GreenDIGIT

Greener Future Digital Research Infrastructures

Deliverable 1.2 Data Management Plan, Ethics and AI use practices

GRANT AGREEMENT NUMBER: 101131207

This project has received funding from the European Union's HE research and innovation programme under the grant agreement No. 101131207

Lead Beneficiary: Mandat International

Type of Deliverable: DMP — Data Management Plan

Dissemination Level: Public

Submission Date: 20.08.2024

Version: 1.0

Versioning and contribution history

Version	Description	Contributions
0.1	Draft template	Angelica da Silva Lantyer (UVA) drafts the template for MI to prepare GreenDIGIT DMP - internal deadline Jul 31st.
0.2	Initial draft	Iida Lehto (MI) and Adrian Quesada Rodriguez (MI) prepare the draft and integrate inputs from GreenDIGIT consortium partners (from DMP survey).
1.0	Final draft	Angelica da Silva Lantyer reviews initial draft and adds UvA inputs.
		Iida Lehto integrates reviews from Yuri Demchenko into the final draft.

Authors

Author	Partner
Sébastien Ziegler	MI
Adrian Quesada Rodriguez	MI
Cédric Crettaz	MI
Iida Lehto	MI
Maria Roglekova	MI

Reviewers

Name	Organisation
Yuri Demchenko	UVA
Angelica da Silva Lantyer	UVA

Disclaimer

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use, which may be made of the information contained therein.

Executive Summary

This Data Management Plan (DMP) aims to illustrate and offer an insight into the types of data that are foreseen to be generated during the GreenDIGIT project. This initial DMP version presents types of data that will be collected, created or processed by GreenDIGIT partners. The DMP demonstrates plans and measures to be undertaken to follow Findable, Accessible, Interoperable, and Reusable (FAIR) principles. The DMP also outlines the main ethical themes that arise from data collection and processing, which will be further analysed in future editions of the deliverable (final version planned on M36 (D2.1)). Lastly, the deliverable describes the project's commitments towards data security, while providing an overview of the use of Artificial Intelligence (AI) and Intellectual Property Rights (IPR).

The DMP is supported by a data management survey (see Annex A) completed by all GreenDIGIT partners. The survey has been created with the goal to present the initial partner prognosis and individual strategies, which will be regularly updated when the data management plans of partners evolve.

Table of contents

List of tables

List of abbreviations

Abbreviation	Description
AI	Artificial Intelligence
API	Application Programming Interface
DMP	Data Management Plan
DPO	Data Protection Officer
DPIA	Data Protection Impact Assessment
FAIR	Findable, Accessible, Interoperable, and Reusable
GDPR	General Data Protection Regulation
GreenDIGIT	Greener Future Digital Research Infrastructures (current project)
HW	Hardware
IPR	Intellectual Property Rights
MRS	SLICES Metadata Registry Service (MRS)
PDI	Preservation Description Information
PID	Persistent Identifier
RI	Research Infrastructure
RO-Crate	Research Object Crate
RSSI	Received Signal Strength Indicator
SLA	Service Level Agreements
SNR	Signal-to-Noise Ratio
TOM	Technical and Organisational Measure
VRE	Virtual Research Environment
WP	Work Package

1 Introduction

The overall objective of GreenDIGIT is to develop and validate a consistent framework, including a package of innovative technical solutions, models, and tools to increase the sustainability of digital research infrastructures (RIs) by lowering their environmental and climate impact throughout their whole lifecycle.

The existence of the GreenDIGIT project is fundamental to the need to make Europe a climate neutral continent by 2050.[1] Climate neutrality is the aim of several European objectives and priorities, namely the European Green Deal. It works through different initiatives: transformation of the economy, making transportation sustainable, building a sustainable energy system, etc. One of the goals of the Deal is the green industrial revolution. Under this goal, the European Commission aims to introduce clean technologies and products to the market and transform the current economy to help reduce carbon emissions. Digital research infrastructures must uphold the Commission's plan in terms of providing environmentally sustainable new technologies.

GreenDIGIT brings together four research infrastructures that represent significant stakeholders of the ESFRI DIGIT area, which aims to boost a sustainable transition: SLICES, SoBigData, EGI, and EBRAINS. In order to reduce environmental impact not only in the DIGIT sector but also throughout the entire ESFRI landscape, it is imperative that energy efficiency within these infrastructures be addressed. For this reason, five objectives have been outlined:[2]

1. Evaluation of the four RIs as well as the broader network with the purpose of creating recommendations and roadmaps for providers.
2. Create design principles, reference architecture, and actionable model for RIs to establish a system for environmental impact assessment and monitoring.
3. Implement novel technologies procedures, and instruments for digital service providers to help them cut down on their energy use and environmental effect overall.
4. Develop the above-mentioned technologies in a manner that is consistent with FAIR principles for data management, Open Science, and reproducibility considerations.
5. Provide guidance and assistance to digital service providers operating in the RI communities about environmentally aware lifecycle management and infrastructure and service operation.

Moreover, the project is set to target three areas of development: (1) scientific applications carrying out scientific workflows and data processing; (2) data and data management infrastructure guaranteeing data storage, access, and sharing; (3) physical infrastructure and datacentres comprising computation, storage, network, and IoT/sensor devices.

However, as the project develops with time, it becomes evident that more specific and clear metrics or benchmarks for energy efficiency and carbon reduction specifically tailored for digital RIs within the

[1] 'Delivering the European Green Deal' (*European Commission*) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_en#transforming-our-economy-and-societies> accessed 29 July 2024

[2] GreenDIGIT Proposal

Green Deal are needed. For this reason, the project will address different gaps and measures to reduce environmental and climate impact through the full RI lifecycle.

1.1 DMP Purpose and Objectives

The Data Management Plan (DMP) of GreenDIGIT aims to provide an overview of the expected types and methods of data collection, storage, or processing by all partners during and beyond the project lifecycle. The DMP supplies a plan for compliance with FAIR principles, including an initial ethics review of the project's data management, data security, as well as consideration of AI usage practices.

The DMP is prepared in accordance with the objectives of WP1 as the overall administration and management of the project. Concrete WP1 objectives include:

1. Managing the project through administrative, financial and innovation coordination; Developing quality control procedures and identifying and managing risks to the project; Ensuring prompt payments to consortium partners;
2. Ensuring the submission of project deliverables and achievement of project results on time and ensuring effective communications within the consortium and with the EC;
3. Preparing and updating the project's Data Management Plan;
4. Preparing Ethics Requirements and monitoring regulatory compliance.

2 Data summary

2.1 Main categories of data

There are three different categories of data that will be relevant to the GreenDIGIT project. All partners will handle the use of non-personal data; some partners will handle the use of personal data in connection to WP task requirements; and the participating RIs of GreenDIGIT may handle the use of other special categories of data. See below the breakdown of expected data categories:

1. Non-personal data (environmental data), including:
    - Carbon intensity of electricity;
    - Workload and computing environment related HW usage metrics and control interventions;
    - Time-series data from monitoring of hardware resources;
    - Performance indicators of software systems;
    - SNR and RSSI readings from a LoRa Network.
2. Personal data (any information relating to identified or identifiable individuals, including for instance email or IP addresses). The personal data will be collected through different means:
    - Directly from data subjects who belong to the research team;
    - Directly from data subjects outside the research team (i.e. DIGIT survey participants and interviewees, beta-testers, early adopters).
    - Indirectly through partners of the project;
    - Indirectly through other organizations external to the project.
3. Special categories of data, which may include personal data revealing sensitive information, such as sexual orientation, racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, as well as any health, genetic or biometric data related to the data subjects.

2.2 Datasets

This section presents the initial list of datasets anticipated at the start of GreenDIGIT. As the project is in its early stages, this list summarizes the expected datasets and their characteristics likely to be used and produced. This list will be continuously updated throughout the project's development, offering detailed descriptions as they become available, adding new datasets as needed, and removing those not used as expected.

For the purpose of finding resources efficiently and comprehending their use, consistency in naming is crucial. Nonetheless, it is the creator's responsibility to provide research outcomes and associated data files' appropriate titles. Naming could include a lenghty name, displaying data format, using leading zeros, using a consistent naming scheme.

In addition, the datasets should describe four aspects: (1) how data will be dealt with; (2) how data will be administered; (3) how data will be stored and shared; (4) how data will be made legally and ethically compliant. The first requirement entails data formats and software; file transfers, file sharing, and remote access; file naming and formatting; version control. The second requirement also includes access restrictions, documentation, back-ups, and security.

The final version of this deliverable (D2.1) will demonstrate details on the kind of data processed as well as the format. Preference to open and standard formats will be given.

Table 1: Research outputs from modelling energy efficiency

Name of the dataset	Research outputs from modelling energy efficiency
Main Involved Partner	**University of Amsterdam**
Dataset Description	**Datasets are not yet defined. Research data are managed by researchers who use SURF Research Drive. For specific research, data may be obtained from open-source repositories.**
Type of personal data *(if relevant)*	**N/A**
Purpose for using/processing the dataset	**Research/experimentation, modelling, ensuring research reproducibility**
Format(s) of the dataset	**Multiple, there is no common dataset format. Commonly used format acceptable by Zenodo. Defined by the researcher or projects.**
Dataset storage	**SURF Research Drive**
Dataset origin	**Experiments and computer modelling**
Dataset ownership	**No rules for ownership, but researchers have the responsibility to make datasets open and published as outcomes of the project and research**
Restrictions for use	**No, datasets are intended to be reused for research reproducibility**
Dataset access	**Datasets are publicly published on Zenodo and available on research Github**
Length of dataset storage	**Not defined**
License to access the dataset	**Open-source licences are used for research datasets.**
Additional comments	**UvA's policy for data management: https://rdm.uva.nl/en/planning/uva-policy/uva-and-auas-policy.html**

Table 2: Carbon intensity of electricity

Name of the dataset	Carbon intensity of electricity

Main Involved Partner	**CESNET**
Dataset Description	**A dataset describing the current and historical carbon intensity (and price) of electricity.**
Type of personal data *(if relevant)*	**N/A**
Purpose for using/processing the dataset	**We will use this dataset as an input for power control of the infrastructure components. In the planning and analytics part of our work, we will combine this dataset with data describing real-world or artificial workload and HW usage metrics (including power consumption).**
Format(s) of the dataset	**Format style inspiration: https://www.electricitymaps.com/get-our-data**
Dataset storage	**Local database – copy of primary source**
Dataset origin	**External source, for example https://www.electricitymaps.com/get-our-data**
Dataset ownership	**We expect to use only public versions of this datasets.**
Restrictions for use	**Data policy from the source, example inspiration - https://www.electricitymaps.com/get-our-data**
Dataset access	**Public**
Length of dataset storage	**For the project lifespan**
License to access the dataset	**To be decided**
Additional comments	**It is a part of the project to select proper sources and clarify usage rules/licences.**

Table 3: Workload description

Name of the dataset	Workload description
Main Involved Partner	**CESNET**
Dataset Description	**Dataset describing workload (id of workflow, software, dataset), computing environment (cluster/HW conf), related HW usage metrics and control interventions (like CPU/GPU frequency limitation).**
Type of personal data *(if relevant)*	**We expect that data describing users or IP addresses of resources will be anonymized (description of users removed and resources addressed by IP transformed into technical IDs).**
Purpose for using/processing the dataset	**The dataset will be collected and used for planning and analytics work. For example, simulation or evaluation of new planning algorithms and controlling strategies.**
Format(s) of the dataset	**Format style inspiration: The MIT Supercloud Dataset https://arxiv.org/pdf/2108.02037**
Dataset storage	**Distributed storage attached to individual computing resources.**
Dataset origin	**Computing environment MetaCentrum, Czech national academic distributed computing infrastructure.**

Dataset ownership	**The dataset is owned by CESNET. We expect work with similar datasets owned and provided by RIs operated by consortium members.**
Restrictions for use	**We don't know about written policy. We expect creation of one.**
Dataset access	**Currently only internal access for CESNET distributed computing environment infrastructure operators.**
Length of dataset storage	**For the project lifespan**
License to access the dataset	**To be decided**
Additional comments	**It is a part of the project to select proper sources and clarify usage rules/licences.**

Table 4: SZTAKI HUN_REN Science Cloud Infrastructure Monitoring Data

Name of the dataset	SZTAKI HUN_REN Science Cloud Infrastructure Monitoring Data
Main Involved Partner	**SZTAKI**
Dataset Description	**The dataset contains time-series data from the monitoring of the hardware resources (CPU, net, SSD, RAM) of the SZTAKI scientific computing cloud infrastructure.**
Type of personal data *(if relevant)*	**N/A**
Purpose for using/processing the dataset	**The purpose is to use machine learning to identify patterns from hardware resource usage data to optimize resource utilization.**
Format(s) of the dataset	**Multivariate time-series with 1-minute sampling intervals in Comma Separated Values (CSV) format.**
Dataset storage	**The data is stored on servers with authentication.**
Dataset origin	**The data is collected using Prometheus monitoring tool's Node Exporter modules, which record hardware resource metrics (e.g., CPU Usage, RAM Usage, power consumption in KWh, fan speeds, etc.).**
Dataset ownership	**HUN-REN SZTAKI HBIT (Network Security and Internet Technologies) department.**
Restrictions for use	**The data is only accessible to authorized personnel. The scope of authorized personnel is determined by the HUN-REN SZTAKI HBIT department but can be extended as needed.**
Dataset access	**Employees of HUN-REN SZTAKI HBIT and HUN-REN SZTAKI LPDS (Laboratory of Parallel and Distributed Systems).**
Length of dataset storage	**There is no time limit. Data is continuously stored during operation.**
License to access the dataset	**N/A**
Additional comments	**N/A**

Table 5: GreenDIGIT WP3 RI survey responses

Name of the dataset	GreenDIGIT WP3 survey responses
Main Involved Partner	**Sorbonne University, Mandat International, EGI**
Dataset Description	**Survey and landscape analysis of RI practices and needs**
Type of personal data *(if relevant)*	**Name and email of participants**
Purpose for using/processing the dataset	**To identify practices, metrics and tools regarding environmental sustainability and impact reduction.** **To contribute to the development of sustainable solutions and the promotion of best practices within the European research digital infrastructure community, also to the development of the digital RIs technical recommendations and policy development.**
Format(s) of the dataset	**Collected via LimeSurvey tool and GoogleDocs**
Dataset storage	**UvA Research Drive**
Dataset origin	**Responses to the survey**
Dataset ownership	**GreenDIGIT consortium**
Restrictions for use	**N/A**
Dataset access	**T3.1 members and other project members using the datasets as input to their project related activities**
Length of dataset storage	**Duration of the project**
License to access the dataset	**N/A**
Additional comments	**N/A**

Table 6: Metrics collected from DIGIT service providers

Name of the dataset	Metrics collected from DIGITservice providers
Main Involved Partner	**EGI**
Dataset Description	**Metrics collected from DIGIT service providers about their environmental impact**
Type of personal data *(if relevant)*	**Name and email of participants**
Purpose for using/processing the dataset	**To identify practices, metrics and tools regarding environmental sustainability and impact reduction.**
Format(s) of the dataset	**GoogleDocs**
Dataset storage	**UvA Research Drive**
Dataset origin	**1-to-1 interviews with RI sites' representatives**
Dataset ownership	**GreenDIGIT consortium**
Restrictions for use	**N/A**
Dataset access	**GreenDIGIT T3.1 members and other project members using the datasets as input to their project related activities**
Length of dataset storage	**Duration of the project**

License to access the dataset	**N/A**
Additional comments	**N/A**

Table 7: Greenspector's historical measures

Name of the dataset	Greenspector's historical measures
Main Involved Partner	**Greenspector**
Dataset Description	**This dataset contains measured performance indicators of software systems.**
Type of personal data *(if relevant)*	**N/A**
Purpose for using/processing the dataset	**The dataset is used to compare performance of software systems, identify trends, detect and explain performance faults.**
Format(s) of the dataset	**Relational database and document-oriented database**
Dataset storage	**The dataset is stored on the organization's internal systems.**
Dataset origin	**This dataset is built on measures performed on the organization's internal testbench.**
Dataset ownership	**Greenspector**
Restrictions for use	**N/A**
Dataset access	**Members of the development team and the R&D team of Greenspector.**
Length of dataset storage	**No depreciation is planned.**
License to access the dataset	**N/A**
Additional comments	**N/A**

Table 8: SNR and RSSI from an urban LoRa Network

Name of the dataset	SNR and RSSI from an urban LoRa Network
Main Involved Partner	**CNIT**
Dataset Description	**The dataset contains various SNR and RSSI readings from a LoRa Network in the city of Portland, Maine.**
Type of personal data *(if relevant)*	**N/A**
Purpose for using/processing the dataset	**Using the "SNR and RSSI from an urban LoRa Network" dataset to analyse transmission performance and identify weak signal areas. This helps optimise network configuration to reduce power consumption, improving energy efficiency while maintaining service quality.**
Format(s) of the dataset	**The dataset consists of a single sheet file where each row corresponds to a specific LoRa node. Note that: i) the first row corresponds to the LoRa Gateway; ii) each row contains information on the node ID, its coordinates, the packet frequency, bandwidth, Transmission Power, Spreading Factor, Coding Rate, SNR and RSSI.**

Dataset storage	**Server available in CNIT laboratory.**
Dataset origin	**Experimental tests.**
Dataset ownership	**CNIT laboratory**
Restrictions for use	**N/A**
Dataset access	**CNIT laboratory**
Length of dataset storage	**Indefinite period.**
License to access the dataset	**N/A**
Additional comments	**N/A**

In addition, SoBigData, coordinated by CNR-ISTI, provides a rich catalogue of 227 datasets. The overview of all SoBigData's available datasets are publicly available here: https://ckan-sobigdata.d4science.org/organization/1630070e-3b22-4629-b49e-3ee4d291c9c5?systemtype=Dataset

2.3 FAIR data

This section describes the measures to ensure the data produced and used in the GreenDIGIT project adhere to the FAIR data principles of the European Commission. The FAIR data principles require that data are Findable, Accessible, Interoperable, and Reusable.

The GreenDIGIT project will manage data in terms of their level of accessibility. In particular, the project will manage:

- **Openly accessible data:** data that are generated within the GreenDIGIT project that will be made available by the end of the project duration, but also data that are already openly available and that can be reused during the project;
- **Confidential data** that can be accessed only by consortium partners;
- **Restricted data** that are accessible only by a restricted amount of people within the consortium.

GreenDIGIT will apply the FAIR data principles to all datasets generated and used in the project. A decision on the access level of each dataset will be finalized in the course of the project. The rest of this section will offer a plan for how each requirement of the FAIR principles will be addressed during GreenDIGIT. This is necessary to implement a data management strategy for gathering, keeping, conserving, and making data available for additional processing.

As it stands, assessment and optimization for energy, impact, and durability for connected data are not included in the current FAIR data management standards. Thus, they should be extended to include environmental impact-related metadata. Moreover, to combine the creation, implementation, and management of experimental data with scientific applications while maintaining FAIR data standards, Reproducibility as a Service (RaaS) needs to be established. Lastly, ensuring data is appropriately stored, tracked, and shared is crucial in guaranteeing its accessibility and suitability for

replication studies. Tools like Zenodo or DataCite can be used to issue unique identifiers and manage data citations, while OpenRefine and Dataverse can be used to clean, organize, and distribute data.

Moreover, it is imperative that the FAIR principles are followed by all organisational roles throughout all data lifecycle stages. When preparing to handle the data, partners can rely on the FAIR Maturity Model, created to establish readiness and FAIRness of the data. Accordingly, there are four stages for each indicator: (0) means it's not applicable; (1) means it's not being considered yet; (2) means it's being considered or is in the planning stage; (3) means it's undergoing implementation; and (4) means it's fully implemented.

The following sections will present each of the principles and provide the relevant individual partners' plans for those managing data with a need to implement the FAIR principles.

2.3.1 Making data Findable

Findability of data can be achieved through well-defined and published metadata, API for registries and handles resolution, and service level agreements (SLA) as well as policies. Additionally, findable data means that both computer programmes and humans can easily find and access the information. Pursuant to this and to FAIR Guiding Principles for scientific data management and stewardship,[3] the project should ensure the following elements are present:

1. Assignment of a globally unique and persistent identifier to (meta)data.
2. Description of data with rich metadata.
3. Explicit inclusion of an identifier of the data that the metadata is describing.
4. Registration or inclusion in index in a searchable resource of the (meta)data.

To ensure data resulting from research and development in GreenDIGIT is findable, the project plans to utilize various strategies and tools. Firstly, GreenDIGIT will publish research outputs, datasets and metadata in trusted certified data repositories, including Zenodo, OpenAIRE, arXiv, and national services, such as Recherche Data Gouv (France) and national EOSC nodes. Researchers can also use GitHub for storing both research and experiment information and data, involving community contribution. GreenDIGIT will investigate if the usage of RO-Crate is possible to make the result files more easily findable.

Datasets and research outputs will also be made available in the catalogues of participating RIs in GreenDIGIT. This includes the SLICES Metadata Registry Service (MRS)[4] and the SoBigData Catalogue[5], where resources can be searched using keywords and presented according to classification.

[3] 'FAIR Principles' (*GO-FAIR*) <https://www.go-fair.org/fair-principles/> accessed 24 July 2024

[4] 'SLICES-RI Metadata Registry System' (*SLICES-RI)* <https://www.slices-ri-eu/slices-ri-metadata-registry-system/> accessed 18 July 2024

[5] 'SoBigData Catalogue' (*SoBigData*) <https://sobigdata.d4science.org/catalogue-sobigdata> accessed 25 July 2024

Metadata and Persistent Identifiers (PID) will be generated for research data throughout the project. GreenDIGIT may use external metadata and PID registries and services (such as operated by EOSC, EGI, or national research infrastructures) and at a later stage create its own metadata and PID registries, which will be interoperable with external services. The future datasets and related metadata will also have unique identifiers, such as unique URLs. This allows to find and reference them in an efficient way. It is also expected that all the datasets and metadata will be searchable in the GreenDIGIT project website. Moreover, all datasets collected or generated within the GreenDIGIT project will be tagged with a set of pre-defined keywords.

The partners' commitment to guaranteeing data findability is summarized in the table below.

Table 9: Partners' strategies to make data findable

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met: Findable
CESNET	We currently don't have a clear plan to publish the datasets. It is to be decided if some testing or evaluating datasets regarding workload itself and workload manager behaviour will be published. If so, we expect to use the currently built Czech EOSC node to FAIRify and publish the data.
CNRS	Data will be made available through the Recherche Data Gouv data repository.
MI	The future datasets and related metadata will have unique identifiers, such as unique URLs. This allows to find them and to reference them in an efficient way. It is also expected that all the datasets and metadata will be searchable in the GreenDIGIT project.
PSNC	In order to make data findable, all datasets collected or generated within the GreenDIGIT project will be tagged with a set of pre-defined keywords.
SoBigData	The SoBigData Catalogue (https://sobigdata.d4science.org/catalogue-sobigdata) is the main application for finding and accessing all the datasets, methods, services, and publications. It contains all the resources, which may be accessed online or physically by visiting the resource's publisher. The catalogue is the primary application for discovering and searching for a dataset inside SoBigData RI. All the elements inside SoBigData RI are discoverable through this service. The user can insert a set of keywords, and the list of the results is visualized. The search result provides a list of items included in the catalogue and their classification (e.g., Method, Training Material, Dataset). The complete description is provided on the dedicated page, accessible by clicking on the item. These features can be added to the search filter, which will be recalculated in real-time. The search results can be sorted alphabetically according to the insertion date or popularity. The catalogue organizes products only by a set of defined categories that the user can navigate by selecting a specific link.
SZTAKI	This is subject to discussion with HUN-REN SZTAKI decision-makers. There are no technical obstacles to making data accessible and searchable with a certain time delay.

TUM	We plan to use the SLICES Metadata Registry Service (MRS) to make the data findable. We investigate if the usage of RO-Crate is possible to make the result files more easily findable.
UvA	Publishing in recognised data repositories, providing searchable/discoverable metadata, registering datasets and metadata with the recognised repositories such as Zenodo, OpenAIRE, arxiv, or national services. Researchers can also use Github for storing both research or experiment information and data. Recommended naming methods should be used to make research, project and data findable.

2.3.2 Making data Accessible

Accessibility is connected to infrastructures and management of data. For this reason, data is made accessible through repositories, implementing usage policies and access control, sovereignty, API management, data access protocols, and compliance with the GDPR to ensure data protection. Pursuant to the FAIR guidelines, the requirements to make data accessible are the following:

1. A standardized communications protocol makes (meta)data retrievable by their identifier.

2. It is freely available, open source, and easily implementable everywhere.

3. When required metadata are available, the protocol permits an authentication and authorization process to be carried out.

4. In the event that the data are no longer available, metadata can still be accessed.

To ensure the research data resulting from GreenDIGIT is accessible, the project will publish and share research outputs, datasets and metadata in trusted data repositories accessible for scientists and researchers, as detailed above with Open Science licences and IPR. In the case of publishing data on public repositories, technologies used by the repositories will allow the discoverability of published data and the appropriate access protocol to access and download data. Standardised protocols will be used to retrieve the data and their metadata. Access to the data and metadata can be limited to avoid unauthorized edition or deletion of datasets. It means that authentication and authorization mechanisms and procedures will be put in place as the data is ready to be shared.

In addition, data will be made accessible in the participating RIs' catalogues, on the project website (for at least five years after the project is concluded), and on GitHub. The SLICES-RI MRS endorses the FAIR principles and allows registered users to access data, services, and software by defining a metadata profile scheme for each digital object, coined the SLICES FAIR Digital Object. Registered users can easily access data in the registry through a web portal to search for objects and access reporting facilities. More information on the metadata scheme is available on the SLICES-RI website. On the other hand, SoBigData RI's Catalogue also promotes open science and is developed under FAIR principles. Three main fields defining the accessibility rules for a dataset on the SoBigData Catalogue include: **Accessibility** – defines how the access to the resource is regulated: Virtual Access,

TransNational Access[6], or Both; **Availability** – outlines how the availability to the resource is offered: on-line by e-infrastructure facilities, on-site by visiting the institution that is the data controller of the dataset; **Accessibility Mode** – describes the nature of the dataset and how access to the resource is implemented. Online access is typically used for connecting to the servers that provide data, e.g., a DBMS or API Access. In this case, the data are accessible by a server interrogable programmatically through API. Furthermore, all the datasets have an ethical and legal section of metadata that defines why the user cannot freely download a dataset or if a dataset can be accessed, e.g., considering the geographical restriction of using a particular dataset. Additionally, GreenDIGIT will investigate the possibility to use public platforms, like Zenodo, to crosslink experimental results with the participating RI's' catalogues.

Partners' plans to make data accessible is outlined in the table below.

Table 10: Partners' strategies to make data accessible

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met: Accessible
CESNET	We currently don't have a clear plan to publish the datasets. It Is to be decided If some testing or evaluating datasets regarding workload itself and workload manager behaviour will be published. If so, we expect to use the currently built Czech EOSC node to FAIRify and publish the data.
CNRS	Data will be made available through the Recherche Data Gouv data repository.
MI	Standardized protocols will be used to retrieve the data and their metadata. The access to the data and metadata can be potentially limited to avoid unauthorized edition or deletion of datasets. It means that authentication and authorization mechanisms and procedures should be put in place.
PSNC	Data collected or generated within the GreenDIGIT project will be published in open platforms, accessible for all scientists and researchers. Specific platforms to be used in the project will be selected during the project.
SoBigData	SoBigData RI promotes open science, and it is developed under FAIR principles. Three main fields define the accessibility rules for a dataset: Accessibility – defines how the access to the resource is regulated: Virtual Access, TransNational Access[7], or Both; Availability – outlines how the availability to the resource is offered: on-line by e-infrastructure facilities, on-site by visiting the institution that is the data controller of the dataset; AccessibilityMode - describes the nature of the dataset and how access to the resource is implemented. Online access is typically used for connecting to the servers that provide data, e.g., a DBMS or API Access. In this case, the data are accessible by a server interrogable programmatically through API. Currently, the RI does not provide any documentation on how to access the dataset via API or DBMS;

[6] ibid [4]

[7] 'Call 2024' (*SoBigData*) <http://sobigdata.eu/calls/transnational-access-2024> accessed 20 July 2024

	by the way, since SoBigData RI adopted large diffuse standards, a user can find all the required information for accessing the database. Furthermore, all the datasets have an ethical and legal section of metadata that defines why the user cannot freely download a dataset or if a dataset can be accessed, e.g., considering the geographical restriction of using a particular dataset.
SZTAKI	This is subject to discussion with HUN-REN SZTAKI decision-makers. There are no technical obstacles to making data accessible and searchable with a certain time delay.
TUM	We plan to use the SLICES MRS to provide data access. We investigate the possibility of using public platforms like Zenodo to crosslink the experimental results with the SLICES MRS.
UvA	In the case of publishing data on public repositories, it is the technologies used by those repositories to allow the discoverability of published data and the appropriate access protocol to access data and download them. Appropriate access control and registration of access and download data should be also Implemented. In the case of SURF Research Drive used by UvA researchers, two factor authentication is used and data access can also be configured by the project or researcher.

2.3.3 Making data Interoperable

Interoperability is a cornerstone to the realisation of the FAIR principles and for a FAIR Digital Object. To make data interoperable, standard data formats, API, FAIR maturity levels and repositories certifications should be present. Interoperability is presented in the FAIR Guidelines as:

1. (Meta)data represent knowledge in a formal, common, comprehensible, and widely applicable language.
2. (Meta)data employ FAIR-compliant vocabulary.
3. (Meta)data provide appropriate references to other (meta)data.

The interoperability of research data in GreenDIGIT will be ensured by using recognised, standardised and interoperable data formats for the publication of open research data. For most data in the project, GreenDIGIT aims to utilise formats commonly used in the domains of networking experiments, computer modelling, environment documenting, and energy consumption. For any custom data formats specific to the GreenDIGIT project, necessary description of the experiments, research or RI/facility operation, as well as corresponding metrics will be provided to allow the use of data by other researchers or for the possibility to be reused in further research or projects.

All datasets will be described with sufficient information in the metadata to make it possible to transform them to standardised categories. Data will be stored in open formats, including txt, json, xml, and csv. For the network-based experiments created in a testbed, GreenDIGIT will rely on well-established data formats, such as pcap or Jupyter to simplify the exchange and conversion of data.

The SoBigData RI has a specific 'workspace' service to make data interoperable and employable to other services, such as SoBigDataLab[8], or JupyterHub. The workspace is an online environment that supports secure and controlled data storage and sharing. Each Virtual Research Environment (VRE) has an associated workspace where users can store, access, and share documents and results related to the activities inside a specific gateway and the VRE. Moreover, each user has a private space where storing data and documents creates folders and a public space, one for each subscribed VRE where they share files. The SoBigDataLab provides many methods that can be selected and executed into the RI. The methods are performed by loading the input data into the user workspace. It is possible to execute a method only if the required input file is already present in the workspace, allowing inter-disciplinary interoperability.

Partners' commitments to make data interoperable are shown in the table below.

Table 11: Partners' strategies to make data interoperable

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met: Interoperable
CESNET	We currently don't have a clear plan to publish the datasets. It is to be decided if some testing or evaluating datasets regarding workload itself and workload manager behaviour will be published. If so, we expect to use the currently built Czech EOSC node to FAIRify and publish the data.
CNRS	Data will be stored in open format (txt, json, xml, csv, …) and they will be accompanied by metadata describing the format.
MI	Recognised, standardized and interoperable data formats will be used for the publication of the open research data
PSNC	In order to make data sets interoperable, all data sets will be described with sufficient information (meta data) to make it possible to transform them to standardized categories.
SoBigData	The SoBigData RI has a specific service to make data interoperable and employable to other services such as SoBigDataLab[9], or JupyterHub. This service is called workspace. The workspace is an online environment that supports secure and controlled data storage and sharing. Each Virtual Research Environment (VRE) has an associated workspace where users can store, access, and share documents and results related to the activities inside a specific gateway and the VRE. Moreover, each user has a private space where storing data and documents creates folders and a public space, one for each subscribed VRE where they share files. The SoBigDataLab provides many methods that can be selected and executed into the RI. The methods are performed by loading the input data into the user workspace. It is possible to

[8] ibid [6]

[9] 'SoBigData Lab' (*SoBigData*) <https://sobigdata.d4science.org/group/sobigdatalab/sobigdatalab> accessed 18 July 2024

	execute a method only if the required input file is already present in the workspace, allowing inter-disciplinary interoperability.
SZTAKI	The analysis of the data may reveal correlations that can be applied to other hardware infrastructures, but it is more likely that the correlations specific to the data and the infrastructure from which it was derived will be relevant only to that infrastructure. This cannot be definitively determined without deeper analysis.
TUM	For the network-based experiments created in our testbed we want to rely on well-established data formats (e.g. pcap or Jupyter) to simplify the exchange and conversion of data. RO-Crate metadata may be helpful to describe the data to make interoperability easier to achieve.
UvA	Interoperability is ensured by the researchers by using dataset formats commonly used in the domain of research. In particular case of the GreenDIGIT project, this is related to networking experiments, computer modelling, environment documenting, energy consumption. In case of custom data format that may be specific for the GreenDIGIT project, necessary description of the experiments, research or RI/facility operation and corresponding metrics must be provided to allow the use of data by other researchers or the possibility to be used in other research or projects.

2.3.4 Improving data Reuse

Crucial elements to successfully improve data reusability must account for provenance and lineage, reproducibility compliance, preservation, personal identifiable data (PID) and API embedded in datasets. Data can be reused when:

1. Richly specified (meta)data have many useful and accurate properties.
2. An understandable and accessible data usage license is provided with the release of (meta)data.
3. Detailed provenance is linked to (meta)data.
4. (Meta)data adhere to community norms pertinent to the domain.

In order to improve data reusability, data collected or generated within the GreenDIGIT project will be published in open platforms for potential reuse by the scientific community, as outlined in the above sections. The published data will be provided with sufficient metadata containing specific attributes dedicated to the reuse of the datasets and research usage guidelines.

According to national laws, specifically the French law No. 2016-1321 of October 7, 2016, for a Digital Republic, data will be released under an Open Licence (either based a Creative Common licence or Etalab License). Time constraints for the reuse of datasets are not planned. User acknowledgement of the usage of datasets will be mandatory. GreenDIGIT will also investigate the possibility of using RO-Crate to make experimental data more reusable.

For the SoBigData RI, the preservation and reusing procedures describe how partners of the RI store the data, which technology is used, and how long the data is available. SoBigData RI monitors that all the partners comply with privacy and licensing restrictions declared for their data and will take care of the costs associated with their long-term preservation. The available and newly gathered datasets are registered in the RI following the specification defined data management plan of the RI: **Publicly Available Data** - several datasets are made available by private and public entities and will be included in the RI resources. **Restricted data** - within the consortium, SoBigData will make available proprietary datasets. Due to the restriction imposed by data owners, such collections will be made available prevalently on-site, through Transnational Access. In such scenarios, the consortium will simplify the procedures needed to grant data access to the researchers who want to pursue experiments on them. The access via online services will be granted for all those datasets whose policies allow open diffusion; conversely, for all the datasets whose access is restricted due to licensing restrictions, access will be provided only through Transnational Access. Moreover, for some datasets, to avoid Term of Usage (ToS) infringements, access will be offered in the form of data crawlers to obtain data directly from the original source and for a specific and time-limited experiment (e.g., Twitter data).

Lastly, enhancing reproducibility may involve developing a platform that makes experiment sharing and reuse simple. This can be accomplished by sharing code, data, and documentation using tools like Jupyter Notebook and version control systems like Git or GitHub.

Partners' responses to the element of being reusable are displayed in the table below.

Table 12: Partners' strategies to make data reusable

Partner	Did you or will you be taking measures in order to comply with the FAIR data principles (making data Findable, Accessible, Interoperable and Reusable)? If so, kindly provide additional information on how each of these principles are being met: Reuseable
CESNET	We currently don't have a clear plan to publish the datasets. It Is to be decided If some testing or evaluating datasets regarding workload itself and workload manager behaviour will be published. If so, we expect to use the currently built Czech EOSC node to FAIRify and publish the data.
CNRS	According to the French law No. 2016-1321 of October 7, 2016, for a Digital Republic, data will be released under an Open Licence (either based a Creative Common licence or Etalab License).
MI	The metadata will contain specific attributes dedicated to the reuse of the datasets generated during the GreenDIGIT project.
PSNC	Data collected or generated within the GreenDIGIT project will be published in open platforms for potential reuse by the scientific community. Time constraints for reuse of data sets are not planned. User acknowledgment of the usage of data sets will be mandatory.
SoBigData	Dataset description includes a unique reference and an assessment of their nature, scale, and available metadata (such as related scientific publications, privacy issues, data governance policies, licensing, or similar resources). The preservation and reusing procedures describe how the partners store the data, which technology is

	used, and how long the data is available. SoBigData RI monitors that all the partners comply with privacy and licensing restrictions declared for their data and will take care of the costs associated with their long-term preservation. The available and newly gathered datasets are registered in the RI following the specification defined the data management plan of the RI: Publicly Available Data - several datasets are made available by private and public entities and will be included in the RI resources. The information about these datasets is not always complete or well described. For this reason, a platform aimed at quickly finding, annotating, and discussing them within the research community is needed. Restricted data - within the consortium, we will make available proprietary datasets. Due to the restriction imposed by data owners, such collections will be made available prevalently on-site, through Transnational Access. In such scenarios, the consortium will simplify the procedures needed to grant data access to the researchers who want to pursue experiments on them. The access through online services will be granted for all those datasets whose policies allow open diffusion; conversely, for all the data sets whose access is restricted due to licensing restrictions, access will be provided only through Transnational Access. Moreover, for some datasets, to avoid Term of Usage (ToS) infringements, access will be offered in the form of data crawlers to obtain data directly from the original source and for a specific and time-limited experiment (e.g., Twitter data).
SZTAKI	The analysis of the data may reveal correlations that can be applied to other hardware infrastructures, but it is more likely that the correlations specific to the data and the infrastructure from which it was derived will be relevant only to that infrastructure. This cannot be definitively determined without deeper analysis.
TUM	Investigate the possibility of using RO-Crate to make experimental data more reusable
UvA	This ensured by the researcher who published the research data and a researcher who will use this data. The published data must be provided with sufficient metadata description and research process and workflow guidelines.

3 Ethics

3.1 Overview of ethical themes

Several ethical themes can be outlined in the project, including data protection, technological development, operation of RIs, data storage and retention, purpose of data collection, additional processing, consent, and informing individuals.

The project must process personal, non-personal, and sensitive data in compliance with applicable EU and national laws on data protection (including authorisations or notification requirements). Pursuant to Article 15 of the Grant Agreement, the project commits to provide data protection under Regulation 2016/679. Data collected for the purposes of administering the project activities will be held securely according to legislation and will not be shared externally. Personal data should be collected only for what is necessary to fulfil the information needs of the project (respecting the principle of data minimisation). Protection of personal data includes the process of its *collection* (when the legal basis is consent, it should be ensured that users give their informed consent), *processing* (the data controller and processor should ensure that technical and organisational measures are in place to secure the data; the data subject should be able to practice their rights), and its *termination of processing* (the processing should be limited to what is strictly necessary, and the controller should employ measures such as minimisation, pseudonymisation, or anonymization and/or delete the data after it is no longer necessary for the project).

Additionally, one of the associated RIs, SoBigData, has indicated that processing of special categories of data might take place when commencing the project. The general rule about special categories of data is that processing is forbidden, unless certain conditions are met. Sensitive data may be processed, i.e., if there are special employment law protections in place or if the data subject has given their express consent. As sensitive data can raise ethical concerns about, i.e., consent and how it is obtained, if it is used for secondary processing, and if it is disposed of after it is no longer necessary, stricter rules on its usage should be in place. Data sensitivity classification is an important step of the data governance policy of the project. Some measures to be implemented so that the dataset is legally and ethically compliant may include restricted access to forms and records, as well as proper technical and organisational measures in place to secure the data, such as encryption. More rules on special categories of data can be found in Article 9 of the GDPR. Accordingly, several rules need to be followed:

- Pseudonymization (replacing the names with codes);
- Avoiding disclosing the sensitive data if there is other information that makes the individuals identifiable;
- Not leaving the documents unattended and keeping them separately from the rest of the data;
- Storing the data in places where they can be locked and secured.

In addition, the project will follow the principles for Open Science. These include open access, open data, open source, reproducibility of research, open educational resources, open notebooks, and

citizen science. GreenDIGIT commits to the principles and will use tools like Zenodo and GitHub, OpenAIRE and ORE.[10]

3.2 Personal data protection principles, rights of data subjects, obligations of controller and processor

3.2.1 Principles of data processing

Article 5 of the General Data Protection Regulation (GDPR) provides for the principles of processing, which every controller and processor should follow. These include:

- Lawfulness, fairness and transparency;
- Purpose limitation;
- Data minimisation;
- Accuracy;
- Storage limitation;
- Integrity and confidentiality;
- Accountability;
- Protection by design and by default.

In addition to these, the controller should decide on a lawful basis for processing, which is outlined in Article 6 of the GDPR. These include:

- Consent;
- Necessary for performance of contract;
- Necessary for compliance with legal obligation;
- Necessary for protection of vital interests of data subject or other individuals;
- Necessary for performance of a task in the public interest;
- Legitimate interest of the controller.

3.2.2 Controller and Processor

Throughout the project, partners may serve both as data processors and as data controllers, whether or not they do so together. This section offers direction on the primary responsibilities that come with taking on either of these roles. Their responsibilities are outlined in Articles 24-43 of the GDPR. Accordingly, controllers have the following obligations:

[10] GreenDIGIT Proposal

- Implementation of appropriate technical and organisational measures (TOMs) and data protection policies. These can be in the form of pseudonymization, data minimization, safeguards.
- Ensuring protection by default and by design.
- Providing proof that the processing is compliant with the GDPR. This can be done through codes of conduct or certification.
- Allocating tasks to processors who apply the appropriate TOMs.

The processor, on the other hand, is also tasked with responsibilities, which include:

- Asking for authorization from controller if they want to involve another processor in the process.
- Have a contractual relationship with the controller, which stipulates all necessary conditions for the processing.
- Adhering to codes of conduct and certification, which can serve as a proof for compliance with the GDPR.
- Respecting confidentiality.
- Deleting or returning data after the contractual relationship with the controller comes to an end.

Both the controller and the processor should keep documentation (records of processing activities) and present it when necessary. In addition, both should cooperate with supervisory authorities and ensure that TOMs provide a sufficient level of security of processing through the implementation of different techniques. In case of a data breach, both should inform the relevant person and authorities, including the data subject if the conditions are fulfilled.

The datasets, as well as the data, must be properly documented to ensure that they are also easily understandable by the user. The information that needs to be included in the documentation is:

- Contextual data and project information: context, past project history, goals, objectives, and hypotheses; publications derived from data gathering
- Techniques and procedures for gathering data: sampling and data collection; tools employed, such as questionnaires, showcards, and interview schedules; compilation of derived variables; temporal/geographic coverage; data validation: cleaning, error-checking; variable labels in a table (which are usually classified as metadata); weighting: factors and variables, weighting procedure; secondary data sources used
- Conditions for data confidentiality, access, and usage: the process of anonymization; the terms and procedures for consent (such as the ethical consent form); and the conditions for data access or usage

3.2.3 Data subject rights

While the processor and the controller have obligations to ensure that the data processing is lawful and follows several principles, they also have obligations towards the data subjects under their requests. For a starter, the controller has the obligation to provide the data subject with all the necessary information about the processing before the processing commence. This includes a description of the processing with purpose, contact, duration, as well as the rights of the data subject. Moreover, as outlined in Articles 12-23 of the GDPR, the data subject has the right to:

- Request information about their processed data;
- Request access to the processed data;
- Request that their data is rectified when the conditions are met;
- Request that their data is erased when the conditions are met ('right to be forgotten');
- Request that their data is portable so they can move it to another provider;
- Object to automated decision-making, including profiling.

3.2.4 Data impact assessment

As displayed in the tables above, several of the partners are foreseen to use datasets with personal data: CESNET, UTH, SU, SoBigData, and EGI. This section will present further information on the processing of personal data, as well as implemented safeguards.

When personal data is processed, it should be ensured that this is done in a proper manner with sufficient safeguards. Two important factors for this include the appointment of a data protection officer (DPO) and the conduct of data impact assessments (DPIA) if necessary. When the processing is 'likely to result in a high risk to the rights and freedoms of natural persons,' the DPIA is necessary, as provided in Article 35(1) of the GDPR. As one of the partners, SoBigData, has outlined the possibility for special categories of data to be processed, DPIA will have to be conducted to determine if the processing presents any risks. Although the rights to privacy and data protection are the main concerns of data subjects, other fundamental rights, such as the freedoms of speech, opinion, and movement, the outlawing of discrimination, and the right to liberty, conscience, and religion, may also be included. Some situations when a DPIA may be necessary include when the processing involves new technologies, systematic processing of large quantities of personal data, processing of personal data, monitoring public spaces, etc. A DPO plays an important role in the conduct of a DPIA and is required to be appointed if certain elements are present: (1) the processing requires systematic monitoring; (2) the processing is done by a public body; (3) special categories of data are processed.

In order to determine whether the processing in GreenDIGIT presents a risk to the rights and freedoms of natural persons, partners were presented with questions on the processing of personal data in the Data Management Survey. Their answers are recorded and displayed in the form of a table.

Table 13. Overview of personal datasets

Partner	For what purpose(s) did/will you collect the aforementioned personal data?
SU	Personal data will be collected to analyse the results of the "Survey of digital Research Infrastructures practices and needs for environmental sustainability and impact reduction" (WP3). The results of the survey will be shared with other ESFRI RIs. The personal data collected (name, email and phone number) will not be disclosed. They will be collected only to facilitate the interview process (once the survey is completed) and be kept by SU's team. When relevant and if the reply is considered a good practice, the name of the RI may be disclosed in the corresponding survey report.
UTH	Participant naming lists for project events. Dissemination material (e.g. photos) also may be collected and posted on the social media pages / website of the GreenDIGIT project.
EGI	Personal data is needed to be able to follow up with individuals for clarifications about discussed matters.
SoBigData	Due to the nature of our research infrastructure, e.g., social mining studies, artificial intelligence, for our research areas we must also collect datasets representing human behaviour that can include personal data.
Partner	**Did/will you process the generated data for any further purposes than the ones it was originally collected for?**
SU	No
UTH	No
EGI	No
SoBigData	Yes
Partner	**If you answered yes to the previous question, then please describe the purpose of this additional processing.**
SU	N/A
UTH	N/A
EGI	N/A
SoBigData	In many cases, due to the multidisciplinary aspects of the data science and social mining we use dataset, or integrate more than one, to make research that are different from the original purposes related to data collection.
Partner	**How did/will you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?**
SU	A statement has been included in the survey. Personal data will be stored by SU's team and will not be disclosed. The personal data will be used by SU's team during the interview to facilitate the contact with the RIs. When relevant, reference to a specific RI may be added in the survey report, but only if it is considered a good practice. A statement has been added on the survey to clarify it.
UTH	Through consent / acceptance of data usage for the purposes of the organized events / dissemination of the project.

EGI	They are informed via the material accompanying the survey and the 1-to-12 interviews.
SoBigData	Now, we do not directly manage the dataset collection but only indirectly through partners of the project, that we use as intermediaries to communicate the use of the personal data inside the RI to the subject.
Partner	**How did/will you plan to collect and document the consent of the data subjects whose personal data will be processed by you?**
SU	The following text has been added to the survey. We consider that by answering the survey, the respondent gives his/her to the collection of his/her personal data. He/she can, at any time, object by contacting survey@greendigit-project.eu: *The data controller for this survey is Sorbonne Université, located at 21, rue de l'école de médecine 75006 Paris – France.* *According to Article 4.1 of the GDPR, personal data is any information that allows you to be identified either directly or indirectly. This survey collects personal data, considered adequate, useful, necessary and limited to the strict minimum. In this case, personal data will be used only to do interviews directly related to your replies to this survey.* *Your personal data is primarily intended for Sorbonne Université. It is processed by the Sorbonne Université staff and the relevant partners of the GreenDIGIT project. This project has received funding from the European Union's HE research and innovation programme under Grant Agreement No. 101131207.* *It is processed solely for the purposes indicated above. Your personal data is hosted exclusively in France and the Netherlands by subcontractors of Sorbonne Université and Universteit Van Amsterdam (as GreenDIGIT coordinator) under strict supervision. They are never communicated to third parties for commercial purposes. Personal data collected by this survey is not transferred to parties outside the European Union. In the event of a contractual transfer of personal data, Sorbonne Université ensures that the recipients are based in the European Union.* *By answering this survey, you consent to the collection of your personal data. You can object to the collection of your personal data, by contacting survey@greendigit-project.eu.*
UTH	Through consent / acceptance of data usage for the purposes of the organized events / dissemination of the project.
EGI	The consent is requested at the time of completing the survey, also during the 1-to-1 interviews.
SoBigData	Now, due to legal aspects (the RI is not a legal entity) the documents are collected by the beneficiaries that originally collected the data.
Partner	**How and where did/will you store the data?**
SU	We used Lime Survey to develop the survey, as recommended SU's IT security department. Once the survey has been done, the results of the survey will be analysed and stored locally on ownCloud (hosted by SU and stored in a secure

	location). A copy of the survey's outcome will be with the partners through UvA's shared folder.
UTH	If the data require storage, the shared folders of the UvA surfnet will be used.
EGI	Data will be stored on UvA premises (UvA Research drive tool).
SoBigData	The data are stored on SoBigData RI data centre powered by D4Science[11]. The data are replicated and distributed along different nodes located in Italy. In any case, SoBigData RI ensures privacy and data integrity between two communicating computer applications: any connections between a client (e.g., a web browser) and a SoBigData RI server have the following properties: (i) the connection is private (or secure) thanks to the adoption of symmetric cryptography to encrypt the data transmitted; (ii) the identity of the communicating parties can be authenticated using public-key cryptography. Data are encrypted and the SoBigData RI authorization is empowered by a token-based authorization system compliant with the Attribute-based access control (ABAC) that defines an access control paradigm whereby access rights are granted to users using policies that are validated in a VRE context.
Partner	**For how long did/will you keep the data?**
SU	When personal data is considered necessary to prove the proper implementation of the project (e.g., WP3 survey) and if it is not possible to anonymise the data, SU will keep a record of this during a period of 5 years after the end of the project. This follows the requirement of the Grant Agreement (Article 18.1). If a personal data is not considered essential and useful to prove the proper project implementation, it will be deleted as soon as it is no longer necessary and required.
UTH	Until project ends (UvA storage is active).
EGI	N/A
SoBigData	Each dataset has its own field "Retention Period" where it is explained the dataset availability in time.

In addition to the above illustrated answers, EBRAINS also provided a response to describe how and where they store their non-personal data, which should also be considered:

The Project Deliverables (that constitute the data that is the subject of this document) will be stored in the central Project Repository, in Microsoft SharePoint. This cloud-based storage is based on a scalable storage approach. As such, no limitation is foreseen linked to a shortage of storage capacity. Storage is in several geographically distributed data centres, thereby ensuring redundancy. Adequate Disaster Recovery Plans have been put in place to ensure data continuity in the unlikely event of a data loss issue. Access and security are core parts of the technical solution put in place for the Project Repository which is implemented using Microsoft SharePoint. As such, the technical implementation of the security layer is based on Azure Active Directory. The actual access policy applied for EBRAINS GreenDIGIT will follow a strict information security classification framework. The EBRAINS GreenDIGIT project information classification scheme requires data to be attributed with one of three classifications

[11] 'D4SCIENCE' (*D4SCIENCE*) <https://www.d4science.org/> accessed 26 July 2024

(excluding public information, which does not need to be marked). The way the information is handled, published, moved, and stored will be dependent on the classification level assigned. The classification levels are:

- *Level 0: Public (or unclassified): Much of the information held by the organisation is freely available to the public via established publication methods. Such information items have no classification and will not be assigned a formal owner or inventoried.*
- *Level 1: Protected: For information not published freely by the organisation, some of this may be classified as Protected. This is typically information that is relatively private in nature, either to an individual or to the organisation; whilst its loss or disclosure is unlikely to result in significant consequences, it would be undesirable. The criteria for assessing whether the information would be classified as Protected include whether its unauthorised disclosure would:*
    - *cause distress to individuals,*
    - *breach of proper undertakings to maintain the confidence of information provided by third parties,*
    - *breach of statutory restrictions on the disclosure of information,*
    - *cause financial loss or loss of earning potential, or facilitate improper gain,*
    - *give an unfair advantage to individuals or companies,*
    - *prejudice in the investigation or facilitating the commission of a crime,*
    - *disadvantage the organisation in commercial or policy negotiations with others.*

  Most EBRAINS GreenDIGIT project staff are likely to handle "Protected" information during their working day.
- *Level 2: Restricted: The level above Protected is Restricted. This information would be more serious if it were disclosed to unauthorised persons and result in significant embarrassment to the organisation and possibly legal consequences. The criteria for assessing whether the information would be classified as Restricted include whether its unauthorised disclosure would:*
    - *affect relations with other organisations adversely,*
    - *cause substantial distress to individuals,*
    - *cause financial loss or loss of earning potential or facilitate improper gain or advantage for individuals or companies,*
    - *prejudice in the investigation or facilitating the commission of a crime,*
    - *breach of proper undertakings to maintain the confidence of information provided by third parties,*
    - *impede the effective development or operation of organisational policies,*
    - *breach of statutory restrictions on disclosure of information,*
    - *disadvantage the organisation in commercial or policy negotiations with others,*
    - *undermine the proper management of the organisation and its operations.*

  Information falling into the classification of "Restricted" or "Internal" will typically be handled by Work package leaders and above, with some employees of lower clearance being given access only in specific circumstances.
- *Level 3: Confidential: The highest level of classification is that of Confidential. This is reserved for information which is highly sensitive and would cause major reputation and financial loss if*

it were lost or wrongly disclosed. The criteria for assessing whether the information would be classified as Confidential include whether its unauthorised disclosure would:

- *materially damage relations with other organisations (i.e., cause formal protest or other sanction),*
- *prejudice of individual security or liberty,*
- *cause damage to the operational effectiveness or security of the EBRAINS GreenDIGIT partners,*
- *work substantially against organisational finances or economic and commercial interests of the EBRAINS GreenDIGIT partners,*
- *substantially undermine the financial viability of the EBRAINS GreenDIGIT partners and major organisations,*
- *impede the investigation or facilitate the commission of a serious crime,*
- *impede seriously the development or operation of organisational policies,*
- *shut down or otherwise substantially disrupt significant business operations.*

Access to information assets defined as "Confidential" will be tightly controlled by the Project Partner Assembly of EBRAINS GreenDIGIT, and in many cases, numbered copies of documents will be distributed according to specific procedures. When deciding which classification to use for an information asset, it is recommended that an assessment is conducted to consider the likely impact if the asset were to be compromised.

A correct classification will ensure that only genuinely sensitive information is subject to additional controls. The following points should be considered when assessing the classification to use:

- *Applying too high a classification can inhibit access, lead to unnecessary and expensive protective controls, and impair the efficiency of the organisation's business.*
- *Applying too low a classification may lead to damaging consequences and compromising of the asset.*
- *The compromising of larger sets of information of the same classification is likely to have a higher impact (particularly concerning personal data) than that of a single instance.*

Generally, this will not result in a higher classification but may require additional handling arrangements. However, if the accumulation of that data results in a more sensitive asset being created, then a higher classification should be considered.

When it comes to data management, relevant regulations to the project include the GDPR, the AI Act, the ePrivacy Directive, the Data Governance Act, Data Act, NIS Directives, and Open Data Directive.

3.3 Intellectual Property Rights (IPR)

The Consortium Agreement outlines the project's intellectual property rights, which are approved by each project partner.

While most of the partners reported in the Data Management Survey that their specific plans for the project at this stage will not give rise to IPR, some partners offered initial insights. CNRS stated that training materials and software will be made available under CNRS copyright. They are going to be

distributed in accordance with the FAIR principles under an Open License. The DIRAC source code, for instance, is available under the GPLv3.

SoBigData offered an elaborative initial overview of how IPR will be involved in their plans:

SoBigData considers the aspects of Intellectual Property (IP), associated with the data collected within the RI and sharing such data. The SoBigData RI is aimed as the European Research Infrastructure for Big Data and Social Mining and accumulates various datasets from different sources, including social media content (like tweets, blogs, etc), call graphs from mobile phone call data, networks crawled from many online social networks, including Facebook and Flickr, etc. Collecting, using and sharing of such data raise questions about intellectual property and data ownership. Starting from 2015, we defined the types of protection, which may be applicable to the datasets in the SoBigData RI, described the legal requirements for protection, which the datasets need to satisfy to be protected, cover the IP issues and legal requirements for sharing the datasets, which have already been accumulated in the project and consider the issues of data ownership. In the centre of consideration are such aspects as: whether social media content can achieve IP protection ("high level creativity"), what type of protection and what legal requirements for protection apply; the conditions under which such IP protected content can be shared in a (virtual) research environment for scientific purposes; the ownership of new data and personal information generated through the analysis of social media, sui generis protection of databases. The basic standards of IP and copyright law that provide protection and grant exclusive rights to the works in return for intellectual input and/or economic investment of owned resources, energy or assets (e.g. algorithms, software, models) are considered and balanced against the basic ideas of data protection and privacy principles.

Within the SoBigData RI, the consortium applied the EU ethical framework in terms of legal compliance (e.g., with respect to the GDPR) and the ethical framework that we built on top of the law. Generally, the RI collects and makes available vast amounts of data for research brings, apart from privacy concerns, also IP issues into play. Significantly, the terms of licensing and copyrights matters are highly relevant. First, several datasets even though publicly available can be copyrighted, e.g. social media content, like Twitter and Facebook, Flickr images, blogs, posts, etc. Second, the use of data items, even though publicly available, is - as a rule - subject to certain license terms, such as Creative Commons (CC) licenses, etc. The reproduction, making available, modification of copyrighted content qualifies as copyright relevant actions and, in principle, require authorization by the right holder, unless exceptions apply. Against this, SoBigData needs a solid IP management module to ensure the collection, use, sharing and making the data resources available for research are done in a copyright compliant way. To this end, the RI aims to build upon the IP management module developed in SoBigData, which works on three pillars: legal, technical and educational. From the legal side, the use of data resources is governed by the RI terms of use, which refer the use of individual data items to the individual license terms. The basic rights and license terms are attached and communicated via metadata. The licensing and access restrictions are - to reasonable extent - enforced by the technical infrastructure.

The legal framework has been translated into concrete implementations (Forgó et al. 2020), such as compliance with intellectual property rights via SoBigData Gateway Terms of Use[12]. More details about

[12] 'SoBigData Catalogue' (*SoBigData*) <https://sobigdata.d4science.org/catalogue-sobigdata> accessed 25 July 2024

intellectual property management can be found in D2.8 IP principles and business models 2[13] and D2.9 IP principles and business models 3[14].

[13] SoBigData, D2.8 (2017), *IP Principles and Business Models 2*
[14] SoBigData, D2.9 (2018), *IP Principles and Business Models 3*

4 Data security

Data security is dependent on the TOMs, which each of the partners have undertaken. This can be found in Article 32(1) of the GDPR which stipulates that the data controller and the data processor must put the required organizational and technical safeguards in place to guarantee a level of security appropriate to the risk, the state of the art, the implementation costs, and the nature, scope, context, and purposes of processing. The article outlines several ways to do so: (1) through pseudonymization and encryption of personal data; (2) through continuous availability, resilience, confidentiality, and integrity of processing systems and services; (3) through the prompt restoration of personal data availability and access, if there is a technical or physical issue; (4) through assessing, testing, and reviewing organizational and technical safeguards for processing security on a regular basis.

In practice, there are multiple measures that can be taken to ensure data security on an individual and organizational level. As explained in previous sections, data security should be present at all stages of the project, especially when special categories of data are processed. These include: technical protection (usage of passwords, power surge protection, firewall protection on all devices and restricted access); data destruction which ensures that the data is not retrievable (overwriting with random data, using special tools for recycling); ensuring back-up storage and back-up strategy are in place (this can be done locally or using cloud services); checking the files regularly and using them in a consistent and transparent manner; implementing data retention policies.

Data storage of internal GreenDIGIT project files is stored on the SURF Research Drive used by University of Amsterdam researchers and GreenDIGIT partners. To keep data secure, two-factor authentication is used across all users of the platform. Data access can also be configured by the project or researcher.

5 AI usage practices

Artificial Intelligence (AI) technologies, methodologies, tools, or algorithms may be used by some partners during the GreenDIGIT project. In the case of AI usage, the purpose and application of the tool should be reported, together with an overview of the data processed, and measures taken to ensure the ethical use and compliance with relevant data protection regulations. By maintaining a responsible use of AI, researchers can harness the power of AI while upholding the standards of ethical conduct and data protection.

During the first stage of the project, no AI use has been implemented. However, at least six GreenDIGIT partners expect to use AI practices in the future stages of the project, however with no binding decisions. Some partners plan to use AI for technical data analytics and to develop machine learning models and frameworks to conduct research on energy and environment optimization. The goal of AI usage is to progress with the complexity of intended models, including multiple factors, metrics, and data. Machine learning models will be trained on extracted data without the use of third-party API services.

At least one partner will use AI technologies in the form of statistical learning approaches, including classification and regression tools to automatically classify existing data or predict new data. The data processed by such tools may contain technical information regarding the performance and properties of ICT systems, obtained through isolated experiments and thus not contain personal information regarding physical persons.

The SoBigData RI will include the use, development, and research of AI technologies as part of the GreenDIGIT project and as part of the RI's research fields related to ethical AI use. The RI presents a white paper, research field related to AI explainability, as well as an ethical MOOC on the topics of ethical use of AI.

Some further forms of AI technologies are under consideration for design and development that may be collected in the SLICES infrastructure. This data will mainly consist of packet captures (pcap) files, measurements related to the network utilization (both control and user plane), and logs from the different service parts of the post-5G blueprint experiments. The measurements will be hosted within the SLICES-RI storage, and provided to external users, described through a detailed SLICES metadata model. The table below includes indicative data examples that may be generated at the SLICES-Gr Node.

Table 14: Indicative data examples generated at SLICES for AI practice use

Dataset Name	Dataset Description/ Purpose of use	Experiment / Test Title (as per the DoA)	Format/Type (.csv,. json)	Size	Open Access
Radio Measurements	RSRP, RSSI measurements from the wireless radio	Distributed Compute Continuum tests among SLICES nodes	.csv	Few KBs	Y

Control and Data Plane captures	Packet captures during the operation of the network	Distributed Compute Continuum tests among SLICES nodes	.csv	Order of 100 MBs	Y
Measurements from the different network services	Measurements on the network and other resource usage of the services deployed for the cloud-continuum use case	Distributed Compute Continuum tests among SLICES nodes	.csv	Order of a few KBs	Y
Logs from the different network services	Logs on the operation of the different services within the network	Distributed Compute Continuum tests among SLICES nodes	.txt	Order of a few MBs	Y

The data will be annotated with the SLICES Metadata model, part of SLICES Data Management Plan, and will be stored within the SLICES-RI Data Management Infrastructure (DMI). The SLICES metadata model has been designed according to the FAIR principles, and complies with them, making the metadata interoperable with services provided by the European Open Science Cloud (EOSC). The data generated over the SLICES testbeds comply with the SLICES Data Management Infrastructure (DMI). DMI has several locations for storage, all managed under the central authority of the SLICES central hub (SLICES-FR). The central storage, along with the policies for accessing and managing the data will be developed within 2025. Currently, there is no organised access control and policy enforcement for the data generated over SLICES. The experimenters authenticate with the SLICES portal and get access to the data that is collected over the testbeds. During 2025, the legal structure of SLICES is expected to be finalised, and therefore, all the access control policies and procedures will be set in place.

For the partners who do not currently plan to use AI or software components using AI, a complete analysis will be conducted in the future if the plan changes or if AI becomes used. For the partners who do plan to use AI, a revised analysis will be conducted to report on the usage.

6 Conclusion and outlook

The present Data Management Plan introduced the initial developments and projections for the processing and/or collection of non-personal, personal, and special categories of data. It did so by presenting partners' answers to a Data Management Survey (see Annex A), which was created to give a direct overview of their datasets. In the survey, partners had to present and detail the data they were planning to collect/process, as each type of data used should be supported by different measures, safeguards, and obligations to the controller/processors depending on whether the data is personal or not.

As a next step, this initial version of the DMP outlined a plan for GreenDIGIT partners to ensure and promote the FAIR and Open Science principles. Partners' inputs on individual plans to follow FAIR principles were also collected and outlined.

In addition to the FAIR principles, the document touched upon ethical and legal aspects that partners need to account for when working with the data. Multiple European and international standards, regulations and directives, which will be further elaborated in the final version of the deliverable, can help partners attune their data management and points of action.

Since the project will make use of AI technologies, special categories of data and possibly IPR, technical and organisational measures are necessary to implement to ensure the safety of the data. The final version of the deliverable will present how it will be managed with more details as the usage of these aspects is still uncertain at the current stage.

The GreenDIGIT will adopt a new recommended approach of relying on the currently well-established research data management practices among ESFRI RIs and EOSC community, which is in many cases is supported by the organisational role of the Data Steward (i.e. Data Management Officer). Resources and staff efforts for data management and quality assurance will be carried out by partners as a part of their research process organisation. Coordination and monitoring of the research data management will be done by WP1 via Executive Board activity and Work Packages coordination.

As the work on the project continues, different gaps between legislation, the European Green Deal and the project objectives continue to emerge. For this reason, when additional details regarding the partners' exact workplans become available, the data included in this deliverable will be further enhanced and supplemented by subsequent iterations. Updates to the data management of partners will be collected using the same survey format found in Annex A.

7 References

'Call 2024' (SoBigData) <http://sobigdata.eu/calls/transnational-access-2024> accessed 20 July 2024

'D4SCIENCE' (D4SCIENCE) <https://www.d4science.org/> accessed 26 July 2024

'Delivering the European Green Deal' (European Commission) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_en#transforming-our-economy-and-societies> accessed 29 July 2024

'European Data Strategy' (European Commission) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en> accessed 24 July 2024

'European Green Deal' (European Commission) <https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en> accessed 24 July

'FAIR Principles' (GO-FAIR) <https://www.go-fair.org/fair-principles/> accessed 24 July 2024

GreenDIGIT Proposal

'SLICES-RI Metadata Registry System' (SLICES-RI) <https://www.slices-ri-eu/slices-ri-metadata-registry-system/> accessed 18 July 2024

'SoBigData Catalogue' (SoBigData) <https://sobigdata.d4science.org/catalogue-sobigdata> accessed 25 July 2024

SoBigData Lab' (SoBigData) <https://sobigdata.d4science.org/group/sobigdatalab/sobigdatalab> accessed 18 July 2024

SoBigData, D2.8 (2017), IP Principles and Business Models 2

SoBigData, D2.9 (2018), IP Principles and Business Models 3

Annex A – Data Protection Coordination and Monitoring Survey

This annex presents the template of the data management survey sent to the partners of GreenDIGIT to effectively follow data management measures for reporting and monitoring in accordance with the objectives of WP1.

General Rationale and Instructions

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

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

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

Partner Organisation

Name:
Address:
Country:
Website:
Privacy policy webpage:

Organizational Contact Person

Name:
Email address:
Phone number:

Organizational Data Protection Officer

Name:
Email address:
Phone number:

Data Processing Activities

1. **Indicate what categories of data you collect(ed) or process(ed) in the context of the project:**

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

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

How did/will you collect these personal data?
☐ directly from data subjects who belong to your research team
☐ directly from data subjects outside your research team (i.e. early adopters, beta testers, etc.)
☐ indirectly through partners of the project
☐ indirectly through other organizations external to the project
☐ N/A (you can skip the last section of this document)

Data Management

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

	Please provide your answers in this column:
Name of the used dataset(s)	
Short description of the dataset(s)	
If the dataset includes personal data, please specify the type of personal data.	
Purpose for which you use/ process the dataset(s)	
Format(s) of dataset(s)	

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

FAIR data

4. Did you or will you be taking measures to comply with the FAIR data principles (making data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable)? If so, **kindly provide additional information on how each of these principles are being met:**

a) Findable
b) Accessible
c) Interoperable
d) **Reusable**

Intellectual Property Rights

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

Ethics and Personal Data Protection

6. **For what purpose(s) did/will you collect the aforementioned personal data?**

7. **Did/will you process the generated data for any further purposes than the ones it was originally collected for? ☐Yes ☐No**

8. **If you answered yes to the previous question, then please describe the purpose of this additional processing:**

9. **How did/will you inform the individuals (the data subjects) about the purpose of the data processing of their personal data in the project?**

10. **How did/will you plan to collect and document the consent of the data subjects whose personal data will be processed by you?**

11. **How and where did/will you store the data?**

12. **For how long did/will you keep the data?**

AI Usage Practices

13. **Did your organization use or plan to use any Artificial Intelligence (AI) technologies or methodologies as part of the project?** If so, please describe the specific AI tools or algorithms employed, the purpose and application of AI in your research activities, the types of data processed by these AI systems, and measures taken to ensure ethical use and compliance with relevant data protection regulations.