HORIZON 2020
H2020 – INFRADEV-2019-3

slices DS

D4.1

Data Management Plan

Project acronym:	SLICES-DS
Project full title:	Scientific Large-scale Infrastructure for Computing / Communication Experimental Studies – Design Study
Grand Agreement:	951850
Project Duration:	24 months (Sept. 2020 – Aug 2022)
Due Date:	28 February 2021 (M6)
Submission Date:	16 March 2021 (M7)
Dissemination Level:	Public
Authors:	UCLan Cyprus, SU, INRIA, UTH, MI, UC3M, UvA, eBOS
Reviewers:	ALL

Co-funded by the Horizon 2020 programme of the European Union

Executive summary

SLICES aims to design and implement a Europe-wide test-platform, to support large-scale, experimental research that will provide advanced compute, storage and network components. We expect that a large number of researchers will take advantage of SLICES to generate data of various sorts (observational, experimental, simulation) and produce research results. Additionally, researchers will have the opportunity to collaborate with other researchers by utilising data and services of other international testbeds that will be seamlessly supported through SLICES. Understanding the data collected and the processes that manipulate them is essential for the development of the necessary policies and procedures for regulating the management and publication of research data within SLICES.

This deliverable analyses requirements from multiple areas of the project, such as the technical and operational requirements drawn from (i) the scientific community, (ii) objectives and constraints of the overall reference architecture, (iii) the governance, management and human resources, and (iv) the legal and ethical policies and regulations, to produce a comprehensive Data Management Plan (DMP) for the future SLICES research infrastructure (SLICES-RI). The DMP provides policies and protocols on how data is governed and managed and how it becomes accessible to other infrastructures and systems. Our approach includes decisions drawn from research and best practices in other infrastructures/systems and are often backed up by appropriate statistics. The deliverable also incorporates recommendations by external projects, such as EOSC and NGI, and adopts well-established principles, such as FAIR, to maximise interoperability and collaboration with external infrastructures/systems.

To ensure interoperability, we placed particular emphasis on the design of metadata, since they are essential to enable reuse, facilitate interoperability and maximise impact. A number of well-known metadata schema standards that are applicable for the purposes of the project have been studied and evaluated, leading to the decision to adopt one of the predominant ones. This is further expanded to accommodate the unique characteristics and peculiarities of SLICES.

The analysis presented in this deliverable also provides input to various parts of the project, such as the reference architecture and services provided, governance analysis, integration and interoperability with EOSC infrastructure and services and relations with international testbeds.

This is the first version of the DMP, which will be updated by the end of the project to also reflect the data management policy in the future SLICES-RI project. Furthermore, the DMP will be aligned with deliverable D7.1, which will fully define the Ethical issues concerning data sharing and compliance with the General Data Protection Regulation (GDPR).

Finally, we also present the Data Management Plan for SLICES-DS.

Table of contents

Acronyms

DGG - Data Governance Group
DMP – Data Management Plan
DPO - Data Protection Officer
DQM - Data Quality Management
EOSC - European Open Science Cloud
GDPR – EU General Data Protection Regulation
NGI – Next Generation Internet
RI – Research Infrastructure
SME – Small-Medium Enterprise

1 Introduction

Researchers and research funders nowadays require that research data is made available for other researchers to examine, experiment and develop further upon. Additionally, preserving the data in conjunction with how conclusions from the data were drawn, protects the researcher from challenges and allows for easier reproducibility of the results. Ensuring interoperability of the data means that the data can reach a wider audience, further strengthening collaboration and facilitating innovation. Fostering a culture of open collaboration, enabled by tools that allow for seamless data exchange and utilisation of advanced compute tools, encourages co-innovation and facilitates inter/multi/trans-disciplinary research.

Policies and procedures for regulating the management and publication of research data are necessary to achieve the aforementioned goals. To this end, SLICES develops a Data Management Plan that governs the management of data and discusses how they can become accessible to users, both within SLICES and outside. The plan demonstrates how SLICES conforms with the European Open Access policy, Open Research Data Pilot and FAIR principles in producing and managing research data. To accomplish this, we analyse requirements drawn from various tasks (see Appendix A - Relationship with other Project Deliverables) and define appropriate metadata (including compatible experiment description) on the data produced by or integrated into the infrastructure with the objective to ensure data accessibility, reusability and interoperability with data produced by similar infrastructures/experiments for enabling complex experiments and multi-domain research. To further enhance interoperability aspects, we study and adopt the recommendations by EOSC FAIRs project, GO FAIR initiative and RDA for FAIR data management, and general European Open Access to research publications and Open Research Data Pilot policies.

Finally, we address compliance with national and international regulations (e.g. GDPR), data protection and privacy, and ethical aspects related to data management.

Deliverable Organisation

The deliverable is organised as follows: Section 2 provides the end-user, technical and operational requirements that relate to the management of data within SLICES. The requirements analysis allows us to propose a holistic data management framework in Section 3, which covers the aspects of data governance (Section 3.1), relationship of the data management architecture with the overall SLICES infrastructure (Section 3.2), procedures to ensure data quality (Section 3.3), metadata management and interoperability (Sections 3.4, 3.5) and support for analytics (Section 3.6). We also discuss, resource allocation (Section 3.8). Then, in Section 4, we discuss how the metadata procedures can be aligned with FAIR principles. Next, Section 5 discusses the compliance of the data management plan with national and international regulations. Sections 6 and 7 consider aspects of Data Security and Privacy and Ethics. Finally, we provide key issues of the DMP following the H2020 DMP template in Section 8.

2 Requirements Analysis

This section analyses end-user, technical and operational requirements that relate to the management of data within SLICES. These requirements stem from analysing the requirements of scientific communities and other related stakeholders (e.g. industry partners), legal compliance and regulation issues at a national, European and international level, interoperability with existing and future platforms and many more.
The analysis has resulted in the identification of the user groups and their characteristics, as well as the data they manage for research. Based on the types and formats of the data, requirements have been drawn for the information model, which is required for storing both data and appropriate metadata to ensure their discovery. Further requirements for interoperability are then drawn, as well as the types of intellectual property rights (e.g. licenses) that SLICES needs to support. The Data Management Framework described in Section 3, designs appropriate procedures, protocols and services aiming to fulfil these requirements.

2.1 User Groups

SLICES infrastructure aims to give opportunities to researchers and industry practitioners who lack access to a RI with sufficient size and diversity to support research and innovation tasks. It is necessary to identify the target user groups that will utilise the SLICES data infrastructure, in order to model their interactions appropriately. To this end, we have identified four user groups, which are illustrated in Figure 1.

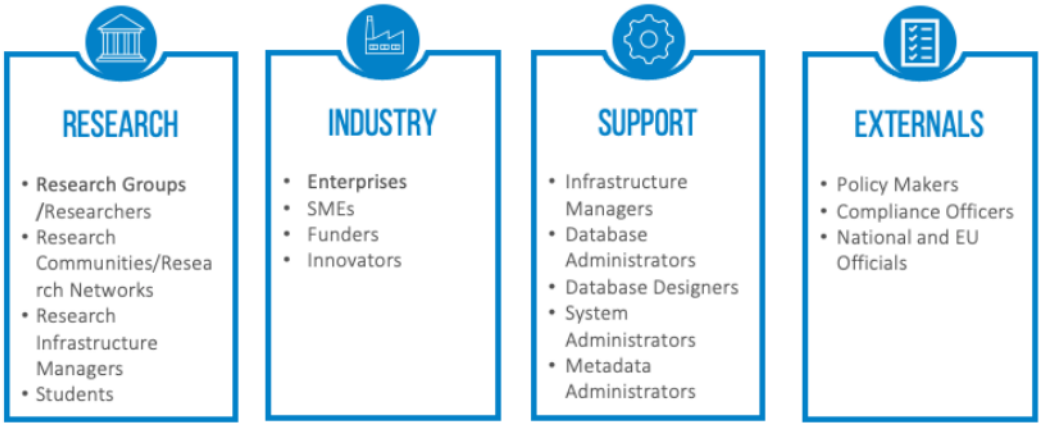

Figure 1: Target User Groups

Below, we provide details for each user group:

A. **Research User Group:** This user group includes *research-oriented organisations*, such as Universities, Colleges and Research Institutes, which primarily focus on the generation of new knowledge through state-of-the-art research and innovation. These organisations typically conduct research using hierarchical structures that include *research groups/clusters* composed of *researchers*, engineers and *students*. Research groups across institutions/organisations often coordinate to form larger *research communities/networks* that collaborate to tackle more complex problems. Users within this group typically utilise institutional or cross-institutional infrastructures to experiment, collect and analyse data to discover new patterns and unveil new knowledge.
   - **Access Frequency:** High, to support research actions
   - **User Type:** Sophisticated, can perform complex operations (e.g. predictive modelling) that may require significant resources

B. **Industry User Group:** This group includes *business-oriented organisations*, such as *enterprises* and *SMEs*, but also applied research and development organisations. The research infrastructure can support the design and execution of ad-hoc experiments. In the case of

SMEs, fast access to resources is essential in order to increase productivity and efficiency and create new opportunities in the competitive global landscape. It is also important to note that through this process, the industry organisations can acquire information about newly available technologies so that they can test them quickly. Furthermore, access to SLICES can also act as a single-entry point for SMEs, which aim to get access to both people and technology services available from the SLICES network. Finally, these collaborations may also be of interest to external *investor* stakeholders, such as *funders* and innovators, that are willing to fund specific endeavours.

- **Access Frequency:** Low, to support ad-hoc innovation actions
- **User Type:** Parametric, can perform complex operations using well-defined functions to draw conclusions fast

C. **Research/Industry Support User Group:** Research infrastructures are typically operated/supported by *research infrastructure managers* and technical personnel who ensure the effective and efficient operation of the infrastructure. Important users include: (i) Database administrators, who are responsible for managing access to the data, coordinating and monitoring the data infrastructure software and hardware resources, controlling its use, and monitoring its efficiency; (ii) Database Designers, who are responsible for defining and managing the metadata, altering the structure and the constraints of the data; and (iii) System Administrators, who are responsible for managing the infrastructure where the data resides (e.g. cloud, server array).

- **Access Frequency:** Very High, to support the 24/7 operation of the platform
- **User Type:** Expert, can perform maintenance operations

D. **External Partners User Group:** This group includes users such as compliance officers, policy makers, educators and civil servants who use information to ensure compliance, enhance regulatory processes, supervise operations, drive new policies or utilise the results of science to enhance their environment and the society in general.

- **Access Frequency:** Very Low, to support a specific task
- **User Type:** Knowledgeable, can perform pre-defined operations

2.2 Types of Data

Understanding the data collected and processed within SLICES is essential to understand data usage and the interaction with target user groups. This, in turn, will allow us to develop an appropriate information model that will represent the data collected from the SLICES testbeds, experimental equipment and applications.

The datasets collected in the SLICES testbeds and experiments will be of value and used for numerous IT/networking infrastructures, and will enable fast and effective implementation of new networking and big data infrastructure technologies. The datasets generated by utilising the SLICES hardware and software infrastructure can be roughly summarised into five main categories:

- **Observational Data:** will be collected using methods such as surveys (e.g. online questionnaires) or recording of measurements (e.g. through sensors). The data will include mostly data related to signal or performance measurements, and network or service log data that allow for experiment evaluation and reproducibility. It is important to note that this data will be often captured in a real-time manner and most probably cannot be reproduced if lost.
- **Experimental Data:** where researchers introduce an intervention and study the effects of certain variables, trying to determine whether there is any correlation/causality. Both observational and experimental data are essential for technology evaluation and will be of interest to the wider research and industry community for making informed decisions about new technology implementations and/or improvements.
- **Simulation Data:** is generated by using computer models that simulate the operation of a real-world process or system. These may use observational data.

- **Derived Data:** involves the analysis (e.g. cleaning, transformation, summarisation, predictive modelling) of existing data, often coming from different datasets (e.g. the results of two experiments), to create a new dataset for a specific purpose. This data is required for data driven experiments and research in Artificial Intelligence or Machine Learning-driven experiments and technologies that are increasingly used in networking, telecommunication and data driven research, as well as facilitated by Industry 4.0 technologies, such as Digital Twins, Automation and Robotics.
- **Metadata:** concerns data that provides descriptors about all categories of data mentioned above. This information is essential in making the discovery of data easier and ensuring their interoperability.

The types of data are explained in Table 1. All types come in a variety of formats, mostly unstructured and semi-structured, thus requiring the development of a non-relational distributed database. The **data model will need to ensure openness and flexibility to accommodate for different format requirements**, while in parallel, **offer high performance storage, processing and retrieval operations**.

Data Category	Sources	Processing	Characteristics/Examples
Observational	- Surveys - Recorded measurements	No processing of data. As collected by the research tools (e.g. questionnaires, sensors).	Signal or performance measurements, network or service logs, real-time captures. Essential for technology evaluation.
Experimental	- Experiments - Interventions	Data characterises variables in an attempt to determine if there is any correlation/ causality.	Essential for technology evaluation, technology improvements.
Simulation	Simulation software environments	Implemented system models generate the data. These may be informed by the use of observational data.	Imitate the operation of a real-world process or system.
Derived	Existing datasets	The analysis (e.g. cleaning, transformation, summarisation, predictive modelling) of existing data, often coming from different datasets to create new ones.	Data driven experiments and research in Artificial Intelligence; Machine Learning-driven experiments and technologies that are increasingly used in networking, telecommunication and data driven research.
Metadata	Descriptors for data from all other categories	Processing follows rules for descriptors.	Essential in making the discovery of data easier and ensuring interoperability.

Table 1: Overview of Types of Data collected in SLICES

2.3 Formats of Data

Different formats of research data exist (e.g. file formats, database formats), which impact directly the ability of a system to manipulate the corresponding data and provide access to other users down the line. These formats can be roughly split into two categories, open and proprietary file formats.
Open file formats[1] (e.g. csv, png) make their specification (i.e. structure) publicly available for others to use or implement. Such formats are essentially standardised by public authorities or international bodies with the objective to improve software interoperability. Moreover, open formats can be either coded in a text format (i.e. human readable format) that can be viewed in a browser, or binary format that is not human readable, but decodable when the format specifications are known.
Proprietary file formats (e.g. rar, sas7bdat) contain a non-transparent structure and its specification is not made publicly available by the software company that developed it. Software companies create such file formats for encoding the outputs of their applications, thus making it only possible to read by using the specific company software, which contains the needed file format specification for decoding.

One of the main objectives of SLICES is to promote interoperability, thus **non-proprietary, unencrypted, uncompressed, and commonly used by the research community formats should be adopted**. When storing the data in open formats may lead to loss of information or structure, it is recommended to also store the data in the original proprietary format. In such cases, it is also recommended to advise users to prepare extra documentation that lists the necessary software (or provides link(s) to download it), the appropriate version and other constraints that the format entails, to further improve data reuse[2]. The library of Cornell University further suggests some file format characteristics that support long term data preservation. These include complete and open formats that are platform/vendor independent, without partial or full encryption and/or password protection. Similar file format characteristics and appropriate file formats are also recommended on the website of The University of Edinburgh[3] and the website of Northwestern University[4]. The topic of recommended formats is further explored in the Recommended Formats Statement, published by the US Library of Congress[5].
In conclusion, we suggest that SLICES recommends the following file format specifications to be used for storing research outputs:

- Open (non-proprietary) file formats
- Uncompressed, or compressed with a free/open format such as .7z
- Unencrypted, or supplemented by appropriate decryption mechanisms
- Commonly utilised by the research community (e.g. csv) or highly interoperable among diverse platforms, systems and applications (e.g. YAML, JSON, XML)
- Developed and maintained by an open standards organisation, with a well-defined inclusive process for evolution of the standard

Some preferred file formats that should be recommended to SLICES users are listed below:

- Containers: TAR, GZIP, ZIP
- Databases and tabular data: YAML, JSON, XML, CSV
- Statistics: ASCII, DTA, POR, SAS, SAV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF

[1] Web Archive: Openformats.org Open vs. Proprietary formats, https://web.archive.org/web/20130219012116/http://www.openformats.org/en1 [Last Accessed 09 February 2021]
[2] Stanford Libraries: Best Practices for file formats, https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats [Last Accessed 09 February 2021]
[3] The University of Edinburgh: Information Services, Research Data Service, https://www.ed.ac.uk/information-services/research-support/research-data-service/after/data-repository/choosing-file-formats [Last Accessed 09 February 2021]
[4] Northwestern University: Library, https://www.library.northwestern.edu/about/administration/policies/file-format-recommendations.html [Last Accessed 09 February 2021]
[5] Library of Congress: Recommended Formats Statement, https://www.loc.gov/preservation/resources/rfs/ [Last Accessed 09 February 2021]

- Video: MOV, MPEG, AVI, MXF
- Audio: WAVE, AIFF, MP3, MXF
- Image: TIFF, JPEG 2000, PDF, PNG, GIF, BMP

2.4 Dataset License Types

There is a growing need for data licensing to avoid legal challenges with data sharing. This is a multi-dimensional and complicated aspect, since different jurisdictions apply different rules and standards for different aspects of data (e.g. record values, attribute names, database model). Licences and waivers are instruments that allow users to permit a second party to access and reuse data. Licenses grant permissions given that specific conditions (e.g. attribution, copyleft and intent), which are set by the data owner, are met. To avoid the complexity of drafting a license from scratch, there are several standard licenses available for data owners[6,7], which we outline below:

- **ODC-By:** Open Data Commons Attribution License (ODC-BY) allows re-users of the data to distribute, copy, transform, build upon and produce works using the data for any purpose. New content or new databases generated as a result of using the licensed dataset must contain a notice mentioning the use of the licensed dataset.
- **ODC-ODbL:** Open Data Commons Open Database License (ODC-ODbL) is an extension of ODC-By, since it adds more conditions. Firstly, the "copyleft" condition is added in case new databases are derived from the original database. Secondly, technological restrictions may apply only to the database or a new database derived from it, only if another copy without the restrictions is made available.
- **ODC-DbCL:** Open Data Commons Database Contents License (ODC-DbCL) removes the copyrights of the contents of a database, but it does not affect the copyright of the actual database.
- **PDDL:** Open Data Commons Public Domain Dedication and License (PDDL) is nearly identical to CC0, but instead it uses wording that is specific to database terms, while it provides a set of community norms to be associated with a database.
- **CC0:** Creative Commons Zero (CC0) allows for waiving all database rights and copy right interests to the public domain. Moreover, it can act as an irreversible loyalty free/ unconditional license for anyone who wants to use the data for any purpose.
- **CC PDM:** Creative Commons Public Domain Mark (CC PDM) acts as a tool that asserts works as already being a part of the public domain, thus allowing public works to be more easily discoverable and recognisable as public. Unlike CC0, it cannot waive work rights.
- **CC BY:** Creative Commons Attribution (CC BY) is one of the open Creative Commons licenses that is only described by a single condition, i.e. attribution. This license specifies that the re-user of the data must provide credit to the licensor when the work is distributed, displayed, performed or used to derive a new work.
- **CC BY-SA:** Creative Commons Attribution Share Alike (CC BY-SA) is an extension of CC BY, since it has an additional condition, i.e. Share Alike, which requires re-users that transform, remix or build upon the licensed dataset to distribute their contributions under the same license as the original.
- **CC BY-NC:** Creative Commons Attribution Non-Commercial (CC BY-NC) is an extension of CC BY, since it has an additional condition, i.e. Non-Commercial, which prevents re-users from using the licensed dataset for any commercial purposes.
- **CC BY-ND:** Creative Commons Attribution No-Derivatives (CC BY-ND) is a more restrictive extension of CC BY, since it has an additional condition, i.e. No-Derivatives, which prevents re-users from making additions, transformations or any type of changes to the dataset.

[6] Ball, A. (2014). 'How to License Research Data'. DCC How-to Guides. Edinburgh: Digital Curation Centre, https://www.dcc.ac.uk/guidance/how-guides/license-research-data [Last accessed 09 February 2021]
[7] DataWorld Documentation, Common License Types for Datasets, https://help.data.world/hc/en-us/articles/115006114287-Common-license-types-for-datasets [Last accessed 09 February 2021]

- **CC BY-NC-SA:** Creative Commons Attribution Non-Commercial Share Alike (CC BY-NC-SA is one of the most restrictive licenses. It allows users to share a dataset only if they provide credit, avoids using the dataset for commercial purposes and makes sure to redistribute their changes using the same license.
- **CC BY-NC-ND:** Creative Commons Attribution Non-Commercial No-Derivative (CC BY-NC-ND), yet another one of the most restrictive licenses, allows users to share a dataset if it is unmodified and not being shared for commercial reasons. No additions or transformations are permitted on the dataset.
- **CDLA-Permissive-1.0:** Permissive Version 1 is one of the Community Data License Agreement licenses and it is very similar to permissive open-source licenses. The re-users of the data can modify, adapt and share the data, as long as they provide credit, while no obligations or restrictions are imposed on derived computation results.
- **CDLA-Sharing-1.0:** Sharing Version 1 is one of the Community Data License Agreement licenses and aims to put the "copyleft" principles in use for data. Re-users are allowed to adapt, modify and share the dataset or their derived changes, but only under the CDLA-Sharing, while also giving credit to the owner of the licensed dataset. No obligations or restrictions are imposed on derived computation results.
- **OGL:** Open Government License (OGL) is a license intended for the UK public sector and cannot be used by licensors outside the UK. It is similar to CC BY since it has the attribution requirement. Moreover, derivative works and commercial uses are also allowed, while no "copyleft" condition exists. A non-commercial variant of this license also exists (NCGL).
- **Bespoke or Custom License:** Can be used when datasets contain high commercial value or when explicit responsibilities to re-users of the data need to be specified. Templates such as the Restrictive License (RL) can guide the preparation of bespoke licenses.

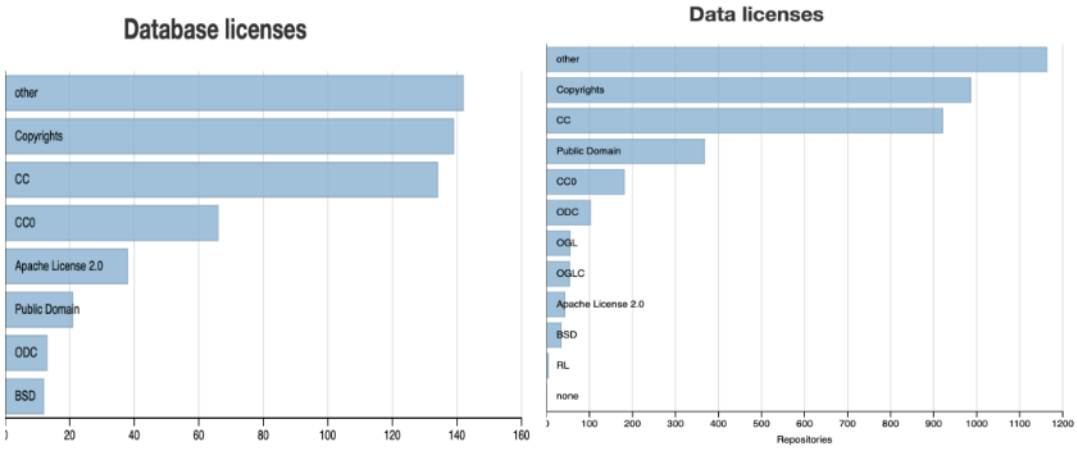

Figure 2: Data and Database Licenses usage statistics by re3data

As can be seen from re3data[8] statistics, some of the aforementioned license types, such as Copyrights, CC and Public Domain, dominate the license utilisation landscape. However, it is also noteworthy that the majority of repositories allow users to specify the other license types too.

SLICES end users will need to have the ability to decide on a suitable license and attach it to their data. This should be explicitly supported by the metadata repository using a library of license types that users can choose from when constructing their metadata. In those cases where a suitable license does not exist, then **Public Domain should be provided as a default option**, but the option to select Other should also be available. Additionally, applications that query for metadata should display this information prominently, so that a second party will immediately realise that the data must be licensed prior to accessing it.

[8] Re3data, Statistics, Data and Database licenses usage, https://www.re3data.org/metrics/dataLicenses [Last Accessed on 03 February 2021]

2.5 Expected Data Size

Testbed as a Service (TaaS) was developed for several years in order to provide a "ready to go" environment for experimental activities by providing easy access to the required communications, computing and storage resources for the experiments. OneLab/FIT[9], Grid'5000[10], GENI[11] and Fed4Fire+[12] are examples of such solutions that have hosted hundreds of thousands of experiments and thousands of users. However, the limitations and bottlenecks of the current Internet have called for a new design supported by recent worldwide initiatives, such as NSF/PAWR[13] and NSF/Fabric[14] in the US, but unfortunately, none of that kind in Europe.

SLICES aims to design and develop a distributed research infrastructure with next generation capabilities that will host thousands of users and their data. Our preliminary estimations include up to 5,000 users and their data, accounting for up to 50GB per user on the individual nodes and up to 1TB on the cloud. This provides us with a preliminary estimation of 0.25PB-1PB of data storage for all datacenters residing on SLICES nodes, and 5PB for the cloud-based datacenter.

2.6 Interaction with Other Infrastructures/Systems

As the SLICES architecture relies on highly modular components, the integration with existing Next Generation Internet (NGI) European and international digital technologies will come in a standardised manner. To this aim, SLICES consortium has studied the APIs and manner of interaction with existing infrastructures and has pinpointed the wide use of Network Function Virtualization (NFV) architectures across the community. SLICES will adopt similar APIs, which will allow the interaction and integration with relevant infrastructure as separate domains. This will allow the infrastructure to summon resources under a unified architecture and enable a plethora of different scenarios/use cases to be evaluated through the use of single-point-of-entry and consistent APIs across the community. The different APIs and policy enforcement across different domains is a topic that will be further analysed, along with agreements, compliance issues and technical issues that can come up during the integration process.

It is important to note that some infrastructures/systems (e.g. EOSC portal) provide specific requirements for onboarding. Below, we provide an indicative list of these requirements that have been taken into consideration for the design of the data management framework presented in the next section:

- Services that will be exposed or integrated into other infrastructures must be operational and provide a specific service that is not trivial; e.g. a metadata discovery service vs. a link to a dataset.
- Resources should provide appropriate documentation related to privacy, licensing, terms, etc.
- Data should include rich metadata to ensure effective discovery.
- Data should also adhere to the FAIR Principles (Findable, Accessible, Interoperable and Reusable).
- Appropriate procedures and processes should exist for compliance with national and international regulations (e.g. GDPR).

[9] OneLab, https://onelab.eu/ [Last Accessed on 03 February 2021]
[10] Grid'5000, https://www.grid5000.fr/w/Grid5000:Home [Last Accessed on 03 February 2021]
[11] Geni, https://www.geni.net/ [Last Accessed on 03 February 2021]
[12] Fed4Fire+, https://www.fed4fire.eu/ [Last Accessed on 03 February 2021]
[13] Platforms for Advanced Wireless Research – PAWR, https://advancedwireless.org/ [Last Accessed on 03 February 2021]
[14] Fabric, https://fabric-testbed.net/ [Last Accessed on 03 February 2021]

3 Data Management Framework

This section proposes a data management framework to support the efficient and effective operation of the SLICES infrastructure and achieve the project's objectives. To accomplish this, the data management framework sets its own design goals, which are summarised below:

- **Data Governance**: A systemic and effective Data Governance structure to support the data management operations through a hierarchical structure with appropriate roles (e.g. Data Manager, Data Protection Officer and Metadata administrator), implement all related policies and processes, and adopt standards and leading practices.
- **Data Architecture**: An agile Data Architecture that can perform efficiently to fulfil the SLICES infrastructure requirements, scales gracefully to accommodate for increased workloads, is flexible to integrate new processes and technologies, and is open to interact with other systems and infrastructures.
- **Data Quality**: Appropriate data transformation mechanisms to ensure Data Quality across multiple dimensions (e.g. accuracy, completeness, integrity), in order to improve data utility (e.g. further processing, analysis).
- **Metadata**: Appropriate metadata management mechanisms to facilitate collaboration between users by providing the means to share their data and also support FAIR data.
- **Interoperability**: Facilitate seamless interaction with other systems and infrastructures.
- **Analytics**: Deployment of statistical, machine learning and artificial intelligence techniques to draw valuable insights from data and appropriate visualisation techniques to interpret them.
- **Data Security:** Mechanisms to protect data from unauthorised access and protect its integrity.
- **Privacy**: Strict controls to manage the sharing of data, both internally and externally.

To achieve the aforementioned objectives, the Data Management Framework must integrate infrastructure, people and systems, as illustrated in Figure 3. We consider that all processes and systems can be adapted, refined and optimised, in order to respond to changing end user requirements. As such, after the infrastructure operates for a specific period of time, appropriate metrics should be evaluated for all dimensions. This will allow for assessing if the objectives are met and drawing recommendations for improvement, but also identifying problems and planning for contingencies.

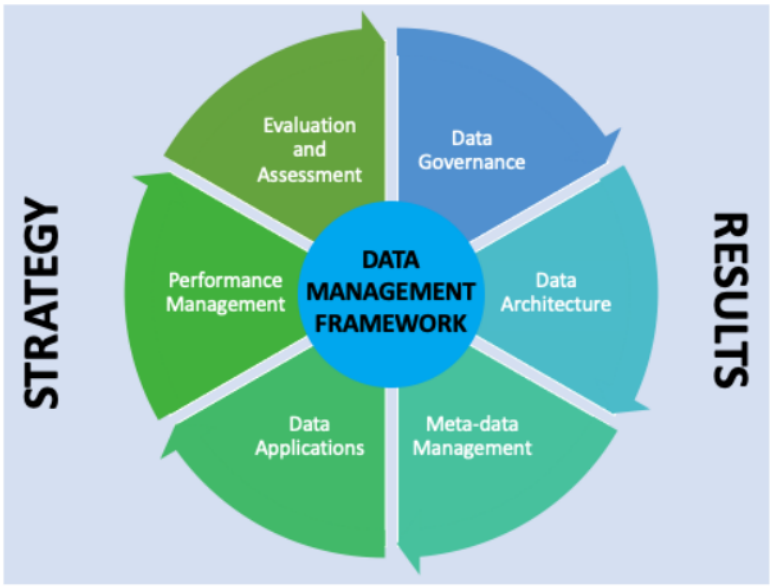

Figure 3: Data Management Framework

The following sections describe how each design goal is met.

3.1 Data Governance Framework

SLICES will collect and manage datasets generated from interconnected testbeds, incorporating a plethora of applications and technologies. This, in turn, allows for researchers and practitioners in Europe and beyond to address numerous challenges in experimental-driven research in networks, distributed computing, big data, etc. However, to effectively exploit the data asset it is imperative to deploy an appropriate data governance framework suited for the current and future SLICES objectives. In this section, we describe the data governance framework, which is part of the larger SLICES governance framework (defined in D3.1). The framework aims to provide appropriate controls to oversight data assets, ensure their value and maximise their impact. It also addresses the legal and compliance dimensions, which are important towards meeting specific regulation requirements, such as GDPR compliance. We provide an overview of the data governance framework, the data that it manages and the involved stakeholders, including their roles and responsibilities in relation to the research infrastructure eco-system that SLICES operates.

3.1.1 Organisational Model

Data Governance strongly depends on a rigorous organisational management model and clearly defined organisational roles, responsible for all aspects of data management. To this end, SLICES defines a structured organisational model, as illustrated in Figure 4, to address all data governance issues.

The Data Governance Group (DGG) is the main decision-making body for data management aspects, and it is part of the Coordination & Management Office of SLICES, as this is detailed in deliverable D3.1. The main responsibilities of DGG include consultation of the management committee for all data related aspects and implementation and overseeing of the data management plan. In order to coordinate all data management activities, DGG elects a thematic director every two years, coined Data Manager, which is responsible to coordinate the data management team and monitor policy implementation.

The Data Manager will collaborate with the data governance group and technical teams to enforce policies, coordinate between the management committee and technology groups, establish appropriate Key Performance Indicators, monitor and report on data quality and data governance metrics, and prioritise and resolve issues with the technical team. Additionally, the Data Manager is responsible for facilitating collaboration within team members and establishing communication channels, both with internal and external partners (e.g. define the means of interaction with other research infrastructures). The Data Manager will also collaborate with the Data Protection Officer (DPO) to facilitate and monitor processes related to the protection of data and compliance to national and international regulations.

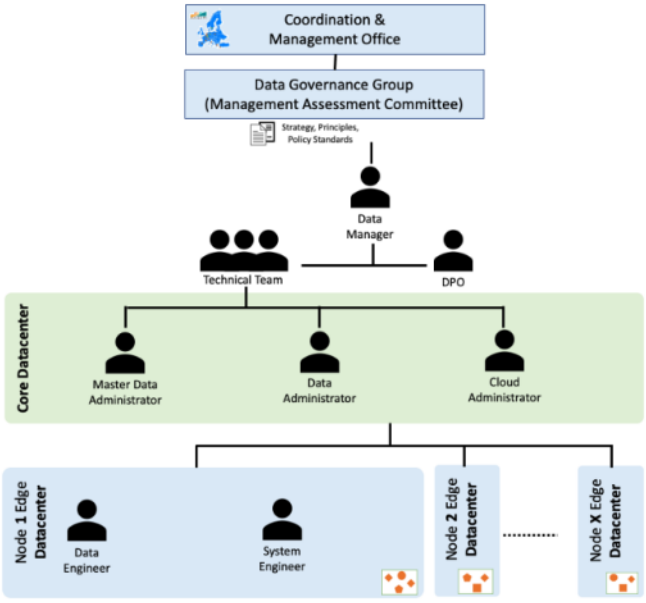

Figure 4: Data Governance Structure

The Data Manager will coordinate the Core Datacenter team, which is the team handling the cloud-based data management infrastructure and the Node Technical Teams, which operate in each SLICES node. The Core Datacenter team will be composed of: (i) the Master Data Administrator, who is responsible for metadata management; (ii) the Data Administrator, who is responsible for the data infrastructure maintenance; and (iii) the System Administrator, who is responsible for maintaining the cloud-based infrastructure that the data and tools reside. Finally, each node will employ a data engineer and system engineer who will be responsible for maintaining the local infrastructure and tools related to data management.

3.1.2 Roles and Responsibilities

The roles and their responsibilities are provided in Table 2. We also provide additional information related to responsibilities for appointments and reporting supervisors for each role.

Role	Appointed By	Reports to	Responsibilities
Data Manager	Data Governance Group	Data Governance Group	Policy Implementation Monitoring/Reporting Internal Communication/Training Coordination of Technical Teams External Communication Policy Monitoring/Improvement Data Lifecycle Management Data Quality Monitoring/Enhancement Change Management
Data Protection Officer	Data Governance Group	Data Manager	GDPR Strategy Implementation Compliance (GDPR, National, International) Risk/Privacy Impact Assessment/Monitoring Communication (Contact point for data subjects)

Master Data Administrator	Data Manager	Data Manager	Metadata Management Metadata Accessibility Monitoring of Data Standards FAIR Compliance
Data Administrator	Data Manager	Data Manager	Monitoring/ Configuration Data Architecture/Modelling Database Security Data Quality Backup and Recovery Troubleshooting
Cloud Administrator	Data Manager	Data Manager	System Monitoring Cloud Resource Planning/ Management Tool Integration Infrastructure Security Descriptive analytics for System Operations
Data Engineer	Node Director, Data Administrator, Master Data Administrator	Node Director, Data Administrator	Metadata Accessibility Monitoring/ Configuration Database Security Data Quality Backup and Recovery Troubleshooting
System Engineer	Node Director, Cloud Administrator	Node Director, Cloud Administrator	System Monitoring Resource Planning/ Management Tool Integration Infrastructure Security

Table 2: Data Governance Roles and Responsibilities

3.1.3 Policy Enforcement and Maintenance

The data governance structure aims to guarantee and safeguard data quality and its effective use. As such, the data governance group and its appointed Data Manager will need to ensure that everyone involved in data management will adhere to the policies and procedures, especially when concerning data quality. To accomplish this, the Data Manager will design and implement appropriate procedures to make the data management policies accessible, so that all involved parties have easy access to them. More information related to the role of the Data Manager will be provided in D3.1.
Furthermore, a procedure should be defined, so that the data management plan is reviewed periodically (e.g. each year) to assess how adherence to the policies has led to a positive impact on SLICES or if inefficiencies have been identified and need to be resolved. Its impact and efficiency will be measured primarily as means of meeting its objectives and achieving its key metrics. This periodic review will also serve as a means for incorporating new/revised objectives in the case where SLICES' strategies change.

3.2 Data Architecture

An agile Data Architecture that can deliver usable data to its end users is essential to fulfil the SLICES objectives. This section describes the environment where data will reside and defines the processes that capture, transform and deliver usable data to end users. Amongst the key objectives that this architecture aims to meet are to scale gracefully and accommodate for increased workloads, to be

flexible in terms of integrating new processes and technologies, and to be open to interact with other systems and infrastructures.
Figure 5 illustrates the Data Architecture for SLICES that complements the testbed infrastructure. This architecture provides a preliminary blueprint for the interconnectivity and utilisation of its primary components. It is expected that the architecture will be constantly updated, in order to accommodate for new technologies that may arise upon the time it will be developed.
The architecture has two levels of data centers: (i) Distributed, located at each SLICES node, supplemented with individual edge/core sites depending on the experimental equipment that the node provides; and (ii) Centralised, located in the cloud with connections from the distributed data centers.

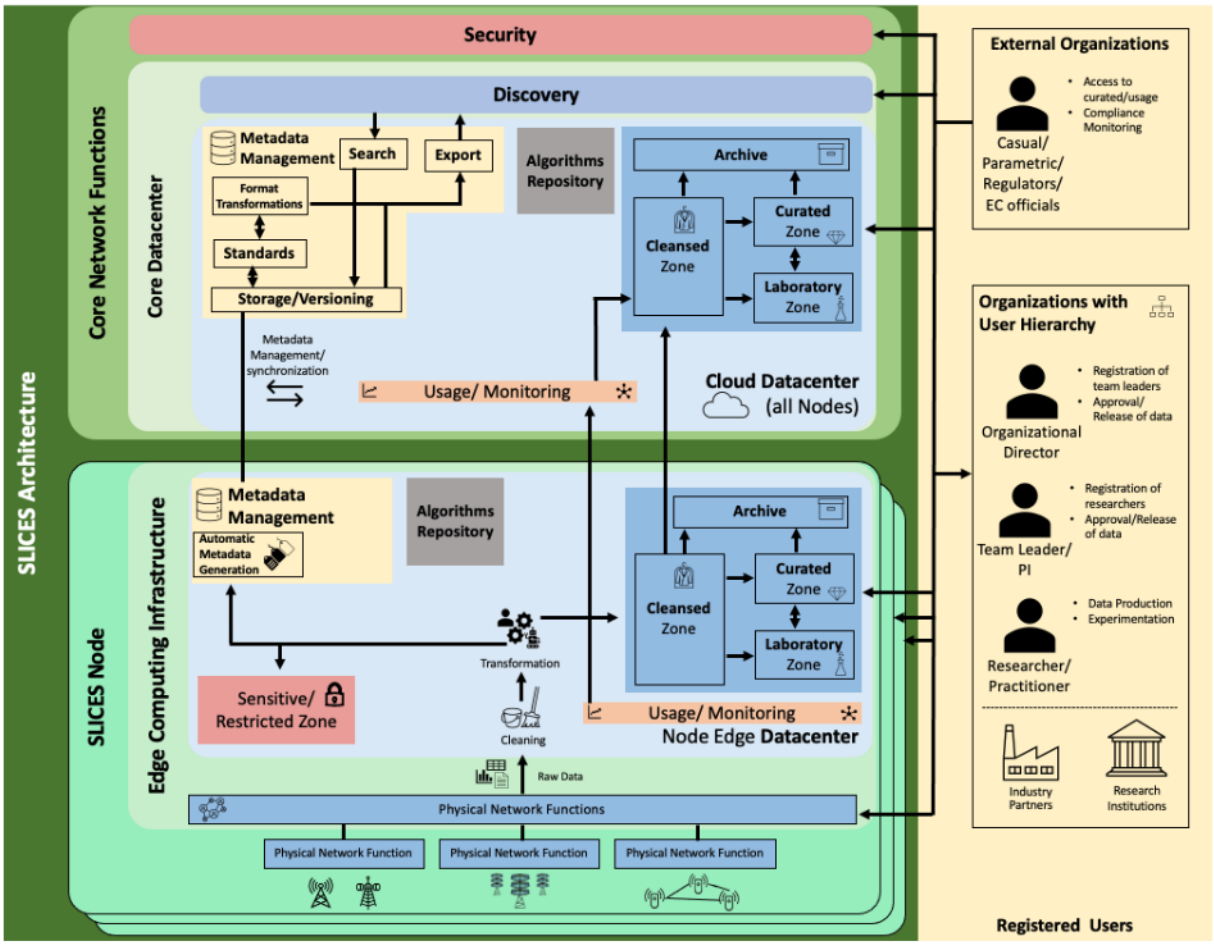

Figure 5: Data Architecture

3.2.1 Node Computing Infrastructure

Starting from the bottom, the raw data is generated from experiments that will use either resource virtualisation or meta-access to the resources and are transmitted through physical network functions. When the data reaches the Node Datacenter, it undergoes a series of predefined or selected pre-processing tasks (e.g. cleaning, integration, reduction) to ensure a defined level of quality.
The data will then be transformed to conform to a well-defined information model and appropriate metadata will be automatically extracted or manually entered beforehand by the user who created them (Creator). Metadata will be stored in the metadata database where they will be made accessible by other mechanisms to facilitate sharing.

Datasets marked by the Creator as sensitive/private or data that contains sensitive information (e.g. personal data) will be stored in a Restricted Zone with more enhanced security mechanisms and its access is limited to its creators only. After these steps, datasets are then stored in the Cleansed Zone. The Cleansed Zone represents an almost identical copy of the source data, however, transformed according to the metadata requirements employed in SLICES. This data will be immutable to changes and can be retained for long periods of time (e.g. a project duration of three years), unless it is archived. The Laboratory Zone will provide the means for very efficient exploration and analysis of the data and will allow for compute-intensive operations on the data. These operations may produce new data that will supplement the original or even generate new summarised datasets, which can be exported to the curated zone. The Curated Xone will contain data that have undergone transformations, such as summarisation, aggregation, modelling and are optimised for information delivery.
The Archive will contain mechanisms for preservation of the data for longer periods of time, such as for four years or more. Additionally, it will allow for backing up all zones for recovery purposes.
Finally, the Usage Monitoring module will deploy components that will measure the resource utilisation of the infrastructure across multiple dimensions, to allow for assessing whether the objectives are met, and accordingly, draw recommendations for improvement.

3.2.2 Core Datacenter Infrastructure

The Core Datacenter will leverage the cloud to collect, store, process and distribute large amounts of data (Big Data). Although it will feature similar components with SLICES nodes, it will offer many additional benefits, such as improved efficiency of computations, seamless scalability and enhanced data security, which will enable researchers at the European level to explore, experiment and tackle challenges of the future Internet.
The cloud-based data management infrastructure will allow researchers to **integrate and analyse big data** from multiple SLICES nodes using advanced storage and compute components to support complex calculations at high speeds and dataset sizes that span from terabytes to petabytes. It will also integrate data from other sources, such as standards, vocabularies and datasets provided by other organisations to enable data triangulation for validating uncovered insights.
Metadata will be assimilated from each individual node and will be transformed/synchronised to support the dataset integrations. A dedicated Discovery module will deploy appropriate APIs (e.g. REST) for exposing metadata information to end-users and supporting advanced querying, which allows users to use free-text search, combine filters and drill down to the individual components of each metadata property. Furthermore, a translation engine will be provided for transforming the implemented metadata format to the majority of well-known formats.
Interoperability with other research infrastructures and systems (e.g. EOSC, NGI) will be facilitated by supporting at least one metadata harvesting protocol (e.g, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[15]) and streamlining data dissemination using popular data exchange formats, such as XML, JSON and YAML for data serialisation (while being extendible to support other formats in the future).
Analytics will allow the discovery of hidden insights and uncover interesting patterns, so that researchers can make better, data-driven decisions faster. The datacenter will integrate a variety of open-source and commercial tools to support real-time data analysis for different application domains, such as data mining, machine learning and artificial intelligence.

3.2.3 Discovery

Discovery components should support a suitable API (e.g. REST) for managing and querying for data. It is important that the API features a number of qualities, such as simplicity, usefulness, consistency

[15] The Open Archives Initiative Protocol for Metadata Harvesting, https://openarchives.org/OAI/openarchivesprotocol.html [Last accessed 01 February 2021]

and predictability. Based on the collected requirements, the following four commands should be the minimum set supported by the SLICES metadata discovery API:

- **upload:** to upload data into the repository or edit existing ones.
- **query:** to search for metadata in the repository using a number of different types of queries (e.g. selection, range, pattern matching); this will also return appropriate links to files.
- **download:** to download specific files.
- **analytics:** to query the platform for statistical information regarding the operation of the platform (e.g. number of metadata containing licenses and their types).

The API will also be accessible through an appropriate, easy-to-use user interface, which provides adequate documentation for external users to test and use it.

3.3 Data Quality Assurance

One of the main objectives of SLICES is to ensure data accessibility, reusability and interoperability with data produced by similar infrastructures/experiments for enabling complex experiments and multi-domain research. To accomplish this, SLICES needs to provide tools that are geared towards ensuring data quality, such as identifying and correcting flaws and inconsistencies. Many factors affect data quality, including accuracy, completeness, consistency, timeliness, believability and interpretability.

SLICES will employ a set of Data Quality Management (DQM) tools to ensure accuracy, consistency and interpretability. Factors such as completeness, timeliness and believability cannot be tackled directly through the tools, but they can be "inferred" partly from the previous measures.

SLICES DQM tools fall into five different categories:

- **Data Cleaning**: will provide tools for dealing with missing and noisy data, such as imputing values based on some measure (e.g. mean, median or mode), replacing values using user-defined cases (e.g. global constants or rules), outlier detection using known metrics (e.g. confidence intervals, boxplots/IQR) and smoothing (e.g. by binning or regression).
- **Data Integration**: will provide tools for addressing semantic heterogeneity and structure, such as entity linking (e.g. based on attributes or rules), redundancy detection and elimination, etc. These tools will also allow for integration with external datasets (e.g. weather measurements, currency conversion) that can be used for normalisation (see Data Transformation).
- **Data Reduction**: will provide tools to obtain a reduced representation of the dataset, either in terms of attributes or volume, such as dimensionality reduction (e.g. correlation analysis, projection to smaller spaces (e.g. PCA), attribute subset selection), numerosity reduction (e.g. regression and log-linear models), sampling (e.g. random, stratified) and summarisation (e.g. aggregation, generalisation).
- **Data Transformation**: will provide tools for transformation/consolidation, such as smoothing, attribute construction (e.g. feature generation), summarisation (e.g. aggregation, generalisation), normalisation and discretisation.
- **Data Interpretation**: will provide tools to generate appropriate interpretations of the data, such as tabular and graphical representations, to evaluate low-level and complex tasks, such as understanding distributions, identifying trends and discovering anomalies/extremums. Additionally, tools to interpret complex modelling techniques can be integrated to provide more insight to end users, such as autonomous examination, network modelling, data mining functions and surrogate modelling with soft decision trees.
- **Data Interpretation**: will by guaranteed by the various security measures, procedures and protocols that will be undertaken to ensure appropriate data/access control, backups and recovery for the data. More information is provided under Section 6.

Completeness of the data cannot be guaranteed by the aforementioned tools. The tools can provide insight on the level of completeness, such as providing the percentage of missing values per attribute

or identifying gaps in a time series, but only the contributor of the data can ascertain that the data are complete. As such, the metadata need to include specific descriptors for completeness.
Timeliness of the data can be determined by two main factors. First, the creator of the data should guarantee that the recorded data timings between data capture and the real-world measurements are accurate. Second, there are cases where the timeliness can only be guaranteed at the time where the data will be utilised (e.g. the date/time range can be utilised for the analysis). As such, the metadata need to include specific descriptors for timeliness; the creator's confirmation that the timeliness is verified, for the cases that the data contain temporal factors.
Finally, believability, which demonstrates the extent to which data originates from trustworthy sources, can only be achieved through implicit means[16], such as identifying the trustworthiness of the source and reasonableness, which cannot be quantified within the platform.

3.4 Metadata Management

Metadata are essential to enable reuse, facilitate interoperability and maximise impact. One of the main objectives of SLICES is to facilitate collaboration of researchers/practitioners, both within SLICES and beyond, by developing an efficient and interoperable metadata repository.
To accomplish this, appropriate Metadata Management procedures should exist, so as to allow for management and sharing of the research data in an intuitive and safe way, respecting the creators' rights, while ensuring compliance with open access principles, such as FAIR.
To this end, we define the following objectives for metadata management:

- **Reusability:** the repository must incorporate appropriate descriptor elements to comprehensively describe metadata and provide appropriate functions to query and retrieve the stored data.
- **Interoperability:** although data will be stored in a particular metadata format, the repository must support transformation of the data to other metadata formats, in order to ensure interoperability with other systems. Additionally, the infrastructure should provide interoperability services to enable researchers/practitioners, content providers, funders and research administrators to collaborate or utilise an existing platform to do so.
- **Data Quality:** is important for ensuring reusability and interoperability with other infrastructures/platforms and applications that may want to consume data from SLICES; this is already covered in Section 3.3.
- **Governance:** the members of the Data Governance Group should have the knowledge and authority to make decisions on how metadata are maintained, what the format being utilised is and how changes are authorised and audited.

To achieve the above objectives, we must first provide an overview of well-known metadata schema standards that are applicable for the purposes of the project. More precisely, we discuss various such standards that focus on different objectives related to record management and archives, providing support on multiple fronts (from aiding the archiving process to discovering, searching and preserving resources). A high-level description of these standards is provided next. A thorough guide to archival standards can be found in[17].

- **AGLS Metadata[18]:** The primary objective of the Australian Government Locator Service (AGLS) Metadata standard is to facilitate the search and discovery of resources, which are supplied by the Australian government. Such resources include either digital or non-digital items.

[16] N. Prat and S. Madnick, "Measuring Data Believability: A Provenance Approach," Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Waikoloa, HI, 2008, pp. 393-393, doi: 10.1109/HICSS.2008.243.
[17] A Guide to Archival Standards, https://www.archives.org.uk/about/sections-interest-groups/archives-a-technology/news-and-events.html [Last accessed 01 February 2021]
[18] AGLS Metadata Standard, http://www.agls.gov.au/ [Last accessed 01 February 2021]

- **AGRkMS[19]:** The Australian Government Recordkeeping Metadata Standard (AGRkMS) is based on AGLS Metadata and primarily deals with national archives. AGRkMS defines the metadata properties based on the metadata standard for record keeping (ISO 23081) that agencies in the Australian government should use when describing entities involved in their business and processes regarding record management.
- **EAD[20] and ISAD(G)[21]:** The Encoded Archival Description (EAD) is a metadata schema that is used for archiving digital resources. It allows for describing the actual content of such resources, but also its overall structure. On the other hand, the General International Standard Archival Description (ISAD(G)) is more generic in that it applies to more traditional archives, not necessarily of a digital nature. EAD is designed to be compatible with ISAD(G).
- **OAIS[22]:** The Open Archival Information System (OAIS) is an international standard that targets challenges related to preserving digital resources long term. More precisely, the OAIS reference model aims at providing access guarantees for archive systems, by outlining a set of functions required for both accessing and ensuring that digital resources are effectively preserved over time.
- **PREMIS[23]:** The Preservation Metadata and Implementation Standard (PREMIS) is yet another metadata standard that at its core deals with matters related to the preservation of digital resources. To this end, PREMIS defines a data model for preservation along with the corresponding data dictionary. The former comprises five distinct entities, namely, the intellectual entity, digital object, agent, rights and event.
- **OpenDOAR[24]:** The Directory of Open Access Repositories (OpenDOAR) is a web-based directory that offers a list of open-access academic repositories. Based in the UK, it offers options for searching resources by locale, content and other measures. OpenDOAR is currently one of the two (2) leading open access directories worldwide.
- **Dublin Core[25]:** Dublin Core (or the Dublin Core Metadata Element Set) defines a set of fifteen core properties, drawn form a larger set of DCMI Metadata Terms, for describing resources. Dublin Core is formally standardised as ISO 15836, ANSI/NISO Z39.85 and IETF RFC 5013. The resources described using Dublin Core include any type of digital resources (e.g. videos, images, and web pages), but also physical resources (e.g. books, CDs, and artworks). The main uses of this standard extend from simply describing resources, to combining metadata vocabularies of different standards, as well as catering for interoperability of metadata vocabularies in the linked data cloud and Semantic Web implementations.
- **DataCite[26]:** DataCite provides persistent identifiers (DOIs) for research data and other research outputs. Organisations can join DataCite as members to assign DOIs to research outputs and make them discoverable to the community. DataCite then develops additional services to improve the DOI management experience, making it easier for members to connect and share their DOIs with the broader research ecosystem and to assess the use of their DOIs within that ecosystem.
- **MINSEQE[27]:** MINSEQE provides a description about the Minimum Information of a high-throughput nucleotide SEQuencing Experiment. This minimum information is required for

[19] Australian Government Recordkeeping Metadata Standard, https://www.naa.gov.au/information-management/information-management-standards/australian-government-recordkeeping-metadata-standard [Last accessed 01 February 2021]
[20] EAD: Encoded Archival Description, https://www.loc.gov/ead [Last accessed 01 February 2021]
[21] ISAD(G): General International Standard Archival Description - Second edition, https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition [Last accessed 01 February 2021]
[22] Open Archival Information System (OAIS), http://www.oais.info/ [Last accessed 01 February 2021]
[23] PREMIS Data Dictionary for Preservation Metadata, https://www.loc.gov/standards/premis/ [Last accessed 01 February 2021]
[24] OpenDOAR - Directory of Open Access Repositories, https://v2.sherpa.ac.uk/opendoar/ [Last accessed 01 February 2021]
[25] Dublin Core Metadata Initiative, https://dublincore.org/ [Last accessed 01 February 2021]
[26] DataCite, https://datacite.org/ [Last accessed 01 February 2021]
[27] http://fged.org/projects/minseqe/

unambiguous interpretation and reproduction of experimental results. The adherence to MINSEQE guidelines enhances the integration of multiple experiments in a wide variety of modalities, thus the value of high-throughput research being maximised.

- **DDI[28]:** Document, Discover and Interoperate (DDI) is a free international standard that describes data produced by surveys and other observational methodologies mainly focusing on the social, behavioural, economic and health sciences. Moreover, the standard supports multiple steps in the research data lifecycle, including conceptualisation, collection, processing, distribution, discovery and archiving.
- **EML[29]:** The Ecological Metadata Language (EML) is a community-maintained specification that provides a thorough vocabulary and a readable XML syntax for the purpose of documenting research data in the earth and environmental sciences. EML enables researchers to preserve and openly document and share their data and findings. Some of EML's core modules allow for (i) identifying/ citing data for describing various formats of data and research methods/ protocols, (ii) describing the structure and content of data, and (iii) annotating data with semantic vocabularies.
- **ISO 19115[30] and FGDC-CSDGM[31]:** These two standards provide means for describing geospatial data. ISO 19115 defines the schema for describing geographic information, as well as information for several properties of digital geographic data, including quality, spatial and temporal aspects and spatial references, amongst others. FGDC-CSDGM stands for Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata.
- **FITS[32]:** Flexible Image Transport System (FITS), currently maintained by the International Astronomical Union, is utilised for the interchange of information between observatories and for archiving. FITS is a file format that facilitates the storage, transmission and manipulation of scientific images and their associated data. Since its conception, FITS was seen as a transport format for more than still images, since it was designed to enable the unambiguous transmission of 1-D spectra, 2-D images or data cubes of three or more dimensions. Tabular data, i.e. two-dimensional data tables can also be stored in FITS.
- **MIBBI[33]:** Minimum Information for Biological and Biomedical Investigations (MIBBI) is a standard defining a set of guidelines for reporting on data extracted from biosciences. MIBBI allows a community to perform easy data verification, analysation and interpretation. If the standard's guidelines are followed, the facilitation of structured databases/ public repositories and the development of data analysis tools is achieved.

Our analysis revealed that Dublin Core is a mature and widely used metadata standard that is domain agnostic. This is also supported by re3data, probably the largest directory of data repositories[34] as illustrated in Figure 6.

[28] https://ddialliance.org/
[29] https://eml.ecoinformatics.org/
[30] https://www.iso.org/standard/53798.html
[31] https://www.fgdc.gov/metadata
[32] https://www.loc.gov/preservation/digital/formats/fdd/fdd000317.shtml
[33] https://fairsharing.org/collection/MIBBI
[34] Re3data, Metrics, Metadata Standards, https://www.re3data.org/metrics/metadataStandards [Last accessed 01 February 2021]

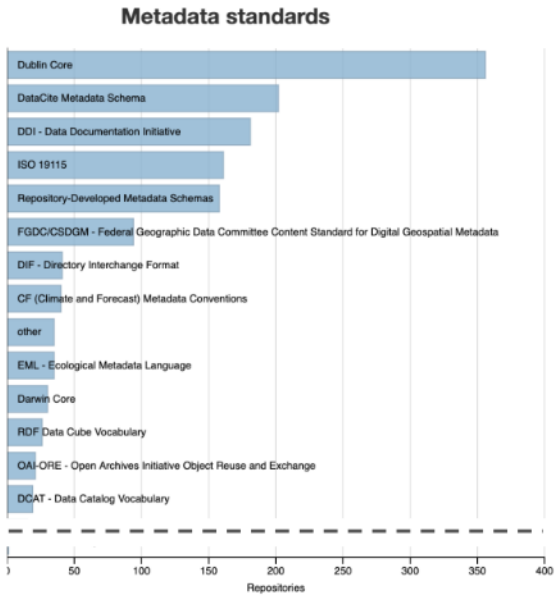

Figure 6: Metadata standards utilised by data management repositories (re3data)

As such, a revised version of Dublin Core, supporting the requirements proposed in previous sections would be an ideal solution for SLICES. Furthermore, automatic metadata generation, e.g. using Machine Learning techniques, should also be used to provide even more reusability and scalability for interacting with other infrastructure and external digital libraries and their collections. The following table proposes the elements suitable for describing a resource and its attributes in SLICES. We indicate which attributes come from the original standard[35] and what are the additions in the Notes.

Category	Label	Definition	Considerations
Instantiation	Date	A point or period of time associated with an event in the lifecycle of the resource. Includes the following sub-elements: • Date Submitted • Date Issued • Date Accepted • Date Copyrighted • Date Modified	Should use the ISO8601 format. Data Modified should be used in conjunction with versioning.
	Date Available	Date that the resource became or will become available.	Appropriate security/publishing mechanisms should be set in place to ensure that no user has access to the resource before publication date.
	Format	The file format, physical medium or dimensions of the resource. Includes the following sub-elements: • Has Format	Can use a list of open formats (e.g. Format Descriptions[36] and openformats.org

[35] DCMI Metadata Terms, https://dublincore.org/specifications/dublin-core/dcmi-terms/# [Last accessed 01 February 2021]
[36] Sustainability of Digital Formats, Format Descriptions, https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml [Last accessed 01 February 2021]

		• Extent • Medium	(accessible through Wikipedia) "Extent" can be used to validate consistency, believability and completeness.
	Identifier	An unambiguous reference to the resource within a given context. Includes the following sub-elements: • Bibliographic Citation • DOI	Tools for translating one bibliographic reference format to the other should be provided. Will produce persistent identifiers (DOIs) for research data and other research outputs.
	Language	The language of the resource.	Should use ISO 639-2.
Intellectual Property	Contributor	An entity responsible for making contributions to the resource.	
	Creator	An entity primarily responsible for creating the resource.	
	Publisher	An entity responsible for making the resource available.	
	Provenance	A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity and interpretation.	
	Rights	Information about rights held in and over the resource. Includes the following sub-elements: • License • Access Rights • Right Holder	License should be selected from a standardised list. Access Rights should be used in conjunction with SLICES property - Privacy Level
Content	Source	A related resource from which the described resource is derived.	
	Subject	The topic of the resource.	
	Title	A name given to the resource.	
	Alternative	An alternative name for the resource.	
	Description	An account of the resource. Includes the following sub-element: • Abstract	
	Type	The nature or genre of the resource.	Recommended practice is to use a controlled vocabulary, such as the DCMI Type Vocabulary.
	Audience	A class of agents for whom the resource is intended or useful. Includes the following sub-elements:	A vocabulary of audiences should be compiled. The user groups defined in Section

		• Education Level • Mediator	2.1, including their subcategories, can be used.
	Instructional Method	A process used to engender knowledge, attitudes and skills that the described resource is designed to support. Includes the following sub-element: • Coverage	
	Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource or the jurisdiction under which the resource is relevant.	
	Accrual Method	The method by which items are added to a collection.	
	Accrual Periodicity	The frequency with which items are added to a collection.	Should be validated against the actual data using ML approaches to improve Data Quality – Timeliness/Believability
	Accrual Policy	The policy governing the addition of items to a collection.	Should be made compulsory, if SLICES property consent for Completeness is provided.
	Relation	A related resource. Includes the following sub-elements: • Conforms to • Has Part • Has Version • Is Version Of • Is Format Of • Is Part Of • Is Referenced By • Is Required By • Is Replaced By • References • Replaces • Requires • Source	A list of standards should be used for "Conforms to". Versioning is essential to develop correct referencing and compatibility, especially for datasets that may change over time "Is Referenced By", "References", "Is Required By" should be calculated.
SLICES-specific	Consents	Creator's consents on aspects such as Completeness and Timeliness. Currently, the following consents have been identified: • Consent for personal data contained in project	

		• Consents of the purposes of the processing operations • Consent for Completeness of data • Consent for Timeliness of data	
	Auto	A structured descriptor added automatically by the system	This may include specific key-value properties related to various types of processing.
	Privacy Level	Privacy level zone as defined by the data management framework. Includes elements such as: • Shared List • Access Modifier	• Private: access only to creator user, overrides any other setting • Shared Organisations: shared with all users of specified organisations, overrides public • Shared Users: access only to selected users, overrides other properties besides private modifier • Public: access to anyone
	Keywords	A list of keywords that can be used for user queries.	

Table 3: SLICES Metadata Properties (based on Dublin Core)

To support the above properties, a list of other standards should be utilised, such as the ones listed to ensure proper exchange using standardised codes. A number of these standards are listed below:

- ISO 3166: The set of codes for the representation of names of countries.
- ISO639-3: The three-letter alphabetic codes for the representation of names of languages.
- ISO 8601-1: Representations for information interchange for date and time.
- DCMI-Period: DCMI Period Encoding Scheme.
- DCMI-Point: DCMI Point Encoding Scheme.

In conclusion, the proposed metadata format, can be implemented in the data management tier of the SLICES infrastructure and enable data reuse, but also support interoperability with other infrastructures/systems to maximise impact. However, the implementation of this layer comes at a cost that should be evaluated when all infrastructure requirements are finalised. An alternative scheme is to use existing platforms for research data management, such as DSpace, CKAN, Zenodo and Figshare. An analysis of these platforms is performed in Section 3.5, where the desired features for metadata management and information sharing are drawn. The SLICES consortium will need to take into account the identified stakeholders' requirements, cost and maintenance factors and decide on the best course of action at the end of the project.

3.5 Intra/Inter-operability

It is of prime importance to efficiently and effectively manage metadata information, both for consistency within the SLICES infrastructure (intraoperability between nodes) and collaboration with other infrastructures/systems (interoperability). In Section 3.4, we have defined a comprehensive metadata format, which is based on well-known and mature standards, to improve data re-use and visibility, and enhance intra/inter-operability; however, appropriate systems need to be deployed to guarantee both.

Intra-operability issues may arise when there are inconsistencies in the data (e.g. because of human error) or incorrect synchronisation between each node and cloud master data management modules. This is more likely to happen in the following cases:

1. There is lack of synchronisation between any node and the cloud master data management
2. A node allows a user to input incorrect data
3. The metadata model is updated and the current records' metadata cannot be correctly transformed to the new model.

Issues 1 and 2 can be easily addressed by ensuring the updates to the master data model/format will be performed at scheduled maintenance date/times, ensuring that there are no records in transit at the time. Issue 3 is more complicated as incompatibilities between models may require complex data processing mechanisms to be applied. However, this can also be avoided by proactively planning the metadata model upgrades and allocating sufficient time for testing and validation.

Interoperability has been partially addressed in Sections 3.3 and 3.4, but can be further strengthened by implementing tools that allow users to efficiently create/transfer the data, search and manage data based on the metadata information, and download data. To this end, we have investigated a number of platforms that support research data management, focusing on features such as API functionality, query and retrieval mechanisms, and supported communities. The platforms are described below:

Figshare[37]	• Allows the discovery, citing, sharing and uploading of research output. • Allows for uploading up to 5GB single file of any format + 20GB of free private space. Provides a desktop uploader application. • Generates a DOI for the researcher's work/ allows for reserving a DOI before releasing the data. • Allows for collaboration through collaborative spaces/ private link sharing. • Enables the customisation of showcase portals for presenting public research outputs. • Provides reporting and statistics information at various levels, e.g. researcher or object. • Provides a REST API for automating research workflows.
Dataverse[38]	• Provides counts for web visibility, academic credit and citations. • Accepts citations for datasets and files, such as EndNote XML, RIS Format, or BibTeX. • Can expose data to other systems using a variety of metadata formats. • Provides REST APIs such as: Search API, Data Deposit API, Data Access API, Metrics API, etc. • Is discoverable by Google with adherence to Schema.ord JSON-LD. • Provides login using institutional providers or GitHub, Google etc. • Supports a Data Explorer tool for preview and analysis of tabular files.

[37] https://figshare.com/
[38] https://dataverse.org/

	• Supports file downloads of tabular data in a variety of formats, such as TSV, RData. • Provides dataset versioning capabilities and file access control. • Allows integration with Dropbox for retrieving already uploaded files. • Can operate using a filesystem or object storage. • Is able to pull header metadata from Astronomy FITS files.
Mendeley Data[39]	• Enables researchers to control the full lifecycle of research data, i.e. collection and discovery. • Enables the creation of projects where researchers can collaborate/ share and annotate their data. • Integrates with Dropbox, Google Drive, Box and Azure for retrieving already uploaded files. • Allows for institutions to retain data on their own servers. • Published dataset metadata are aggerated to DataCite's metadata index and to the OpenAIRE portal. • Supports the harvesting of public dataset records using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard. • Data is stored on Amazon's S3 servers and archived with Data Archiving and Network Services (DANS).
Open Science Framework[40]	• A collaboration tool/ workflow system that enables researchers to collaborate on projects and publish findings for dissemination. • Provides a centralised repository for project files, data and code, with version control capabilities. • Supports file access control for selecting which parts of a project are public or private to the team. • Enables the team to keep logs, notes and track their progress. • Offers project analytics for measuring the impact of the project using citation, downloads and project access count. • Integrates with Dropbox, GitHub or Figshare. • Can publish reports in Google Scholar, Crossref and ORCID. • Supports search integration with platforms such as Mendeley, DataCite and ZOTERO.
Zenodo[41]	• Accepts research outputs of all research fields and all file formats. • Assigns publicly available DOI to all works and supports harvesting of all content via the OAI-PMH protocol. DOI is available before publishing. • Integrates with OpenAIRE. • Enables uploading with a variety of licenses and access levels. • Citation data is sent to DataCite. • Statistics enable the tracking of visits, visitor type, country and referrer domain. • Includes citation sources such as: NASA Astrophysics Data System, DataCite and Crossref.
Code Ocean[42]	• Facilitates the creation, organisation and dissemination of computational research in a collaborative manner. • Standardises research workflows and reproduction of computational discoveries.

[39] https://data.mendeley.com/
[40] https://osf.io/
[41] https://zenodo.org/
[42] https://codeocean.com/

	• Provides reproducible Capsules, which are entities that contain code, data, environment setup and any associated results, while also being versioned. • Integrates with data tools such as RStudio, Jupyter, Shiny Terminal and Git. • Allows access to any size/ type of data – can generate docker environments. • Allows researchers to create and share results in easy-to-use web analytic apps. • Allows researchers to utilise public capsules from a repository that provides outputs from the global community. • Deployed on the AWS cloud with a dedicated VPC.
4TU.ResearchData[43]	• International data repository for science, engineering and design. • Enables the curation, long-term access and preservation of research datasets. • The empowering technology of this platform is Figshare, while all data is hosted by the TU Delft Library. • DataverseNL enables researchers to create project spaces. • Project spaces provide data file management, metadata and documentations, version control, collaboration tools, storage and backup. • Every dataset is provided with a DOI for linking or citing the dataset in publications. • A DOI can be reserved prior to dataset publication. • Allows researchers to find and reuse a large number of published datasets through a digital library. • Uses OPeNDAP, which is a protocol that allows the use of data from a server without the need of downloading the data files.
ANDS[44]	• The Australian National Data Service, which helps Australian researchers to publish, discover and access research outputs. • Enables access to data coming from more than 100 Australian research organisations. • The platform does not hold the actual data, but descriptors to these data. The original data is hosted by the publishing partners and contributors. • Institutions have to provide their own metadata to the Research Data Australia registry. • The four main services provided are the research data discovery portal, DOI Service (DataCite), handle service and the research vocabularies service. • Other tools include connecting and linking data, assistance with harvesting, through the utilisation of various protocols and schemas. • Data information are syndicated to global data citation indexing systems.
Dryad Digital Repository[45]	• Completely open source with code available on GitHub and is based on Stash, which is a data publication platform. • The preservation of data is up to the underlying repository with which Dryad is integrated. • Each dataset has its own landing page that presents all descriptive and administrative metadata – can be also downloaded as PDF.

[43] https://data.4tu.nl/info/en/
[44] https://www.ands.org.au/
[45] https://datadryad.org/stash

	• Integrates with any SWORD/OAI-PMH-compliant repository. • Supports DataCite and can also be configured to support other schemas. • DOIs are provided to all datasets. • Easy sign-in can be enabled via institutional providers and ORCID. • Supports Drag and Drop uploads with simple navigation. • Allows search by subject, filetype, keywords, campus, location, etc. • Allows for building relationships between datasets from other publications.
Re3data [46]	• Global data registry repository offering data from a vast range of academic disciplines. • Provided by DataCite as a service. • Provides access to datasets to researchers, publishers, funding bodies and scholarly institutions. • Indexes more than 2450 research data repositories. • Allows developers to access information via a REST API.

Table 4: Features of Platforms for Research Data Management

Our analysis revealed that the following features are necessary to maximise interoperability:

- **FAIR principles:** implementation should support FAIR principles (discussed in the next section).
- **APIs**: development of APIs (e.g. REST) for exposing metadata information to end-users.
- **Repository Interoperability**: support at least one metadata harvesting protocol (e.g. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[47]) to streamline data dissemination.
- **Flexible Referencing:** develop a flexible engine for transforming the implemented metadata format to the majority of well-known formats.
- **Serialisation:** support XML, JSON and YAML for data serialisation, while being extendible to support other formats in the future.
- **Complex Querying:** support both simple and advanced queries that allow users to use free-text search, combine filters and drill down to the individual components of each metadata property.
- **User Experience:** develop a user friendly, intuitive UI for designing and enhancing the user experience according to user experience design (UXD) principles. The UI must seamlessly offer the functionality of the underlying infrastructure in an efficient and pleasant manner.

We found that amongst the most predominant platforms that support research data management is Figshare and Zenodo. They both support the aforementioned features and more. An important characteristic of Zenodo is the fact that it is hosted by CERN, which increases the probability for long term sustainability. A thorough comparison of these platforms can be found in[48], according to which the only additional feature that Zenodo supports is the ability to upload from Github, the predominant code hosting platform for version control and collaboration.

In conclusion, **SLICES can design its own dedicated research data management platform or use an existing research data management platform, such as Zenodo or Figshare**. The decision needs to take into account the cost of developing the platform and the final requirements from all stakeholders. If the decision is to implement a new platform, then the aforementioned objectives need to be fully met, so that the solution is more easily adopted by the community.

[46] https://www.re3data.org/

[47] The Open Archives Initiative Protocol for Metadata Harvesting, https://openarchives.org/OAI/openarchivesprotocol.html [Last accessed 01 February 2021]

[48] Mislan, K.A.S. & Heer, Jeffrey & White, Ethan. (2015). Elevating The Status of Code in Ecology. Trends in Ecology & Evolution. 31. 10.1016/j.tree.2015.11.006.

3.6 Analytics

Data analysis is a vital requirement for any organisation that wants to allow users to gain valuable insights and discover hidden patterns within their data. This becomes even more promising with projects such as SLICES that have the potential to empower collaboration using "larger", federated datasets to facilitate ground-breaking research. To accomplish this, there needs to be support for a plethora of data analysis tools that allow researchers and practitioners to obtain insights and unveil complex patterns that are hidden within the data, in order to make informed decisions. In the majority of tools, the data analysis process begins by integrating one or more data sources form the dataset that will be used as the basis for analysis. In most cases, this also includes some cleaning and pre-processing steps, according to a specific data model that will be utilised. Platforms then allow users to utilise a comprehensive collection of statistical methods and techniques (e.g. algorithms for predictive modelling), as well as the means to visualise them using different types of visualisations that can be customised in every aspect, with the objective to "tell a story".

SLICES aims to integrate a variety of open-source and commercial tools to support data analysis for different application domains, such as data mining, machine learning and artificial intelligence. We use the four types of analytics mentioned in[49], to provide a rough categorisation of the analytics that should be supported by SLICES:

- **Descriptive Analytics:** focus on understanding what has happened and what is happening currently in the data. This category uses standard statistical descriptors and visualisations to describe properties of the data, such as distribution, dispersion, mode, etc.
- **Diagnostic Analytics:** focus on understanding why something happened. To accomplish this, they use descriptive analytics to distinguish irregularities.
- **Predictive Analytics:** focus on understanding what will happen by observing the trends and patterns in historical information. Some of the most popular algorithms included in this category are: Regressions, Forecasting, Classification, Clustering, Association, Outlier Analysis, Text Mining.
- **Prescriptive Analytics:** focus on selecting the best course of action when multiple solutions are available, based on resource optimisation. Some of the most popular algorithms included in this category are: Optimisation, Multiple-Criteria Decision Analysis, Simulation.

In general, all algorithm families can be used for various parts of data management and analysis. For example, in some cases, a predictive model can be used to impute missing values for an attribute during data pre-processing, while in others, to predict the class of a new entry. As such, the best course of action is to develop an **algorithm repository,** which will be accessible by various software components of the infrastructure.

3.7 Other Data Management Issues

This section describes important, but not required, objectives that should be provided by the data management infrastructure.

3.7.1 Naming Conventions

Naming consistency is important for efficiently locating a resource and understanding its use. However, it is up to the creator to provide proper names for research outputs and related data files. SLICES will offer Naming Conventions/Guidelines to improve the structure/consistency of files. The draft guidelines include the following recommendations:

- **Name Length:** Provide a maximum length for file names, as long filenames may not be interoperable with some systems.

[49] The 4 Types of Data Analytics, KDNuggets Gold Blog, https://www.kdnuggets.com/2017/07/4-types-data-analytics.html [Last accessed 03 February 2021]

- **Date Format:** Allow the display of dates in a chronological order, even over the span of many years; use the YYYYMMDD format.
- **Leading Zeros:** Used to make ascending order of numbers correspond with alphabetical order.
- **Naming Scheme:** Use a consistent naming scheme throughout; do not use spaces or punctuation symbols as these may not be interoperable with some systems.
- **Order:** the naming scheme should provide easy distinction between different file groups and also provide which elements should go first.

3.7.2 File Organisation

File organisation is important for efficiently locating a resource, even in the cases where there is no predefined structure available. SLICES will offer File Organisation Guidelines to improve the consistency of the structure of data. The draft guidelines include the following recommendations:

- **Hierarchical Structure**: adopt a hierarchical structure that includes at least the following folders:
  - **Data**: includes all input data, in the cases where data is not related to experiments
  - **Experiments**: includes a folder for each experiment (e.g. **exp01, exp02**). Each experiment folder should include at least the following folders:
    - **input data**: contains all data required for the experiment
    - **software**: contains all software components, models for the experiment or appropriate links
    - **deployment**: contains the steps on how to conduct the experiment
    - **output data**: contains the results of the experiment
  - **Relationships**: contains references to relationships with any other research data, according to the metadata specification (see Section 3.4)
  - **Dissemination**: contains any material related to dissemination, such as presentations, press releases, articles, etc.
  - **Miscellaneous**: contains any other material
- **Folder Naming:** use the naming scheme provided in Section 3.7.1

3.7.3 Data Storage

SLICES aims to design and develop a distributed research infrastructure with next generation capabilities that will host thousands of users and their data. Our preliminary estimations include up to 5,000 users and their data, accounting for up to 50GB per user on the individual nodes and up to 1TB on the cloud. This provides us with a preliminary estimation of 0.25PB-1PB data storage for all datacenters residing on SLICES nodes and 5PB for the cloud-based datacenter.

3.8 Resource Allocation

To ensure the sustainability of the infrastructure, some constraints will be placed on registered users, such as the overall size of research data stored. Costs related to the opening of data will be covered by the owner. The resources required for long-term preservation are analysed in D3.2 – Cost analysis. The infrastructure will support long-term data retention according to a suitable policy (e.g. for the duration of the project).

4 Alignment with FAIR Data Principles

With the advancement of technology and the plethora of electronic data being generated and available online, there is a need to ensure the longevity of such data as well as its access to the wider research community. Wide access to scientific data can facilitate further knowledge discovery and research transparency. Moreover, with the rapid expansion of the digital ecosystem, the use of machines to process the vast volume of available data is crucial, as humans cannot efficiently and effectively perform the relevant data processing (e.g. find, access, reuse), without additional computational support.
It is with the above in mind that the FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles were developed. The FAIR principles are intended to be used as guidelines for data producers and publishers, with regards to data management and stewardship. One important aspect that differentiates FAIR from any other related initiatives is that they move beyond the traditional data and they place specific emphasis on automatic computation, thus considering both human-driven and machine-driven data activities. Since their publication, FAIR became widely accepted and used.

To this end, SLICES wants to fully endorse and adopt the FAIR principles, acting as a catalyst to enable and foster the data-driven science and scientific data-sharing in this area. In what follows, we provide how SLICES adheres to the FAIR guiding principles using the categorisation provided in[50].

Code	Principle	SLICES adherence to principle
Findable		
F1	(meta)data are assigned a globally unique and persistent identifier	The data will be assigned a Digital Object Identifier (DOI) at the time of upload.
F2	data are described with rich metadata (defined by R1 below)	Metadata will be provided in a consistent format with appropriate properties based on an enhanced version of the well-established format DublinCore. Furthermore, automatic metadata generation, e.g. using Machine Learning techniques, should also be used to provide even more reusability and scalability for interacting with other infrastructures and external digital libraries and their collections.
F3	metadata clearly and explicitly include the identifier of the data they describe	The data will be assigned a Digital Object Identifier (DOI) at the time of upload. The metadata will include the identifier in a dedicated Property.
F4	(meta)data are registered or indexed in a searchable resource	A dedicated resource discovery component will allow for searching data resources through the metadata and keywords (assigned by the user or automatically generated by the platform). The Discovery component will support simple and advanced queries that allow users to use free-text search, combine filters and drill down to the individual components of each metadata property.

[50] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Accessible		
A1	(meta)data are retrievable by their identifier using a standardised communications protocol	Metadata are retrievable using at least one metadata harvesting protocol (e.g. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)) to streamline data dissemination. Support for REST APIs with XML, JSON and YAML capabilities for metadata (while being extendible to support other formats in the future).
A1.1	the protocol is open, free and universally implementable	OAI-PMH and REST are open, free and universally implementable, as well as supported by the vast majority of applications.
A1.2	the protocol allows for an authentication and authorisation procedure, where necessary	Authentication and authorisation procedures will be implemented for data that are not openly available. Metadata will be open and will not require authentication and authorisation procedures.
A2	metadata are accessible, even when the data are no longer available	Metadata will be retained for the duration of the project and the lifetime of the infrastructure. In case a research data management platform is used, the lifetime will be connected with the lifetime of the host repository. Metadata will be stored in a dedicated data store.
Interoperable		
I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	Metadata will be provided in a consistent format with appropriate properties based on an enhanced version of the well-established format Dublin Core. Dublin Core has been formally standardised as ISO 15836, ANSI/NISO Z39.85 and IETF RFC 5013. The metadata will be accessible through the REST APIs, which will export XML, JSON and YAML) and also will have the ability to translate the metadata format to others based on a mapping engine.
I2	(meta)data use vocabularies that follow FAIR principles	The following vocabularies/standards will be utilised: •ISO3166: The set of codes for the representation of names of countries. •ISO639-3: The three-letter alphabetic codes for the representation of names of languages. •ISO 8601-1: Representations for information interchange for date and time.

		•DCMI-Period: DCMI Period Encoding Scheme. •DCMI-Point: DCMI Point Encoding Scheme. We also anticipate more standards to be adopted when the final SLICES design is available.
I3	(meta)data include qualified references to other (meta)data	This information will be provided by the defined metadata property ("Relation") and its sub-properties.
Reusable		
R1	meta(data) are richly described with a plurality of accurate and relevant attributes	Metadata will be provided in a consistent format with appropriate properties based on an enhanced version of the well-established format DublinCore. Furthermore, automatic keyword generation, e.g. using Machine Learning techniques, will be used to enhance discovery.
R1.1	(meta)data are released with a clear and accessible data usage license	Rights-License is part of the DublinCore properties. A list of licenses and their definitions will be provided to the user to select from when constructing their metadata. In cases where a license does not exist, Public Domain should be provided as a default option, but the option to select Other should also be available. Data downloaded by other users will be subject to the specified license.
R1.2	(meta)data are associated with detailed provenance	Provenance is part of the DublinCore properties and requires the user to provide a statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity and interpretation.
R1.3	(meta)data meet domain-relevant community standards	Dublin Core is one of the most widely used standards (ranks 1st in re3data) and has been formally standardised as ISO 15836, ANSI/NISO Z39.85 and IETF RFC 5013. It is cross-domain and not domain-specific.

5 Compliance

This section will introduce, at a high level, the overall principles and measures that the project will follow in order to ensure compliance with the relevant ethical and personal data protection frameworks. Further details on compliance will be presented by Task 1.3 as part of deliverable D1.3 (M24).

5.1 Compliance with GDPR

The project will ensure compliance with the fundamental principles of data protection in Article 5 of the GDPR when designing the proposed innovations. More details on GDPR will be provided by Task 7.1 as part of in deliverable D7.1.

5.1.1 Lawfulness, fairness and transparency

Pursuant to Article 5(1) (a) of the GDPR, personal data shall be "processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness, transparency')". For the processing of personal data to be **lawful**, the partners will identify specific grounds for the processing of personal data carried out within the project. Partners shall ensure that the processing is carried out on a 'lawful basis', pursuant to one of the six options set forth by Article 6 of the GDPR. The choice of one of these legal bases for the processing will depend on the project's purposes and the specific relationship with the individuals involved. Since special categories of personal data and data relating to criminal convictions and offences will not be processed during the project, there will be no more specific additional conditions for processing, pursuant to articles 9 and 10 of the GDPR. Appropriate consents for the lawfulness of the generated/submitted data have been proposed in Section 3.4 to be included in the metadata of the research data.

Fairness means that partners shall ensure that the processing of personal data must always be fair as well as lawful. If any aspect of the project's processing is unfair, it will be in breach of this principle – even if it is shown that a lawful basis for the processing has been identified.

In general, fairness means that the project should only handle personal data in ways that users would reasonably expect and not use it in ways that have unjustified adverse effects on them. Assessing whether the project is processing personal data fairly depends partly on how it obtained them. If users are deceived or misled when the personal data is obtained, then this is unlikely to be fair. Personal data may sometimes be used in a way that negatively affects an individual without this necessarily being unfair. What matters is whether such detriment is justified.

The project should also ensure that it treats individuals fairly when they seek to exercise their rights over their data. This ties in with the obligation to facilitate the exercise of individuals' rights.

Transparency is fundamentally linked to fairness. Transparent processing is about being clear, open and honest with individuals from the start about the description of the project and how and why their personal data are processed. If individuals know at the outset how the project will use their personal data, they will be able to make an informed decision about whether to enter into a relationship, or perhaps to try to renegotiate the terms of that relationship.

Transparency is important even when there is no direct relationship with the individual and their personal data are collected from another source. In some cases, it can be even more important, as individuals may have no idea of who is collecting and using their personal data, and this affects their ability to assert their rights over their data (so called 'invisible processing').

The partners must ensure that the project informs individuals that the processing of their personal data is carried out in a way that is easily accessible and easy to understand using clear and plain

language. The project will also ensure that there is transparency and accountability in the way that the users are able to access their data through appropriate user interfaces.

5.1.2 Purpose Limitation

Before collecting any personal data, partners will define the purposes of the processing operations in a sufficiently specified manner. They will also adopt processes and measures to ensure that data of SLICES users are only processed for such purposes. The collected data will not be used for commercial purposes.

5.1.3 Data Minimisation

Partners shall avoid and prevent any unnecessary collection and use of personal data in SLICES. **In case personal data would be nevertheless accessed, the research project will minimise such processing and will strictly abide to the GDPR and to the ePrivacy Directive (and corresponding regulation once it has come into effect).** Partners will develop applications, tools and services that are grounded on a **"privacy by design" approach**. Personal data protection and the operationalisation of the data protection by design and by default requirement in the GDPR is part of the key project requirements:

- The project will comply with the principle of **data minimisation**, by limiting the collection and/or storage of sensitive data to the extent that it is necessary and will not store any personal data for a longer period than needed.
- Wherever applicable, **personal data will be parsed and anonymised,** and personal identifiers will be either hashed or randomly generated.
-

5.1.4 Data Accuracy

The project will implement appropriate processes to record the source of the personal data and provide data subjects with the possibility to rectify their data.

5.1.5 Storage Limitation

Personal data will not be stored for longer than it is necessary for the defined processing purposes. Regarding the data retention period, the project plans to differentiate among different categories of data (personal and non-personal, sensitive data):

1. **Special categories of personal data**, as defined by the GDPR. Such data shall not be collected nor processed by the project.
2. **Personal data directly controlled by the users**, such as identifiers, passwords and optional email addresses that can be modified and deleted by the users will be stored as long as the related users want to keep them.
3. **(Regular) personal data** will be minimised and managed according to the GDPR. The data retention period will be minimised and specified on a case per case basis by assessing the necessity and interest to process and store such data.
4. **Fully anonymised data** with no means (once anonymised) to identify the person behind such data fall outside the scope of the GDPR. As stated by the GDPR in Recital 26 *"This Regulation does not […] concern the processing of such anonymous information, including for statistical or research purposes."* In the absence of any risk for personal data protection, such data can be stored as long as they are relevant for research purpose.
5. **Non-personal data that are non-critical,** with no relation to any physical person. In the absence of any risk for personal data protection, such data can be stored as long as they are relevant for the research project and scientific audits.

6. **Non-personal data that are critical data** are data that may have an impact on security or may negatively impact the exploitation potential (of the consortium as a whole or of any individual members of the consortium). Such data will be kept confidential and will not be released to the public.
7. **Server logs** collected for statistical and security reasons will be kept for 3 full years following the year of collection.

5.1.6 Integrity and Confidentiality

Partners may use data processors for certain processing operations (e.g. storage). Should this be the case, they will diligently select processors that adopt adequate security measures and ensure that SLICES data will be stored in servers in the EEA.

5.1.7 Accountability

According to the accountability principle, controllers shall be responsible for and be able to demonstrate compliance with all principles. Accountability requires the allocation of **clear responsibilities for the processing**. It is a must to define the main roles in each processing activity, as different requirements may apply to each under GDPR. It is not always easy to identify who acts as controller, who acts as processor, situations of joint controllership or separate controllership. Allocation of responsibilities may be particularly challenging in the case of a project like SLICES, involving several partners under a consortium agreement and external stakeholders.

During the first months of the project, legal and technical partners will work together to map data flows and respective responsibilities in the processing.

Lawful basis for data processing in SLICES - consent

The sharing of data in SLICES will be entirely voluntary. Partners will develop a dissemination plan and encourage citizens' and companies' involvement in SLICES by pointing out the benefits of an extended privacy-by-design data marketplace. However, the ultimate decision lies with the users, who will be called to consent to the processing. The project will abide to the **"Prior informed consent"** rule by, among other things:

- Providing clear and transparent information (all the information required under Arts. 13 and 14 –where relevant, for personal data obtained indirectly- of the GDPR will be provided). The website will be fully transparent regarding the data processor and data controller.
- Permitting the users to choose what information they want to provide and for what purpose (consent requests will be granular and specific).
- Requesting consent via a clear affirmative action (opt-in).

Participating end-users will be duly informed ahead of their acceptance to participate in SLICES about the personal data protection policy of the project. This information will be provided in a layered, readable and user-friendly mode, and it will be accessible at any time from the website. Legal partners will draft **informed consent forms** and **data protection information sheets,** and communicate those to the European Commission. Proactive actions will be undertaken to guarantee that the users fully understand and give their consent to the data protection policy of the project. Users will be granted the right to withdraw their consent to the processing at any time, in a simple and effective way.

5.2 Data Management Compliance with GDPR

The GDPR gives data subjects an array of rights they can use vis-a-vis data controllers with a view to keep control over their personal data. These include the rights to information, access, rectification, erasure, restriction of processing, objection, data portability and rights in relation to automated decision-making.

During the project's duration, the DPO will act as a central point for responding to any requests from data subjects. With input from the DPO, partners will also put in place processes (including organisational and state-of-the-art technical measures) enabling the project to effectively manage the rights of individuals.

Documentation and involvement of processors and sub-processors

In line with Article 30 of the GDPR, project partners will **keep written records** of all processing activities involving personal data undertaken during the development and testing of SLICES. The records will include the information required under Article 30 paragraphs (1) *(for controllers*) and (2) *(for processors)*. Preferably, partners will make use of templates prepared by national DPAs to meet the GDPR documentation requirements.

Whenever the project/any consortium partner **uses processors or sub-processors** to carry out any data processing operations related to SLICES, a Data Processing Agreement will be signed in accordance with Article 28(2) of the GDPR.

Territoriality and normative scope

The research project is to take place in EU-member states and/or in associated countries with equivalent level of privacy protection. The data will be stored in secured servers in Europe and no personal data will be transferred abroad or stored in non-European based cloud infrastructure.

Moreover, the European personal data protection norms and the ethical standards and guidelines of Horizon2020 will be rigorously applied, regardless of the country in which the users are located. The project will voluntarily and universally apply and respect the European norms and standards regardless of the territorial location of the users. The project will abide by all applicable and any future EU and national legislations, as well as by the relevant international guidelines, and it will be conducted in accordance with the Declaration of Helsinki.

Rights of the Data Subjects

Effective systems should be put in place in SLICES to safeguard the rights of data subjects. The systems should allow both content owners and data subjects whose personal data are utilised within SLICES to manage their data using appropriate procedures that clearly indicate how and when the rights will be exercised. The proposed systems will ensure the rights of the user in the ways presented below. Note that this is a minimal set of functions that should be provided, and we expect this to be enhanced when the final design of the systems is available:

- **The Right to Be Informed:** the system will provide individuals with information including how SLICES processes their personal data, the personal data retention period, how it may be shared and with which entities/processors it may be shared.
- **The Right of Access:** the system will provide registered users with the ability to easily access their data using dedicated navigation menus/options, using a clear and plain language.
- **The Right to Data Portability**: the system will allow the registered users to download their data in common data interoperability formats, such as the ones described in Section 3.5.
- **The Right to Rectification**: the system will provide different functions to allow users to rectify their data. In the case of registered users, there are two high level types of data: the profile information and the research data that have been uploaded. The profile information will be accessible and updateable through the platform. For other data uploaded (e.g. research data), changes will be allowed, e.g. through erasure or updating via versioning.
- **The Right to Erasure**: the system will provide different functions to allow users to erase their data, as long as these satisfy the requirements of the data retention periods.
- **The Right to Restrict Processing**: the system will provide different functions to allow users to restrict processing of their personal data, as well as resources that contain personal data (e.g. datasets), and appropriate consents that have been provided in the metadata.

- **The Right to Object:** the system will allow users to refuse to have their personal data processed by the system under certain circumstances, e.g. authorship disputes.
- **Rights Related to Automated Decision-Making and Profiling**: the system will adopt appropriate data minimisation and anonymisation techniques so that automated decision-making and profiling will only be performed using summarised and anonymised data.

Furthermore, to cater for any additional requests of the data subjects, the system will offer dedicated UI components to allow for asking questions to the DPO and receiving appropriate answers. These functions will be available within the portal and they will allow users to exercise their rights.

5.3 Compliance with National Regulations

Countries or institutions may impose additional regulations to ensure additional properties on research data. For example, ethics are considered an integral part of research for activities funded by the European Union. In several countries, which are also part of this consortium, there are committees that impose additional regulations to data collected from research activities, such as the Cyprus National Bioethics Committee, the French Comites de Protection des Personnes, the Medical Research Ethics Committee in the Netherlands, and the Swedish Ethical Vetting Board[51]. The research data uploaded or generated within SLICES should comply with the national and institutional regulations and should be supplemented with evidence of potential approvals required by formal bodies.

[51] European Conference of National Ethics Committees (COMETH), https://www.coe.int/t/dg3/healthbioethic/cometh/national_ethics_committees/ [Last accessed 20 February 2021]

6 Data Security and Protection of Personal Data

This section will introduce, at a high level, the overall principles and measures that the project will follow in order to ensure data security policies are in place. Further details on compliance will be presented by Task 3.2 as part of deliverable D1.3 (M24).

The enforcement of security and trust management policies, and also Acceptable Use polices, permits a good and secure deployment of the research infrastructure and linked applications. The usage of the research infrastructure and the related tools and applications must follow the policies defined to ensure the security and the protection of personal data. The tailoring of such polices will be done in the SLICES-DS project in Task T3.2 and in particular, in deliverable D3.6 “Data Protection Policies” (to be published in M24). A strong focus will be given to the protection of personal data encountered notably in the datasets of the experiments done in the SLICES research infrastructure.

Nevertheless, some principles can already be mentioned in the Data Management Plan and will be presented in more in detail in the relevant deliverable D3.6. The security and trust management policies will describe all the processes that correctly enforce the security and protection of data, specifically, the personal data provided to the research infrastructure providers or managers. In this context, the different entities managing the components of the research infrastructure, typically the testbed providers, should be trustable by the end-users, namely the researchers. Some security and technical mechanisms will be put in place in order to make the testbed providers trustable, and will be explained in the security and trust management policies, including how to enforce them correctly. For example, the creation of certificates by each node of the SLICES research infrastructure can be used to ensure the trustworthiness of the distributed nodes. Only valid and recognised certificates will allow the exchange of data between the different nodes and also the researchers. Of course, the data in transit will be encrypted to avoid any data leakage or interception during data transmissions. Secure protocols will be used between the distributed nodes and the manner to utilise them will be documented to avoid any vulnerabilities in data exchanges. The definition of such technical measures to enforce the security and protection of data will be mentioned in the guidelines concerning the data protection by design. The different policies elaborated in Task T3.2 will follow the standards and best practices concerning the security and protection of personal data.

The Acceptable Use policy defined in the SLICES-DS project will describe the rules and guidelines on how to use the SLICES research infrastructure in conformity with the laws, and in particular, privacy regulations like the GDPR. This policy will establish the consequences of a violation of the rules written in the policies, like the suspension of an account or the legal action to be undertaken if required.

7 Ethical Aspects

All activities in SLICES will be carried out in compliance with ethical principles, standards and guidelines, as well as any applicable international, EU and national law. Specifically, SLICES-DS will abide by the ***All European Academies (ALLEA) European Code of Conduct for Research Integrity*** **principles of reliability, honesty, respect and accountability**[52]**. Specifically, the Code of Conduct emphasises that its** purpose is to help realise the basic responsibility of the research community, which is to formulate the principles of research, define the criteria for proper research behaviour, maximise the quality and robustness of research, and respond adequately to threats to or violations of research integrity. In doing so, the Code of Conduct recognises that Interpretation of the values and principles that regulate research may be affected by social, political or technological developments and by changes in the research environment.

The interpretations of the four ethical principles considered by the ALLEA European Code of Conduct are described next. **Reliability** is about the consistency of the research methodology and the analysis, as well as the credibility of the resources. **Honesty** is about developing, undertaking, reviewing, reporting and communicating research in a transparent, fair, full and unbiased way**. Respect** is about attitudes and behaviours towards colleagues, research participants, society, ecosystems, cultural heritage and the environment**.** Finally, the principle of **Accountability** for the research refers to the journey from idea to publication, its management and organisation, training, supervision and mentoring, and its wider impacts.

Careful attention will be given to ensuring that the data collection and data management processes respect the abovementioned principles, in alignment with applicable law. The processes, methods and techniques for data collection and data management will respect the project's requirements, while also applying the utmost consideration for research subjects, study participants and all stakeholders involved. The project partners will take all measures necessary to refrain from practicing any form of plagiarism, data falsification or fabrication.

As far as data collection from human participants is concerned, SLICES will always inform participants about what data will be used, who will have access to the data, in what format the data will be accessed, which data protection rights apply to the data, and how long the data will be kept for. Any such activity will be coupled with the collection of voluntary informed consents from all participants, including consent for long-term storage of data or archiving. Privacy will be respected in this process and all personal data will be protected according to GDPR, national regulations, institutional regulations and data management standards as outlined in Section 7, Compliance. Protection of personal data is further elaborated in Section 8, Data Security and Protection of Personal Data. Moreover, data security and protection will reinforce data quality in SLICES.

Finally, consideration of ethical aspects must include the identification of ethical risks that may emerge from data collection and data management, including any data processing that needs to occur. These risks fall under different categories, according to the European Commission's Ethics and Data Protection report[53], a document that aims to raise awareness in the scientific community with regards to these issues. Risks related to different types of personal data may include data about racial or ethnic origin, political opinions, religious or philosophical beliefs, genetic, biometric or health data, sex life or sexual orientation, and data about trade union membership. Risks with regards to specific data subjects may include data about children, vulnerable people or people who have not given their explicit consent to participate in a project. There are also risks regarding: (i) the scale or complexity of data processing, e.g. when we have large-scale processing of personal data; (ii) the data collection or processing technique used, e.g. when privacy-invasive methods are used or when artificial intelligence

[52] The European Code of Conduct for Research Integrity, (2017), https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf [Last Accessed on 09 February 2021]

[53] Ethics and data protection, European Commission, https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection_0.pdf [Last Accessed on 09 February 2021]

is used to analyse personal data; (iii) the involvement on non-EU countries, e.g. when collection of personal data is done outside of the EU or when there is transfer of personal data to non-EU countries. SLICES will consider all indicators of data collection and management (including data processing) operations that may entail higher ethics risks. When such high risks are identified, a detailed analysis of any issues raised must take place, which must cover the following aspects, in alignment with EU's Ethics and Data Protection report: (i) an overview of all planned data collection and processing operations; (ii) identification and analysis of the ethics issues that these raises; and (iii) an explanation of how these issues will be mitigated in practice. The analysis will be included in the research protocol and any relevant documentation for ethics approvals. Following ethics approval, records documenting informed consent procedures must be kept, so that these are available if requested by data subjects, funding agencies or data protection supervisory authorities.

8 Data Management Plan Summary

The following table provides a synopsis of key issues of the DMP following the H2020 DMP template[54].

DMP component	Issues to be addressed
1. Data Summary	**State the purpose of the data collection/generation** This information will be provided by the defined metadata fields (Title and Description). More information can be found in Section 3.4 -Metadata Management. **Explain the relation to the objectives of the project** The mission of SLICES is to provide the research and engineering community with a fully controllable, programmable virtualised digital infrastructure test platform. It aims to answer the fundamental questions regarding digital infrastructures in an evolving environment, enable new technologies to support the vision (5G and beyond), support ICT breakthrough discoveries with the use of OTS and ad-hoc programmable technologies together with advanced design and execution cloud-based solutions. This initiative should act as a catalyst to enable and foster the data-driven science and scientific data-sharing in this area. Open research data should be considered together with the test platform and ultimately contribute to the deployment of a data repository where all data produced by the platform, under some policies, could be made available under the FAIR principle (Findable, Accessible, Interoperable, and Reusable). Based on the issues addressed in this section and the methodology and planned actions discussed within this document, the objectives of the project related to data collection, processing and availability are met. **Specify the types and formats of data generated/collected** This information will be provided by the defined metadata field (Format). More information can be found in Sections 3.4 - Metadata Management and 2.3 - Formats of Data. **Specify if existing data is being re-used (if any)** Existing data come in two different forms: (i) data utilised within a research project; or (ii) external data, such as a well-known online datasets, a standardised vocabulary, etc. For (i), it is up to the creator to: (a) include the external data as part of the data package and (b) specify appropriate relationships to external data using the metadata field Relation (see more in Section 3.4 - Metadata Management). **Specify the origin of the data** This information will be provided by the defined metadata field (Source). More information can be found in Section 3.4 - Metadata Management. **State the expected size of the data (if known)** According to the estimates presented in Section 2.5 - Expected Data Size, the data size provided by a single user, for all data uploads, has a limit of up to 50GB.

[54] TEMPLATE HORIZON 2020 DATA MANAGEMENT PLAN (DMP), https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf [Last accessed 15 January 2021]

	Outline the data utility: to whom will it be useful The data is useful to a number of stakeholders described in Section 2.1 - User Groups.
2. FAIR Data	
2.1. Making data findable, including provisions for metadata	**Outline the discoverability of data (metadata provision)** Metadata will be provided in a consistent format with appropriate properties based on an enhanced version of the well-established format DublinCore. Furthermore, automatic metadata generation, e.g. using Machine Learning techniques, should also be used to provide even more reusability and scalability for interacting with other infrastructures and external digital libraries and their collections. For more information, see Section 3.4 - Metadata Management. **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?** SLICES will support persistent identifiers (DOIs) for research data and other research outputs, either through custom metadata management software or the adoption of a research data management platform. For more information, see Section 3.4 - Metadata Management. **Outline naming conventions used** The naming conventions for files are outlined in Sections 3.7.1 - Naming Conventions and 3.7.2 - File Organisation. **Outline the approach towards search keywords** Metadata will be provided in a consistent format with appropriate properties based on an enhanced version of the well-established format DublinCore. Furthermore, automatic metadata generation, e.g. using Machine Learning techniques, should also be used to provide even more reusability and scalability for interacting with other infrastructures and external digital libraries and their collections. **Outline the approach for clear versioning** This information will be provided by the defined metadata field ("Has Version" and "Is Version Of"), in the cases where versioning is applicable. More information can be found in Section 3.4 - Metadata Management. **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how** Metadata will be provided in a consistent format with appropriate properties based on an enhanced version of the well-established format Dublin Core. Dublin Core has been formally standardised as ISO 15836, ANSI/NISO Z39.85 and IETF RFC 5013. The core properties are part of a larger set of DCMI Metadata Terms.
2.2 Making data openly accessible	**Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.** The creator will indicate the access level of the data during creation using a dedicated metadata property named Privacy Level, which includes the following access levels: • Private: access only to creator user, overrides any other setting • Shared Organisations: shared with all users of specified organisations, overrides public

	• Shared Users: access only to selected users, overrides other properties besides private modifier • Public: access to anyone More information can be found in Section 3.4 - Metadata Management. **Specify how the data will be made available.** The data will be made available via dedicated discovery functions that are described in Sections 3.4 and 3.5 and are outlined below: • Utilisation of established platforms for research data management, such as Figshare and Zenodo. • Design and development of a metadata discovery system fulfilling the interoperability objectives described in Section 3.5. **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?** In the case where SLICES will utilise established platforms for research data management, such as Figshare and Zenodo, then these already provide extensive documentation for their APIs. On the other hand, if the metadata discovery software will be implemented within the infrastructure, then extensive documentation will be provided to support the access process. **Specify where the data and associated metadata, documentation and code are deposited** The data and associated metadata, documentation and code will be deposited in the data infrastructure as described in Section 3.2. **Specify how access will be provided in case there are any restrictions** When sound reasons that restrict the openness of the research data (e.g. due to intellectual property rights, sensitive or personal information) exist, the creator will need to indicate this by setting the Privacy Level metadata property appropriately.
2.3. Making data interoperable	**Assess the interoperability of your data.** The data will be made available via dedicated discovery functions that are described in Sections 3.4 and 3.5 and are outlined below: • Utilisation of established platforms for research data management, such as Figshare and Zenodo. • Design and development of a metadata discovery system fulfilling the interoperability objectives described in Section 3.5. In either case, the following principles will be ensured: • Support for FAIR principles. • Development of appropriate APIs for exposing metadata information to end-users. • Support of at least one metadata harvesting protocol (e.g. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)) to streamline data dissemination. • Support of a flexible engine for transforming the implemented metadata format to the majority of well-known formats. • Support for XML, JSON and YAML for data serialisation, while being extendible to support other formats in the future.

	• Support of both simple and advanced queries that allow users to use free-text search, combine filters and drill down to the individual components of each metadata property. • Development of a user friendly, intuitive UI for enhancing the user experience by seamlessly offering the functionality of the underlying infrastructure in an efficient and pleasant manner. **Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.** Metadata will be provided in a consistent format with appropriate properties based on an enhanced version of the well-established format Dublin Core. Dublin Core has been formally standardised as ISO 15836, ANSI/NISO Z39.85 and IETF RFC 5013. Additional vocabularies/standards that will be utilised: • ISO 3166: The set of codes for the representation of names of countries. • ISO639-3: The three-letter alphabetic codes for the representation of names of languages. • ISO 8601-1: Representations for information interchange for date and time. • DCMI-Period: DCMI Period Encoding Scheme. • DCMI-Point: DCMI Point Encoding Scheme. We also anticipate more standards to be adopted when the final SLICES design is available. **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?** Both intra- and inter-operability will be facilitated by storing the metadata in a consistent format with appropriate properties based on an enhanced version of the well-established format Dublin Core. Intraoperability will be additionally enhanced by employing a synchronisation mechanism that ensures that the updates to the master data model/format will be propagated to the individual node modes. Interoperability will be further enhanced by supporting at least one metadata harvesting protocol (e.g. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)) to streamline data dissemination. Furthermore, a flexible translation engine will be developed to transform stored records to the majority of well-known metadata formats. Finally, the APIs will support various data exchange formats, such as XML, JSON and YAML for data serialisation, while being extendible to support other formats in the future.
2.4. Increase data re-use (through clarifying licenses)	**Specify how the data will be licensed to permit the widest reuse possible.** Intellectual property rights – e.g. the data may contain commercially sensitive information and opening it may impair the protection of IPR. A list of licenses and their definitions will be provided to users to select from, when constructing their metadata (Rights-License is part of the Dublin Core properties). In case a license does not exist, Public Domain should be provided as a default option, but the option to select Other should also be available. Data downloaded by other users will be subject to the specified license. **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed.**

	Datasets will be open and immediately available unless they are licensed under a particular scheme, as these are mentioned in Section 2.4. **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.** Datasets will be usable by third parties under appropriate licenses and access rights. This information will be provided by the defined metadata fields ("License" and "Access Right"), in the cases where versioning is applicable. More information can be found in Section 3.4 - Metadata Management. **Describe data quality assurance processes.** The data quality assurance processes are described in Section 3.3. However, the creators will also need to provide consent on the data quality level of the uploaded data. **Specify the length of time for which the data will remain re-usable.** The data will be available for re-use until the end of the project unless additional funds are secured to allow for continuation of the infrastructure use.
3. Allocation of resources	**Estimate the costs for making your data FAIR. Describe how you intend to cover these costs.** The work to be done in making the data FAIR will be covered by the assigned budget for the project. **Clearly identify responsibilities for data management in your project.** The data governance and the responsibilities of all roles are defined in Section 3.1. **Describe costs and potential value of long-term preservation.** The costs for long term preservation are part of the cost analysis for the design and development of the infrastructure.
4. Data security	**Address data recovery as well as secure storage and transfer of sensitive data.** Data Recovery as well as secure storage and transfer of sensitive data has been partially covered in Section 6 and will be further addressed in the deliverables of WP2, such as the reference architecture and services offered.
5. Ethical aspects	**To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former** The ethical considerations are described in Section 7.
6. Other	**Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)** No other procedures need to be put in place for data management.

Appendix A - Relationship with other Project Deliverables

This deliverable **receives input** from the following Tasks (T)/ Deliverables (D):

- T1.2/D1.2 Technical and operational requirements from the scientific community
  - Identification of user needs and key technical and operational requirements challenges, which are difficult or even impossible to solve using state-of-the-art experimental facilities, with focus on data management infrastructures and mobile networking environments.
- T1.3/D1.3 Analysis of legal compliance and regulation issues in Europe
  - Identification of administrative and compliance procedures and how these can technically and operationally support the infrastructure, with regards to data management.
- T2.1/D2.1/D2.3 SLICES reference architecture
  - Physical Infrastructure requirements and constraints for the overall SLICES infrastructure.
  - Run-time monitoring and control requirements for the overall SLICES infrastructure.
- T2.2/D2.2/D2.4 Service delivery scheme
  - Identification of user workflows that must be supported.
  - Identification of mechanisms to deliver services to the scientific community (suitable protocols and client tools).
- T3.1/D3.1/D3.4 Governance, management and human resources
  - Understand the governance for operations with identified responsibilities and reporting lines, including the International Scientific Advisory Board and the User Committee.
  - Define the SLICES human resources policy for implementation and capacity building, including the process for hiring and training for infrastructure development and operation.
- T3.2/D3.6 Data Protection Policies
  - Identification of data security, data protection/privacy and trust management policies and how these will be facilitated by the data management infrastructure.
  - Guidelines and architectural considerations on data-protection-by-design to ensure compliance with the European General Data Protection Regulation (GDPR) and potentially to complementary Directives, such as the NIST and ePrivacy Directives.
- T4.2/D4.2 Integration and interoperability with EOSC infrastructure and services
  - Options to connect SLICES to EOSC, in terms of exposing SLICES infrastructure and services via EOSC infrastructure.
  - Requirements on how to effectively interact with other RIs via the EOSC infrastructure and services, including API, metadata, data exchange and AAI identity federation.
- T4.3/D4.4/D4.5 Relations with international testbeds
  - Identification of how SLICES will interact with Next Generation Internet (NGI) European testbeds and other international digital technologies.
- T5.1 Stakeholders continuous engagement
  - Definition of key target groups of the proposed RI.
- D7.1 Ethics Requirements
  - Definition of ethics requirements and GDPR policy.

This deliverable **provides input** to the following Tasks (T)/ Deliverables (D):

- T1.2/D1.2 Technical and operational requirements from the scientific community
  - Identification of user needs related to data management and dissemination of data/metadata as well services required by the research community.
- T1.3/D1.3 Analysis of legal compliance and regulation issues in Europe

    - Provides a framework, mechanisms and guidelines to ensure compliance with specific regulations, e.g. GDPR.
- T2.1/D2.1/D2.3 SLICES reference architecture
    - Physical Infrastructure requirements for data management infrastructure.
    - Software and applications infrastructure, for different aspects of data management, such as data storage, data pre-processing and analytics.
    - Backend IT services and human resources to operate the data management infrastructure.
    - Data security and data protection, with emphasis on compliance with regulations, such as GDPR.
    - Run-time monitoring and control, and appropriate reporting for all components of the data management infrastructure.
- T2.2/D2.2/D2.4 Service delivery scheme
    - Identification of data workflows that must be supported.
    - Architectural design of data services that should be offered by SLICES.
    - Identification of mechanisms to deliver data-related services to the scientific community, such as data sharing, analytics, experiment reproducibility and dedicated data zones.
    - Identification of mechanisms for interoperability with existing infrastructures and other systems, such as incorporation of specific metadata standards and relevant interoperability components.
    - Mapping of FAIR data management principles to specific actions.
- T2.2/D2.5 Use case studies of the evaluation of candidate architectures
    - Identification of data related services and workflows that must be evaluated.
- T3.1/D3.1 Governance, management and human resources
    - Design of the Data Governance Group and its human resource hierarchy with identified responsibilities and reporting lines.
- T3.3/D3.2 Cost analysis and sustainability
    - Cost for implementation and support of the data management infrastructure operations and related human resources.
- T4.1 Data Management Policy and FAIR principles adoption
    - This deliverable is the main output of this Task.
- T4.2/D4.2/D4.5 Integration and interoperability with EOSC infrastructure and services
    - Analysis of existing research data management platforms and how to interact with them.
    - Analysis on how to interact with EOSC infrastructure and services, including API, metadata and data exchange.
- T4.3/D4.4 Relations with international testbeds
    - Analysis on how to interact with international testbeds.
- D4.3 Definition of the SLICES metadata profiles to support FAIR principles
    - Proposal of metadata format and properties to describe datasets.

Appendix B - Data Management Processing Form

The following structure illustrates a draft data processing form that highlights the fields that will need to be completed by data contributors (e.g., researchers or industry practitioners) as part of electronic data submission.

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/Accesss level	Open	○
	SLICES Node level	○ *(select node from list)*
	Shared	○ *(provide a list of one ore more organizations from a list)*
	Private	○
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 resourcce:	*e.g., link to terms (in list format)*
Provide any License(s) that are related to the resourcce	*specific licenses that apply to the resourcce*

Resource Characteristics (to be completed for each resource)		
What is the resource type?	Collection/Project	◯
	Single resource	◯
	Part of a collection	◯ *If yes, select collection/projeect from a list*
What is the measruement type (if any)?	Observational	☐
	Experimental	☐
	Simulation	☐
	Derived	☐
	Other: (please specify)	☐
What is the format of the resource	Open file format	◯
	Proprietary file format	◯ *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	◯*Yes* ◯*No* *e.g., if yes, provide requirements*
	Storage intensive	◯*Yes* ◯*No* *e.g., if yes, provide requirements*
	Network intensive	◯*Yes* ◯*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	☐
	Sensitive data	☐
	Data subject to license	☐
	Derived	☐
Provide appropriate consents for the resource:	Do you verify the completeness of the data?	◯*Yes* ◯*No* ◯*N/A*
	Do you verify the timelineess of the data?	◯*Yes* ◯*No* ◯*N/A*
	Have you obtained appropriate consents for the use/processing of personal	◯*Yes* ◯*No* ◯*N/A* *If yes, provide description and/or link to resource*

	data contained in the resource?		
	If you are using external resources, have you obtained appropriate licenses/rights to use them?	○*Yes* ○*No* ○*N/A* *If yes, provide description and/or link to licenses/rights*	
	Additional Consents (please specify)	○*Yes* ○*No* ○*N/A*	
Data Quality Assurance	*Do you verify the quality of the data during data collection?*	○*Yes* ○*No* ○*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)?	○*Yes* ○*No* ○*N/A* *Select yes if the instruments used for data collection provide quality assurances*	
	Have you used the recommended directory structure?	○*Yes* ○*No* *If no has been selected, please describe the structure*	
	Have you used the recommended naming conventions?	○*Yes* ○*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	☐
		Automatic numbering/Date-Time/Version number in the structure (directory/filename)	☐
		"Track changes" feature in software	☐ *Specify software and method*
		Dedicated version control software:	☐ *Specify software and method*
		Other	☐ *Specify method*

	Data Security	*Have any measures been taken to secury 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 measruement type (if any)?	Observational	☐
	Experimental	☐
	Simulation	☐
	Derived	☐
	Other: (please specify)	☐
What is the format of the resource	Open file format	◯
	Proprietary file format	◯ *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	◯*Yes* ◯*No* *e.g., if yes, provide requirements*
	Storage intensive	◯*Yes* ◯*No* *e.g., if yes, provide requirements*
	Network intensive	◯*Yes* ◯*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)*	

HORIZON 2020
H2020 – INFRADEV-2019-3

slices
DS

D4.1

Annex: SLICES-DS Data Management Plan

Project acronym:	SLICES-DS
Project full title:	Scientific Large-scale Infrastructure for Computing / Communication Experimental Studies – Design Study
Grand Agreement:	951850
Project Duration:	24 months (Sept. 2020 – Aug 2022)
Due Date:	28 February 2021 (M6)
Submission Date:	16 March 2021 (M7)
Dissemination Level:	Public
Authors:	UCLan Cyprus, SU, INRIA, UTH, MI, UC3M, UvA, eBOS
Reviewers:	ALL

Co-funded by the Horizon 2020 programme of the European Union

Executive summary

SLICES aims to design and implement a Europe-wide test-platform, to support large-scale, experimental research that will provide advanced compute, storage and network components. We expect that a large number of researchers will take advantage of SLICES to generate data of various sorts (observational, experimental, simulation) and produce research results. Additionally, researchers will have the opportunity to collaborate with other researchers by utilising data and services of other international testbeds that will be seamlessly supported through SLICES.

The main objective of the SLICES Design Study project, coined SLICES-DS, is to adequately design SLICES in order to strengthen research excellence and the innovation capacity of European researchers and scientists in designing and operating Digital Infrastructures. SLICES-DS will build upon the experience of the existing core group of partners, to prepare (in detail) the conceptual and technical design of the new leading-edge research infrastructure (RI), coined SLICES-RI, for the next phases of its lifecycle. This will be accomplished by meeting specific project objectives, which are briefly summarized below:

- Analyse the needs of the scientific and industry communities and translate them into an architecture and a roadmap for the long-term evolution of SLICES-RI
- Carry out all preparatory work and planning required for realising SLICES-RI
- Define the governance and management principles underpinning the new RI
- Define models for the financing of the new RI
- Define stakeholder and engagement strategy on community-based research

This addendum provides the Data Management Plan for SLICES-DS, which describes the data that will be generated during the course of the project to meet the aforementioned objectives, how it will be managed and what mechanisms will be used for sharing data.

This is the first version of the Data Management Plan, which targets only the SLICES-DS project. The DMP will be updated as the project progresses and will be finalized by the end of the project.

Table of contents

Acronyms

DGG - Data Governance Group
DMP – Data Management Plan
DPO - Data Protection Officer
DQM - Data Quality Management
EOSC - European Open Science Cloud
GDPR – EU General Data Protection Regulation
NGI – Next Generation Internet
RI – Research Infrastructure
SME – Small-Medium Enterprise

1 Introduction

The overarching objective of the SLICES Design Study project is to design a pan-European experimental platform, which aims to strengthen research excellence and the innovation capacity of the European research community in the ICT field. To accomplish this, the project needs to understand the key requirements necessary for preparing SLICES-RI and ensuring that multiple objectives are met. To this end, the project has adopted four categories of actions: (i) Requirements Collection, (ii) Design, (iii) Qualification and (iv) Impact. It is important to note that the project is a design study and as such, it is not expected to generate large amounts of *research* data. However, during the different actions that will be carried out during the project's lifetime, such as collecting requirements from user communities and communicating with external partners, some data will be collected.

The focus of this addendum is to understand the different types of data to be collected in the SLICES-DS project, as well as the corresponding data management policies that manage them.

2 Data Summary

- What is the purpose of the data collection/generation?
- What is the relation to the objectives of the project?
- What types and formats of data will the project generate/collect?
- Will you re-use any existing data and how?
- What is the origin of the data?
- What is the expected size of the data?
- To whom might the data be useful ('data utility')?

In order **to facilitate the creation and adoption of a Digital and Data Research Infrastructure in Europe,** the project will collect and analyse information on the current state of relevant technologies and infrastructure developments. Furthermore, it will deliver recommendations and practical approaches to facilitate research in the area of digital technologies for science. Finally, it will develop data exchange policies, and relevant practices and services. To this end, the project will collect and manage different types of data, which are listed below:

- **Data collected or generated** using (i) *primary research* methods, such as surveys (e.g. using online questionnaires), interviews and workshops, and (ii) *secondary research* methods involving the analysis and synthesis of existing research results, such as reports on best practices, design guidelines and recommendations. The majority of data will be collected in WP1 and WP5, through activities/tasks which are described below:
    - (secondary research) Analysis of the technological status and capabilities of state-of-the-art facilities in Task 1.1.
    - (primary research) Collection and analysis of technical and operational requirements from the scientific community in Tasks 1.2 and 5.1. This also includes data related to the organization of workshops, such as lists of participants.
    - (secondary research) Analysis of legal compliance and regulation issues in Europe in Task 1.3.
    - (secondary research) Analysis of trends and the technical evolution of key ICT and communication technologies in Task 1.4.
    - (secondary research) Analysis of the standardization activities/processes in the related technology/industry domains, such as Industry 4.0, Big Data in Task 1.4.
    - (secondary research) Analysis of the interactivity of SLICES with other infrastructures/systems, such as EOSC and NGI testbeds in Task 4.2 and 4.3.
    - (primary research) Organization of workshops and analysis of technical and operational requirements from the scientific community in Task 1.2.
    - **Dissemination and Communication activities**, which entail generating content for various dissemination/communication channels, such as the website, social media, flyers, posters and videos (e.g., webinars, project

presentations) and the bi-annual e-newsletter (as described in Task 5.2).

- **Project-related data** such as deliverables consisting of the results of the project, which may include synthesized data, meeting minutes, etc.
- **Software Code** for the purposes of: (i) integrating with the EOSC services (Task 4.2); (ii) FAIR data management of produced datasets (Task 4.1); and (iii) delivering and validating tools for SLICES-DS services. The planned software products and tools will use well-established software and APIs to ensure easy integration and interoperability.

The data collected in SLICES-DS will be **useful** for a diverse set of stakeholders, such as research/industry/funding organizations, policy makers in European Research Area (ERA), and for different types of infrastructures.

When possible, we **reuse existing information** during preparation, extracting information from public reports and documentation, EOSC services descriptions, and EOSC marketplace resources. We will use openly accessible sources and reuse existing, openly licensed survey data wherever feasible (e.g., Horizon2020 template for Data Management Plans). The desk research results, combined with the survey data, will be published as part of the project's deliverables.

The following activities will lead to data collection, analysis, storage and publication:

- **Survey/Interview data** will be gathered in WP1. The number of interviewees and the final handling of the data will be determined at a later stage. Any decisions made will be based on the information needs identified after the completion of the desk research and survey data analysis. This approach minimises the need to ask for personal data in the survey and maximises efficiency, as interviews are time consuming for all parties involved. In the beginning of the project, the results will be stored on the instance of NextCloud that is running at Sorbonne University, named DropSU[55]. DropSU uses industry-standard SSL/TLS encryption for data in transfer. Additionally, data at rest (in storage) can be encrypted using a default military grade AES-256 encryption with server-based or custom key management. Optionally and on a per-folder base, data can be end-to-end encrypted on the client side, with the server assisting in sharing and key management operations using a Zero-Knowledge model.
- **Dissemination and Communication activities** will be implemented in WP5 to support outreach activities. A contact list of interested parties will be collected through the project's website, hosted by UTH, and stored in a secure location. Dissemination data will be available through the website or selected social media accounts, which will be maintained by UTH.
- **Project-related data** such as deliverables with public access rights will be made available through the website. Confidential deliverables and other project related data will be stored on DropSU. Management data (e.g., financial data) will be stored on the project NetBoard platform.
- **Software** tools will be developed using a plethora of development tools, but the code will be stored in the well-known GitHub[56] repository. Publishing software will be accomplished using the standard upload connection from GitHub to Zenodo[57], where a persistent identifier (i.e., DOI) is assigned.

[55] DropSU, https://intranet.sorbonne-universite.fr/_resources/Universite/Cybersecurite/guide%2520RGPD-PSSI-BAT-17-07.pdf?download=true
[56] GitHub Code Repository, https://github.com/
[57] Zenodo, https://zenodo.org/

The file formats chosen are open and comply with good practices and the sustainability preferences of the repositories that will host the data and software code, as described in the appropriate sections of D4.1.

The **estimated volume of the data to be archived and published** is less than 50 GB. All types of data, along with their corresponding projected contributions to the overall volume, are described in Table 1.

Table 5: List of all datasets to be generated and published by SLICES-DS, with volume indication

Origin	**File format**	**Methods or software needed (if any)**	**Estimated volume**	**Task number or WP number**	**# in the description above**
Surveys	.csv	standardoffice software	<10 GB	WP1	1
Survey documentation (e.g. questionnaire, codebook)	.txt, .pdf/A	standardoffice software	<10 GB	WP1	1
Interviews	.txt, .pdf/A	standardoffice software	<10 GB	WP1	1
Dissemination Material	Various media formats	Media production suites	<20GB	WP5	t.b.d.
Software code	t.b.d.	t.b.d.	<10 GB	WP2, WP4	2

3 FAIR Data Management

3.1 Making Data Findable

- Are the data produced and/or used in the project discoverable with metadata?
- Are the data identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
- What naming conventions do you follow?
- Will search keywords be provided that optimize possibilities for re-use?
- What is the approach for clear versioning?
- What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

At the end of the project, the data will be published in a certified digital repository. This will be one of the repositories that is recommended and recognised by the EC Data Management initiative and the research community, such as the EASY archive by DANS (https://easy.dans.knaw.nl/), which is a CoreTrustSeal-certified repository for research data. The repository will provide Dublin Core metadata, which contain keyword information (under the "Subject" field). Each dataset in the repository will have a unique and persistent identifier. EASY is designed for long-term data preservation and availability. Furthermore, a SLICES-DS community space will be set up in the Zenodo archive (https://zenodo.org/) for project reports and software code, which will be accompanied by associated documentation (through GitHub). Zenodo also assigns persistent identifiers and is compliant with the DataCite metadata schema.

Naming consistency is important for efficiently locating a resource and understanding its use. However, it is up to the creator to provide proper names for research outputs and related data files. SLICES adopts certain Naming Conventions/Guidelines to improve the structure/consistency of files. The draft guidelines include the following recommendations:

- **Name Length:** Apply a maximum length for file names, as long filenames may not be interoperable with some systems.
- **Date Format:** Allow the display of dates in a chronological order, even over the span of many years; use the YYYYMMDD format.
- **Leading Zeros:** Use leading zeros to make an ascending order of numbers correspond to alphabetical order.
- **Naming Scheme:** Use a consistent naming scheme throughout; do not use spaces or punctuation symbols as these may not be interoperable with some systems. Order / confirm which element should go first, so that files on the same theme are listed together and can be found easily. Project deliverables based on (and referring to) files, as well as other documentation (see below) will provide more context information.

File organisation is important for efficiently locating a resource, even in cases where there is no predefined structure available. SLICES utilizes certain guidelines to improve the consistency of the structure of the data. The initial guidelines include the following recommendations:

- **Hierarchical Structure**: Adopt a hierarchical structure that includes at least the following folders:
    - **Data**: includes all input data, when data is not related to experiments
    - **Experiments**: includes a folder for each experiment (e.g. **exp01, exp02**). Each experiment folder should include at least the following folders:
        - **input data**: contains all data required for the experiment
        - **software**: contains all software components, models for the experiment or appropriate links
        - **deployment**: contains the steps on how to conduct the experiment
        - **output data**: contains the results of the experiment
    - **Relationships**: contains references to relationships with any other data
    - **Dissemination**: contains any material related to dissemination, such as presentations, press releases, articles, etc.
    - **Miscellaneous**: contains any other material
- **Folder Naming:** Use the naming scheme provided in the previous paragraph

3.2 Making Data Accessible

- Which data produced and/or used in the project will be made openly available as thedefault?
- If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
- If there are restrictions on use, how will access be provided?
- Is there a need for a data access committee?
- Are there well described conditions for access (i.e. a machine readable license)? How willthe identity of the person accessing the data be ascertained?
- How will the data be made available (e.g. by deposition in a repository)?
- What methods or software tools are needed to access the data?
- Is documentation about the software needed to access the data included? Is it possible toinclude the relevant software (e.g. in open source code)?

The SLICES-DS grant agreement states in Article 29 that "Unless it goes against their legitimate interests, each beneficiary must — as soon as possible — '**disseminate**' its results by disclosing them to the public by appropriate means (other than those resulting from protecting or exploiting the results), including in scientific publications (in any medium)". It also states that "Each beneficiary must ensure open access (free of charge online access for any user) to all peer-reviewed scientific publications relating to its results."

The data produced will be made openly available (by default) and will be disseminated in various ways. In particular:

- **Survey/Interview data**, including documentation such as questionnaires and codebooks, that do not contain personal data or have been anonymized, will be made available via certified repositories and/or via the SLICES-DS website. For non-anonymized data, appropriate consents will be drafted in order to assure the rights of the data subjects. Survey/interview data concerning interviewees that do not provide consent to share, object to processing of their data, or withdraw their consent at any point, will not be shared. Data that will be published will be licenced by a CC BY-SA licence. The EASY research data archive[58] will be used to preserve the data. This is a CoreTrustSeal-certified repository for research data, which accommodates both Open Access with the CC licence, and Restricted Access for the interview data with a stricter licence, for only those users who receive permission to access these data. All metadata in EASY are public. They adhere to the Dublin Core metadata standard and implementation-controlled vocabularies are used, for e.g., media type, language, and the relation of the dataset to other resources.
- **Dissemination and Communication data** will be open by default and will be accessible/disseminated in various formats through different outreach activities. Any personal data, such as contact data of mass dissemination lists, will only be stored if appropriate consents have been provided, while appropriate protocols/procedures will be made available for users to exercise their rights.
- **Project-related data,** such as deliverables with public dissemination level, will be made available through the website. Confidential deliverables and other project related data will be stored on the DropSU system and will be made available after 4 years, as part of the obligations of Article 36.1 of the grant agreement ("During implementation of the action and for four years after the period set out in Article 3, the parties must keep confidential any data, documents or other material (in any form) that is identified as confidential at the time it is disclosed ('confidential information').").
- **Software** tools will be published in the Github repository and will be accompanied by an appropriate license, e.g. GPL-3.0, MIT. Publishing software will be accomplished using the standard upload connection from Github to Zenodo, where a persistent identifier is assigned. Database files will also be accompanied by an ODC-By license.
- **Scientific publications** that may arise from the project results will be published in open access venues and shared through Zenodo.

[58] EASY research data archive, https://easy.dans.knaw.nl/

3.3 Making Data Interoperable

- Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering tostandards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets fromdifferent origins)?
- What data and metadata vocabularies, standards or methodologies will you follow tomake your data interoperable?
- Will you be using standard vocabularies for all data types present in your data set, toallow inter-disciplinary interoperability? If not, will you provide mappings to more commonly used ontologies?

The file formats we will deliver (see table above) are open and allow for straightforward reuse. The Dublin Core and DataCite metadata specifications are widely accepted and implemented.

Furthermore, interoperability is the goal of WP4 ("Integration and compatibility with EOSC, FAIR Data management and External RIs"), which will actively participate in domain-specific and cross-disciplinary initiatives involved in semantic interoperability.

3.4 Making data reusable

- Are data quality assurance processes described?
- How will the data be licensed to permit the widest re-use possible?
- When will the data be made available for re-use? If applicable, specify why and how longa data embargo is needed.
- Are the data produced and/or used in the project usable by third parties, in particularafter the end of the project? If the re-use of some data is restricted, explain why.
- How long is it intended that the data remains re-usable?

To the extent that data form the basis of project deliverables, internal quality review procedures will apply. These are described in the SLICES-DS Project Partner Guide (D6.1). Other quality aspects, such as an explicit methodology, are covered by the FAIR aspects addressed above, and in particular by providing sufficient documentation to interpret the data and the code correctly, which relates to FAIR principle R1.2: "(Meta)data are associated with detailed provenance". Such documentation, e.g. sample questionnaires, codebooks and templates of informed consent forms (see Section "Ethical Aspects"), will be published along with the data and code.

Licensing issues have been addressed in Section 3.3

Data and code will be made available through certified digital repositories that support the relevant types of licences, and that will be able to preserve the data for the long term (in principle "indefinitely"). We will set up a SLICES-DS community space in the Zenodo archive for the project results; at the end of the project, this community space could be handed over to an EOSC organisation for future extension and maintenance, ownership arrangements permitting.

4 Allocation of Resources

- What are the costs for making data FAIR in your project?
- How will these be covered?
- Are the resources for long term preservation discussed (costs and potential value, whodecides and how what data will be kept and for how long)?
- Who will be responsible for data management in your project?

An important goal for the project's team is that of delivering data that is as FAIR as possible. Therefore, no plan exists for employing a separate step or explicitly allocating budget for the purpose of making data FAIR. The amount of research data envisaged in the project is very modest, therefore a cost/benefit analysis for the long-term storage of each component is not needed. The costs for long-term preservation in a trustworthy archive are covered.

Each WP leader is responsible for data management in their respective WP, including the implementation of and, if necessary, updates to this DMP. The Project Coordination Office is responsible for the overall data management and for evaluating the implementation of this DMP. Evaluations will take place in months 12, 18 and 24.

5 Data Security

- What provisions are in place for data security (including data recovery as well as securestorage and transfer of sensitive data)?
- Is the data safely stored in certified repositories for long term preservation and curation?

During the project's lifecycle, research data is stored in the DropSU system, in separate folders per WP (see D6.1 SLICES-DS Project Partner Guide, Chapter 4.2 for more details). This allows for file sharing across partners and keeping track of revisions. A retention rule is set for the project to keep files indefinitely. Access to the DropSU Drive is managed by the Project Coordination Office. DropSU network is protected from external attacks. Data belonging to project customers is stored at rest in two types of systems: disks and backup media. DropSU also stores data on offline backup media to help ensure recovery from any catastrophic error or natural disaster at one of their datacenters. With respect to data protection, DropSU is committed to complying with the EU General Data Protection Regulation (GDPR).

Management data will be stored on the project NetBoard platform. The platform is hosted on Absiskey remote servers hosted by Altitude Telecom, which uses a backbone IP in a bunker zone providing high security level services. This network runs more than 12 000 local loops conveyed on an IP MPLS network. Access to the platform is only allowed through a secured connection (SSL encryption) with a personal login and password for each user. The subscription to the platform will last for the whole project duration, plus 6 months extension to end the final report. At the end of each project, the data will be accessible to the project coordinator and to partners on demand for 5 years (legal period imposed by the possible audit, ordered by the funding authority).

For conducting and analysing surveys, SLICES-DS will use software such as LimeSurvey[59], EUSurvey[60], Qualtrics[61] or Online Surveys[62]. Qualtrics and Online Surveys have the ISO 27001 certification for an Information Security Management System[9]. All survey tools are GDPR-compliant. The CoreTrustSeal certification requirements of the EASY research data archive include a pertaining requirement: "The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users"[63].

Near the end of the project, the research data and code will be published and safely preserved in a certified digital repository such as Zenodo, with good and transparent data security processes in place.

[59] LimeSurvey, https://www.limesurvey.org/
[60] EUSurvey: https://ec.europa.eu/eusurvey/
[61] Qualtrics: https://www.qualtrics.com/
[62] Online Surveys: https://www.onlinesurveys.ac.uk/
[63] Core Trustworthy Data Repositories Requirements, https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

6 Ethical Aspects

- Are there any ethical or legal issues that can have an impact on data sharing? These canalso be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).
- Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

SLICES-DS reaches out to many organisations, individuals, and other projects, and will organise surveys, interviews, workshops and other meetings, and repository calls. The project must process personal data under the Agreement in compliance with applicable EU and national law on data protection (including authorisations or notification requirements). Data collected for the purposes of administering the project activities will be held securely and according to legislation, and will not be shared externally. SLICES-DS will only collect personal data that is necessary to fulfil the information needs of the project (respecting the principle of data minimisation), and we will not collect "special categories of data" in terms of the GDPR. Although administrative personal data are outside of the scope of the DMP, we describe next how they are managed:

Any personal data will be protected, as stated in Article 39 of the Grant Agreement. SLICES-DS will provide a Privacy Policy, which addresses personal data (processing, data subject's rights, opt-out, cookies used on the website and in social media, etc). For example, personal data processed for applications collected via the SLICES-DS website will be kept by SU as Data Controller, in terms of GDPR, for up to 5 years (as indicated by Article 18.1 Obligation to keep records and other supporting documentation), to allow for possible external audits, as requested by contractual provisions the Data Controller is subjected to. The data retention period is not specified according to GDPR, and has thus been specified based on project needs, e.g. project audits. In addition to the Privacy Policy Statement, a Terms of Use statement will be compiled when services become available through the project website.

SU, the project leader, has already appointed a Data Protection Officer. The contact details of the Data Protection Officer are made available to all data subjects involved in the research.
Survey, interview and workshop participants (WPs 1, 2 and 5), as well as repositories responding to the SLICES-RI coordination call will be provided with a clear statement of the purposes of data collection, how the data will be used, and with whom it may be shared. Participants will have the opportunity to decide if they want to provide any personal information such as their name and email address. Those who choose to provide this information and agree to be contacted for interviews or for news and updates on the project work, will be added to the contacts database managed by SU or WP5, as the Data Processor in terms of the GDPR.

We developed a project-wide template for informed consent regarding interviews. The template can only be used in combination with an information sheet (in language and terms understandable by participants). Detailed information on the informed consent procedures in regard to data processing will be kept on file and archived. Also, templates of informed consent forms and information sheets will be kept on file and archived.
In addition to the information provided, the ethics requirements, defined in the Ethics Summary Report, will be fully addressed in deliverable D7.1.

7 Other Issues

- Do you make use of other national/funder/sectorial/departmental procedures for datamanagement? If yes, which ones?

Countries or institutions may impose additional regulations to ensure additional properties on research data. For example, ethics are considered an integral part of research for activities funded by the European Union. In several countries, which are also part of this consortium, there are committees that impose additional regulations to data collected from research activities, such as the Cyprus National Bioethics Committee, the French Comités de Protection des Personnes, the Medical Research Ethics Committee in the Netherlands, and the Swedish Ethical Vetting Board[64]. The research data uploaded or generated within SLICES should comply with the national and institutional regulations and be supplemented with evidence of potential approvals required by formal bodies. SLICES-DS partners also agree with the principles and good practices in the European Code of Conduct for Research Integrity[65].

[64] European Conference of National Ethics Committees (COMETH), https://www.coe.int/t/dg3/healthbioethic/cometh/national_ethics_committees/ [Last accessed 20 February 2021]
[65] The European Code of Conduct for Research Integrity, https://allea.org/code-of-conduct/