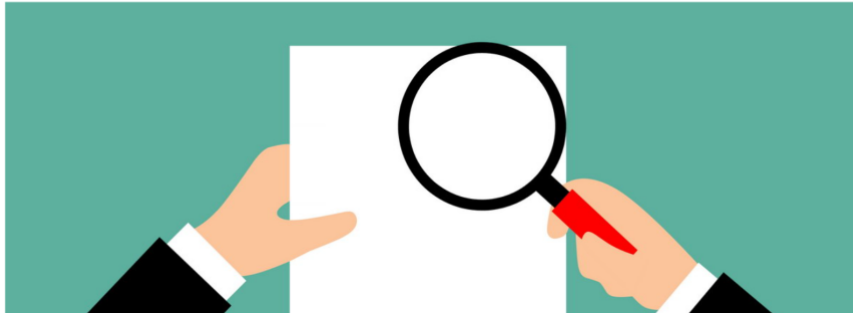


Figure 21. Data Protection Office Rules and Operation

Finally, the Data Protection Office is meant to publish reports, when applicable, in order to inform the public of its activities, as well as the project's data-related actions, policies and updates. Said reports will be made available through the Data Protection Office website, as follows.



SLICES Data Protection Office Reports



This page contains the reports published by the Data Protection Office. The reports may provide additional information and updates on the SLICES data-related activities, its policies and additional measures implemented.

Figure 22. SLICES Data Protection Office Reports

Said information provides an additional layer of transparency regarding the SLICES data management and protection activities, while contributing to the publicly available information that will be further analysed in the following section.

11.4. Public Information for Data Management and Protection

11.4.1. SLICES-SC website

In addition to the above-described activities and mechanisms, and in order for a large-scale research project like SLICES to ensure an effective process of data management and an adequate level of data protection, publicly available resources are of grave importance. Through such publicly available information, transparency, integrity and trust are enhanced within the whole of the SLICES infrastructure. Public information is especially important with regards to end-users and individuals. Those parties may not even be aware of their rights in some instances, so it is a necessity for SLICES to inform individuals about such relevant circumstances (Articles 12 – 14 lay down the information rights in the GDPR).

Such a beneficial public platform can be achieved through a comprehensive, explanatory and approachable website. The SLICES website contains much relevant information in this regard, such as, general information about the mission and objectives of the project, what kind of personal data is collected and why, and more relevant information about the project as a whole.

Being able to access information about the project's data management publicly online helps to instill trust within individuals and data subjects and, as such, positively contributes to the overall success of the SLICES infrastructure.

11.4.2. SLICES Portal

Another way to promote publicly available information and, thus, enhance trust, integrity and accountability in the SLICES project is the SLICES portal.

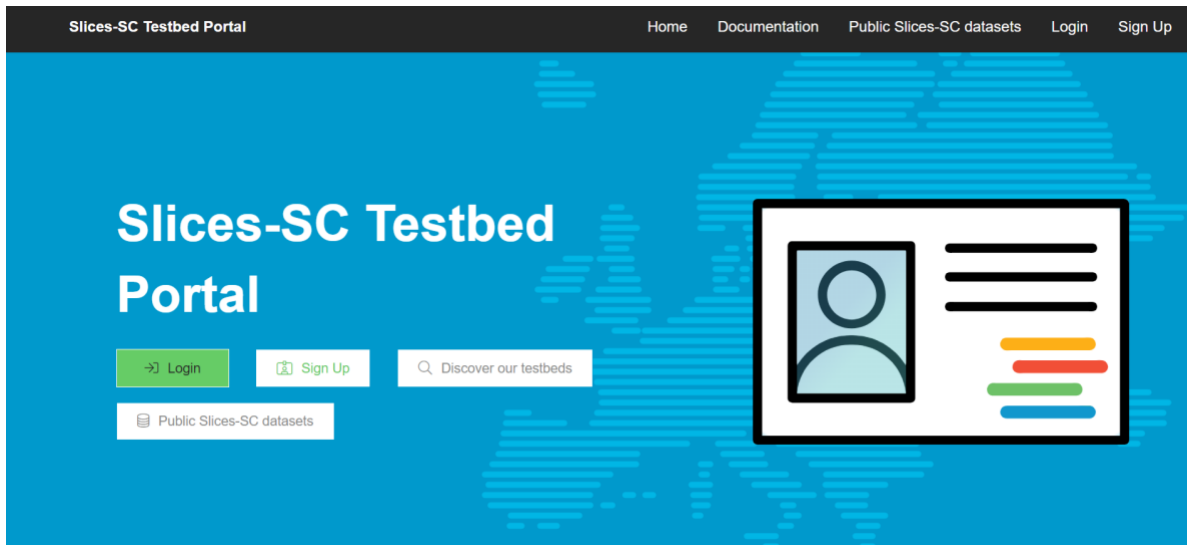


Figure 23. SLICES Portal Home-page

The SLICES portal permits an overview of the testbeds developed throughout the project, thus it also includes a 'discover our testbeds' option. Through this option, publicly available SLICES testbeds are visible. This permits individuals an insightful view of the activities conducted throughout SLICES. Hence, the trust of individuals can be strengthened by being informed of and reviewing the actual developments that SLICES produces, while the portal also serves as an opportunity to awaken the interest of like-minded people, for example, researchers, thus broadening the outreach of SLICES.

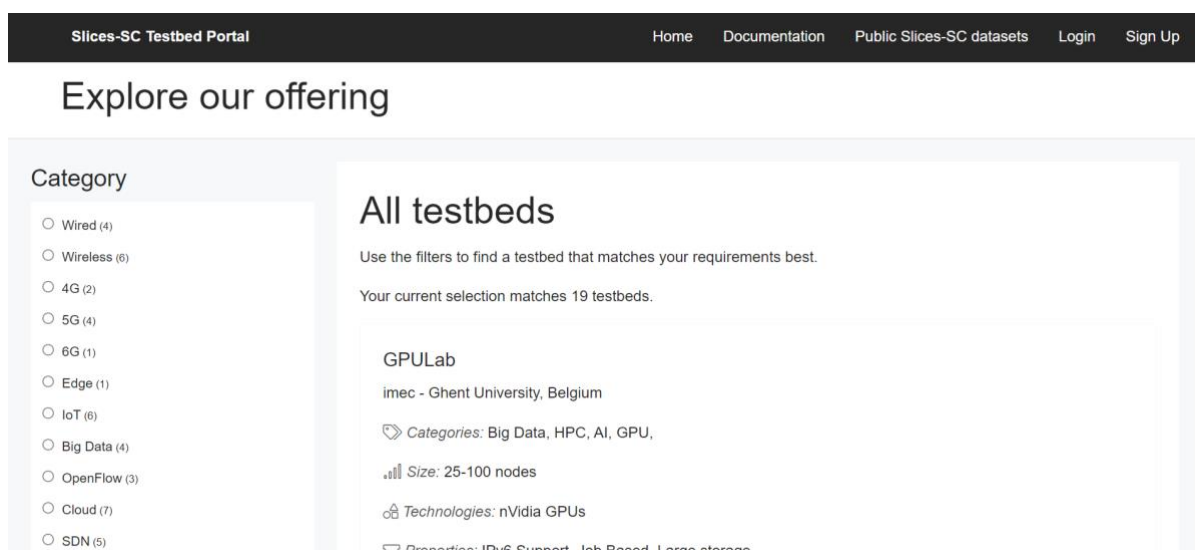


Figure 24. A look at the 'discover our testbeds' site on the SLICES portal



12. Data Protection Monitoring

12.1. Data Protection Coordination

The coordination of data protection activities is essential for the success of SLICES. As explained in the first version of the SLICES DMP, partners have been focusing since the beginning on ensuring that all processes and systems deployed in the context of SLICES are sufficiently flexible to be updated and optimised in the function of new requirements defined by the researchers. Hence, this structure enables continuous monitoring of the SLICES RI at all points of the project. The coordination of SLICES data protection monitoring is, as such, embedded in the governance structure of the project.

An example of the coordination of data protection within the project is the Data Protection and Coordination Committee. As described above, this body is essential for the SLICES infrastructure and it highlights the efforts of coordination among consortium partners that has been undertaken throughout the project. This Committee has assisted significantly with the coordination of DPOs, thus, ensuring a unified approach is taken among the partners. The Data Protection Office, assuming this coordination role, is another instance of coordination in the field of data protection within SLICES, attesting to the partners' commitment to ensure compliance in a coordinated and homogenous manner.

Successful coordination in regard to data protection has many advantages not only for the project but also the SLICES infrastructure as a whole. One of those benefits is a more efficient risk management, since coordination of best practices could lead to less data and security incidents. Furthermore, a data environment that is reliable can help to foster innovation and competition, as researchers themselves are more inclined to trust the SLICES ecosystem. Along the same lines, the data quality and accuracy are enhanced through successful coordination of data protection activities, which raises the level of integrity of the project as a whole. The latter also positively impacts participants of the research project, who are more likely to trust the environment of the project and are, thus, also more likely to participate in said project, furthering the process of the research and innovation. Lastly, the coordination of data protection activities can also have a societal impact, meaning that, in the long term, society will be reinforced to trust in scientific research projects.

12.2. Data Protection Workshop

As a means to further promote data protection cooperation and alignment of the activities, the project's DPO organised a workshop on "Making Research GDPR Compliant". During this workshop, SLICES partners were able to delve deeper into the primordial GDPR Obligations that are applicable when performing research involving personal data, including particularly the following:

- Principles for data collection and processing (Article 5)
- Obligation to inform the data subject about the processing of his/her data (Article 12)
- Possibility of further processing (Recital 50)
- Data protection by design (Article 25)
- Security of processing (Article 32)
- Designation of a DPO (Article 37)
- Conducting a DPIA (Article 35)
- Safeguards and derogations for research (Article 89)



Based on the above, the partners further discussed their roles and obligations under the GDPR, as well as the importance of a solid data protection coordination and monitoring, as detailed in this deliverable. To that end, the functions of the DPCC and Data Protection Office were also clarified.

Additionally, the most relevant legislation in the field of data and research infrastructures in the European Union was further reviewed to ensure a holistic compliance approach, striking a balance between innovation and data protection. Thus, the partners agreed upon the distributed data protection strategy, divided into two layers:

1. Partner layer, remaining responsible for all of their respective organisation's data protection compliance activities, including the performance of a DPIA;
2. Project layer, ensuring coordination and providing guidance regarding further compliance activities, as well as being the point of contact for all of the project's data-related activities.

As the workshop's main outcome, partners better comprehended their respective obligations and the overall SLICES approach to compliance, while exploring additional tools (discussed below) that could assist in these proceedings. Finally, a number of recommendations were provided for researchers' compliance activities, including the following focal points:

- Abide by the principles for data collection and processing in the GDPR.
- Get accustomed with and respect the rights of data subjects.
- Ensure data protection by design.
- Ensure security of processing and storage.
- Designate a DPO and conduct a DPIA.
- Ensure a strong data protection organisational structure.
- Monitor compliance on a continuous basis, using appropriate tools as aid.
- Keep an eye on regulatory updates and upcoming legislation.

12.3. Data Protection Monitoring

As mentioned in previous sections, the Data Protection Office has, as one of its goals, the permanent monitoring of compliance with the complex regulatory landscape. This means that constant attention to data protection principles and legal requirements is necessary in all processing activities within a research project. In order to aid in fulfilling this goal and to ensure an easier and more transparent monitoring process, a survey was created and distributed among partners (see Annex C). This survey looks into reviewing the current level of compliance of SLICES-SC with regulations and addressing potential risks for data subjects. The office has also contributed directly to the SLICES D7.2 mapping of regulatory compliance issues.

In this context, data protection policies, established procedures and mechanisms, as well as the technical and organisational measures must be duly monitored. This approach also seeks to foster the idea of leading by example and creating best practices, which can be replicated in future projects.

12.4. Data Protection and Compliance Tools for SLICES

In order to best align the data protection and overall compliance activities within SLICES, a number of tools has been explored and tested. During this process, the following tools have been identified, which could be used in order to ensure, monitor or promote data protection compliance within the SLICES-RI:



- a) **DP-ID** - It is a global registry of public information on data processing activities, as a way to enable compliance with the requirement of transparency set out by the GDPR. It eases the process for companies, public administrations and research infrastructures to enhance transparency and comply with some of their legal obligations by creating a unique identifier and a public record for their data processing activities (Data Processing ID or DP-ID). It enables to:
- Inform data subjects on the processing of their personal data and their rights and, in turn, comply with the legal obligation to inform data subjects (Art. 13 and 14 GDPR);
 - Facilitate the management of data processing internally and with third parties;
 - Map and monitor interdependent data processing involving multiple companies;
 - Use QR Codes, hyperlinks (URL), and widgets to share data processing required information.
- b) **Privacy App** - It is a mobile app enabling users and cities to:
- Monitor Internet of Things deployments and data protection together
 - Identify and share information on all communication devices deployed in public space.
 - Identify and report any new deployment that they identify in their environment.
 - It enables cities and public administrations to inform their inhabitants and visitors about the solutions deployed in their public space.

The application is designed to support the implementation of the GDPR in smart cities, but could be used also for the deployment of devices in the context of a research infrastructure.

- c) **Privacy Pact** - It is a tool for companies, aiding them in expressing their commitment to respect and to abide by the GDPR rules and principles, regardless of their location. It is contractually binding and Blockchain authenticated. It is not a certification scheme per se; however, it can be used to prepare a certification process such as Europrivacy.

13. Recommendations

The recommendations concern the open research data management and the reproducibility of experiments in the context of the SLICES Research Infrastructure, based on the different trials and experimentations done during the whole duration of the SLICES-SC project.

First of all, the CKAN open data server has demonstrated the need of SFDO (SLICES Fair Digital Object) which is a metadata model enforcing the FAIR principles across all the SLICES nodes and sites. Furthermore, it is also recommended to study in the context of the SLICES-PP the gaps between what the SLICES-SC project has produced in terms of tools and services dedicated to the open research data management and the reproducibility of experiments and the developments planned in the preparatory phase of SLICES (SLICES-PP). Indeed, some interesting progresses were realised in the SLICES-PP project on the future Metadata Registry System (MRS) and Data Management Infrastructure (DMI). Both of them can benefit of the experience gained in SLICES-SC on the real open research data generated and used by the researchers, notably in the different open calls.



The elaboration of the data management framework for the SLICES Research Infrastructure is still in construction, but it should be reinforced by a strong data management policy currently developed in the SLICES-PP project. This data management policy should also ensure that the privacy, the licences and the copyrights are fully followed by all the nodes and sites of the SLICES Research Infrastructure.

All the technical and organisational work related to the open research data management and the reproducibility of experiments should also ease the integration to EOSC, which is an important objective as mentioned previously in this deliverable.

Finally, the utilisation of Artificial Intelligence (AI) to better manage the open research data, in particular the metadata, and the experiment reproducibility should be studied in more details. Indeed, automation of some task related to the data management and the reproducibility could be potentially given to future components and services using AI.



14. Conclusion

This deliverable provides the final perspectives on SLICES-SC's WP3 work on Data Management and replicability, while reporting on the work performed to establish the Data Protection Office as an independent oversight body for regulatory compliance for the SLICES-RI.

As denoted in the previous sections, the work performed by these tasks has been extensive, and required direct coordination with relevant SLICES-SC stakeholders to ensure the project meets FAIR principles while maximizing end-user trust in the solutions and infrastructure being developed. In particular, it is recommended that the activities performed towards the establishment of the Data Protection Office are further continued in the future through its integration in the governance structure of the SLICES ERIC. This would enable a continued coordination across the projects, a well-defined data management and data governance process, and particularly, a scalable regulatory compliance approach ensured by an independent entity in line with current and upcoming regulatory frameworks.



Annex A: Data Management Processing Form

This annex presents a data management processing form which should be completed by all the data providers, typically the researchers. It will permit to submit electronically data in compliance with the data management plan. This form should be accompanied by the completed and signed data processing agreement which is presented in the Annex B of this deliverable.

Identification/Instantiation	
Internal ID	Generated by the resource manager, i.e., DOI
External IDs	Other identifiers for the resource (e.g., links, bibliographic citations)
Privacy/Access level	Open <input type="radio"/>
	SLICES Node level <input type="radio"/> (select node from list)
	Shared <input type="radio"/> (provide a list of one or more organizations from a list)
	Private <input type="radio"/>
Version	

Content	
Creator	Organization/Person Name
Creator ID	Identifier used to recognize creator, e.g., ORCID, DAI, LinkedIn
Title	
Alternative Title	
Description	
Subject	
Keyword(s)	Multiple-selection from a list (e.g., frequent keywords) or free text
Language (s)	Multiple-selection from a list
Duration (if applicable)	Selection from date/time pickers
Location(s) (if applicable)	(Ideally) multiple-selection from hierarchical location lists
Funder(s) (if applicable)	Multiple-selection from a list (e.g., known funding authorities) or free text



Publishers(s) (if applicable)	
-------------------------------	--

Date	
Create date	Automatically generated from the system
Date Submitted	Date of submission of the resource
Date Issued	Date of formal issuance of the resource
Date Accepted	Date of acceptance of the resource
Date Copyrighted	Date of copyright of the resource
Date Modified	Date on which the resource was changed
Availability of the resource	Minimum date that the resource should become available
Expiration of the resource	Maximum date that the resource should be available

Relationships (for each relationship)	
Relation Type	Selected from a list
Reference to resource	e.g., DOI
Description	e.g., uses external dataset for specific purposes

Rights Management	
Provide any right(s) that are related to the resource:	e.g., link to terms (in list format)
Provide any License(s) that are related to the resource	specific licenses that apply to the resource

Resource Characteristics (to be completed for each resource)		
What is the resource type?	Collection/Project	<input type="radio"/>
	Single resource	<input type="radio"/>
	Part of a collection	<input type="radio"/> If yes, select collection/project from a list





What is the measurement type (if any)?	Observational	<input type="checkbox"/>
	Experimental	<input type="checkbox"/>
	Simulation	<input type="checkbox"/>
	Derived	<input type="checkbox"/>
	Other: (please specify)	<input type="checkbox"/>
What is the format of the resource	Open file format	<input type="radio"/>
	Proprietary file format	<input type="radio"/> If yes, provide additional information below
	Proprietary file format details	e.g., link to required software to access the resource
Specify the size of your resource	Automatically assessed by the system. Large files may require different upload processing.	
Specify any special requirements for the resource	Computationally intensive	<input type="radio"/> Yes <input type="radio"/> No e.g., if yes, provide requirements
	Storage intensive	<input type="radio"/> Yes <input type="radio"/> No e.g., if yes, provide requirements
	Network intensive	<input type="radio"/> Yes <input type="radio"/> No e.g., if yes, provide requirements
Provide any other characteristics for the resource	(key, value) pairs, where key is selected from a list, e.g., (Software code, python), (Tabular data, csv)	

Compliance/Data Quality		
Does the resource contain:	Personal data	<input type="checkbox"/>
	Sensitive data	<input type="checkbox"/>
	Data subject to license	<input type="checkbox"/>
	Derived	<input type="checkbox"/>
	Do you verify the completeness of the data?	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A



Provide appropriate consents for the resource:	Do you verify the timeliness of the data?	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	
	Have you obtained appropriate consents for the use/processing of personal data contained in the resource?	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A If yes, provide description and/or link to resource	
	If you are using external resources, have you obtained appropriate licenses/rights to use them?	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A If yes, provide description and/or link to licenses/rights	
	Additional Consents (please specify)	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	
Data Quality Assurance	Do you verify the quality of the data during data collection?	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A Select yes if the instruments used for data collection provide quality assurances	
	Are the data provided in raw format (i.e., no pre-processing has been performed on the data)?	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A Select yes if the instruments used for data collection provide quality assurances	
	Have you used the recommended directory structure?	<input type="radio"/> Yes <input type="radio"/> No If no has been selected, please describe the structure	
	Have you used the recommended naming conventions?	<input type="radio"/> Yes <input type="radio"/> No If no has been selected, please describe the naming convention	
	How will the data be documented?	(key, value) pairs, where key is selected from a list, e.g., (configuration, link to read.me), (jupyter notebook, link to notebook)	
	How will versioning be managed?	No versioning, new resource will overwrite the previous	<input type="checkbox"/>
Automatic numbering/Date-		<input type="checkbox"/>	



		Time/Version number in the structure (directory/filename)	
		“Track changes” feature in software	<input type="checkbox"/> Specify software and method
		Dedicated version control software:	<input type="checkbox"/> Specify software and method
		Other	<input type="checkbox"/> Specify method
	Data Security	Have any measures been taken to secure the data	(key, value) pairs, where key is selected from a list, e.g., (anonymization, details), (encryption, technique)

Data Security (to be completed for each resource)

What is the nature of any security requirements?	Brief summary of requirements, or a link to where they are specified.	
Have any measures been taken to ensure security?	List potential risks	
What is the measurement type (if any)?	Observational	<input type="checkbox"/>
	Experimental	<input type="checkbox"/>
	Simulation	<input type="checkbox"/>



	Derived	<input type="checkbox"/>
	Other: (please specify)	<input type="checkbox"/>
What is the format of the resource	Open file format	<input type="radio"/>
	Proprietary file format	<input type="radio"/> If yes, provide additional information below
	Proprietary file format details	e.g., link to required software to access the resource
Specify the size of your resource	Automatically assessed by the system. Large files may require different upload processing.	
Specify any special requirements for the resource	Computationally intensive	<input type="radio"/> Yes <input type="radio"/> No e.g., if yes, provide requirements
	Storage intensive	<input type="radio"/> Yes <input type="radio"/> No e.g., if yes, provide requirements
	Network intensive	<input type="radio"/> Yes <input type="radio"/> No e.g., if yes, provide requirements
Provide any other characteristics for the resource	(key, value) pairs, where key is selected from a list, e.g., (Software code, python), (Tabular data, csv)	



Annex B: Data Processing Agreement

This annex presents the data processing agreement to be signed by the data provider or subject before any data processing by a legal entity of the project.

Agreement to process data in the SLICES-SC project

Name:

Home organisation:

Email address:

Phone number:

SLICES-SC is an EU-funded project aspiring to foster the community of researchers around the SLICES Research Infrastructure (SLICES-RI), create and strengthen necessary links with relevant industrial stakeholders for the exploitation of the infrastructure, advance existing methods for research reproducibility and experiment repeatability, and design and deploy the necessary solutions for providing SLICES-RI with an easy to access scheme for users from different disciplines. A set of detailed research activities has been designed to materialize these efforts in tools for providing transnational (remote and physical) access to the facility, as well as virtual access to the data produced over the facilities. The respective networking activities of the project aspire in fostering the community around these infrastructures, as well as open up to new disciplines and industrial stakeholders. For further information: <https://slices-sc.eu/>

Date:

Data to be processed by the SLICES-SC project: internal ID and eventually external IDs from the data management processing form

I agree / I don't agree to the process of the above-mentioned data in the SLICES-SC project, following the rules and guidelines described in the Data Management Plan and the related policies.



Annex C: Data Protection Coordination and Monitoring Survey

This annex presents the template of the survey sent to the partners, in accordance with the objectives of T 3.3 and in order to simplify data protection monitoring.

HORIZON 2020

H2020 - INFRAIA-2020-1

Survey SLICES-SC Data

Protection Coordination and Monitoring – WP3

Date: XXXXXXXX

Partner: XXXXXXXXX

GENERAL RATIONALE AND INSTRUCTIONS

This final survey aims at collecting final information on data management, ethics and data processing activities of partners in the present research project. Completion of this report is obligatory as per the Grant Agreement dispositions.

All questions shall be understood as referring to your data management and processing activities in the context of the current research project.

You are kindly requested to answer all questions. Should you not process any personal data in the framework of your involvement in the project, you **should at least complete the sections up to (and including) FAIR data management**.

Please return the completed survey to ccretta@mandint.org and aguesada@mandint.org **no later than 21 June 2024**.

Partner Organisation:

Name:

Address:

Country:

Website:

Privacy policy webpage:

Organizational Contact Person

Name:

Email address:

Phone number:

Organizational Data Protection Officer

Name:



Email address:
Phone number:

Data Processing Activities

1. Indicate what categories of data you collect d or process d in the context of the project:

- non personal data** (i.e. environmental data).
- personal data** (any information relating to identified or identifiable individuals, including for instance email or IP addresses)
- special categories of data** (personal data revealing sensitive information such as sexual orientation, racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, as well as any health, genetic or biometric data related to the data subjects)

2. Describe the categories of personal data you will be collecting and/or processing:

How did you collect these personal data?

- directly from data subjects who belong to your research team**
- directly from data subjects outside your research team (i.e. early adopters, beta testers, etc.) => Researchers working in open calls.**
- indirectly through partners of the project**
- indirectly through other organizations external to the project**
- N/A (you can skip the last section of this document)**

Data Management:

- 3. Please list all datasets which your organization currently uses or has obtained in the context of the SLICES SC project (Feel free to duplicate this table if multiple datasets will have been used):**

Please provide your answers in this column:

Name of the used dataset(s)	
Short description of the dataset(s)	
If the dataset includes personal data, please specify the type of personal data.	
Purpose for which you use/ process the dataset(s)	





Format(s) of dataset(s)	
Where will you store the dataset(s)?	
What is the main source of the dataset(s)?	
Who owns the dataset(s)?	
Origin of the dataset	
Are there any restrictions for the use of the datasets?	
Who has access to the datasets?	
How long will you keep the datasets?	
Under which licence did you obtain access to the datasets?	
Additional comments	

FAIR data:

4. Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, **kindly provide additional information on how each of these principles are being met:**
 - a) Findable
 - b) Accessible
 - c) Interoperable





d) Reusable

Intellectual Property Rights:

- 5. Did your organization generate any foreground IPR as part of the project? If so, please describe its type (patents, copyrights, trademarks, know-how, trade secrets, etc) and provide a brief description.**

Ethics and Personal Data Protection:

- 6. For what purpose(s) did you collect the aforementioned personal data?**
- 7. Did you process the generated data for any further purposes than the ones it was originally collected for? Yes No**
- 8. If you answered yes to the previous question, then please describe the purpose of this additional processing:**
- 9. How did you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?**
- 10. How did you plan to collect and document the consent of the data subjects whose personal data will be processed by you?**
- 11. How and where did you store the data?**
- 12. For how long did you keep the data?**
- 13. What technical and organizational measures (TOMs) are in place to protect and secure the personal data?**
- 14. Describe the measures in place to anonymize and/or pseudonymise the personal data whenever possible?**
- 15. Did you (plan to):**
 - share personal data with other partners within the project? Only to the data subjects themselves.**

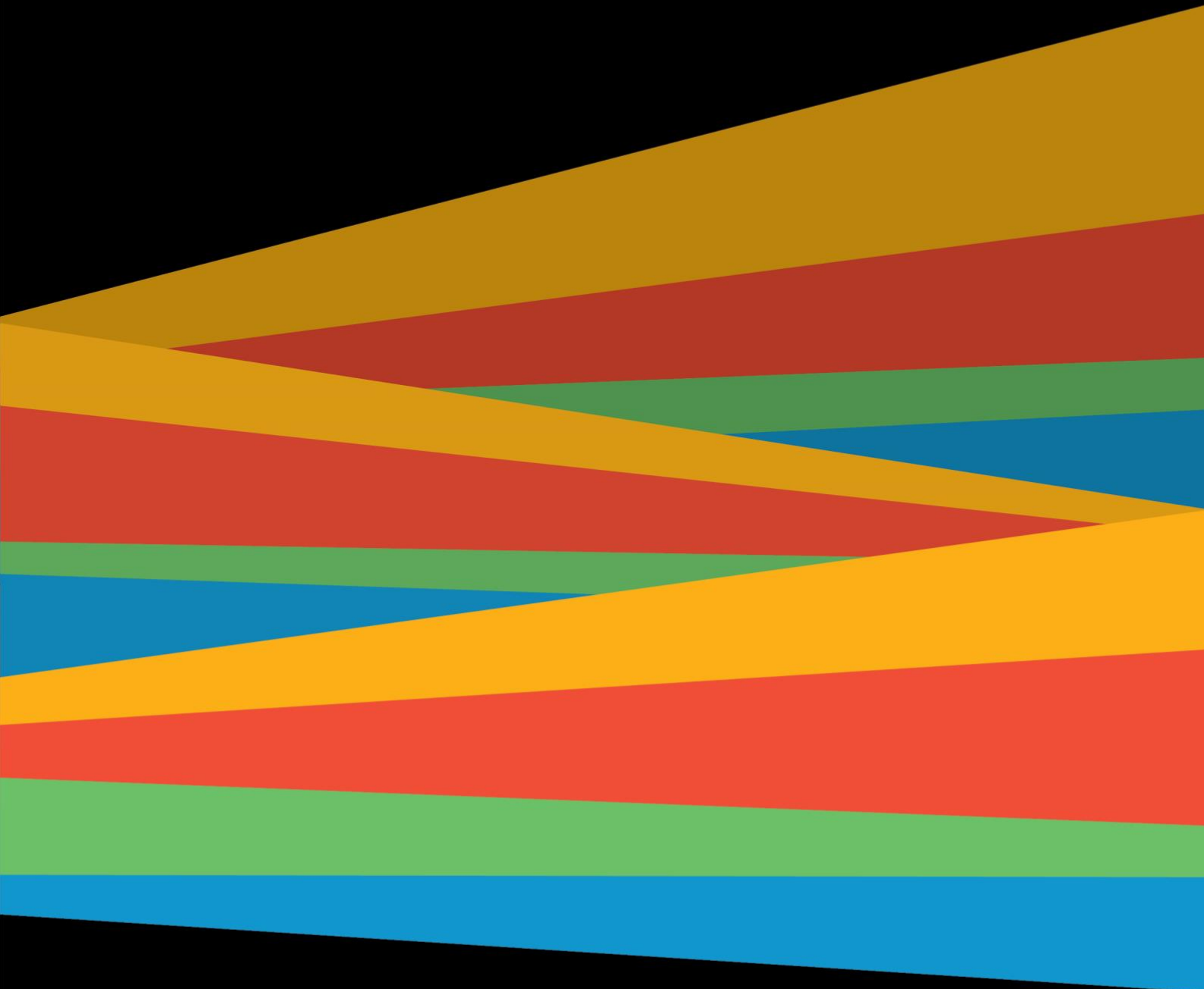


- share personal data with data processors** (third parties who will process data under your control)
- transfer personal data to countries outside of Europe**
- make personal data available to third parties for further research or processing**
- perform a data protection impact assessment?**

16. What risk do you foresee for the individuals (data subjects) whose data will be processed by future instances of the SLICES project in the future?

17. What would be your suggestions to minimise the risks for the data subjects?

18. Any comment or suggestion?



slicessc