interTwin

D1.2 First Data Management Plan

Status: FINAL
Dissemination Level: public

Funded by the European Union

Disclaimer: Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them

Abstract

Key Words Data Management, FAIR, GDPR, meta data, re-usable data

This first Data Management Plan (DMP) introduces a report that specifies how research data will be collected, processed, monitored, and catalogued, following the FAIR principles. This deliverable is viewed as a living report that will advance throughout the life of the project.

Document Description

D1.2 First Data Management Plan

Work Package number WP1

Document type	Deliverable		
Document status	FINAL	**Version**	1
Dissemination Level	Public		
Copyright Status	This material by Parties of the interTwin Consortium is licensed under a Creative Commons Attribution 4.0 International License.		
Lead Partner	EGI.eu		
Document link	**https://documents.egi.eu/document/3918**		
DOI	**https://doi.org/10.5281/zenodo.8036955**		
Author(s)	• Sjomara Specht (EGI)		
Reviewers	• Donatello Elia (CMCC) • Diana Gudu (KIT)		
Moderated by:	• Sjomara Specht (EGI)		
Approved by	Malgorzata Krakowian (EGI) on behalf of AMB		

Revision History

Version	Date	Description	Contributors
V0.1	10/11/2022	ToC	Sjomara Specht (EGI.eu)
V0.2	12/01/2023	Initial First input	Sjomara Specht (EGI.eu)
V0.3	01/05/2023	Incorporated input from WP leaders and Use Case representatives	Malgorzata Krakowian (EGI.eu) Xavier Salazar (EGI.eu) Andrea Manzi (EGI.eu) Levente Farkas (EGI.eu) Charis Chatzikyriakou (EODC) Kalliopi Tsolaki (CERN) Sara Vallero (INFN) Yurii Pidopryhora (MPG) Donatello Elia (CMCC) Matthias Schramm (TU Wien) Christian Pagé (CERFACS)
V0.4	22/05/2023	Document ready for review	
V0.5	26//05/2023	Feedback external review	Donatello Elia (CMCC) Diana Gudu (KIT)
V0.6	08/06/2023	Incorporated feedback and finalized deliverable	Sjomara Specht (EGI.eu)
V0.7	13/06/2023	Approved by AMB	Malgorzata Krakowian (EGI.eu)
V1.0	13/06/2023	**Final**	

Terminology / Acronyms

Term/Acronym	Definition
AI4EU	Artificial Intelligence for Europe
CREM	Climate Research and Environmental Monitoring
DMP	Data Management Plan
DOI	Digital Object Identifier
DT	Digital Twin
DTE	Digital Twin Engine
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findability, Accessibility, Interoperability, Reusability
GDPR	EU General Data Protection Regulation
GW	Gravitational-wave Astrophysics
HEP	High Energy Physics
IP	Intellectual Property
JLA	Joinup Licensing Assistant
RA	Radio Astronomy
WP	Work Package

Terminology / Acronyms: **https://confluence.egi.eu/display/EGIG**

Table of Contents

Table of Tables

Executive summary

The deliverable D1.2 – Data Management Plan – defines the structure within which interTwin will create, manage, and collect data during the project's operations. Furthermore, it specifies how the data will be used or made available for verification and re-use, as well as how the data will be curated and stored once the project is completed.

In addition, the Data Management Plan (DMP) outlines the FAIR design of the data, agreements on data security are created, and ethical issues associated to data collection/generation are addressed.

interTwin adheres to Horizon Europe Open Science FAIR[1] principles and strives to make data as open and as closed as appropriate. The beneficiaries share the project's data in such a way that it is valuable to partners and their users outside the project, while ensuring that the privacy of third parties that participated in the data collection/generation is not violated.

Under these conditions, the data will be managed and released in compliance with the certifications and safeguards of the EU General Data Protection Regulation (GDPR[2]). Every dataset is examined (in terms of sensitivity, privacy, and security) before an official decision is made on whether or not to make that specific information public.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

[2] EU General Data Protection Regulation (GDPR)

1 Introduction

The current initial version is composed of preliminary information and frameworks that will be followed. Hence, it is subject to updates in the future upon developments and changes during the project. In principle, the DMP describes the standards that will be used, how the project's research data will be stored and published for verification and reuse. interTwin aims for full open access to data, following the FAIR principles.

1.1 Purpose and scope of the document

Open Science practices are implemented as an integral part of the proposed methodology, and can be described as follows.

Firstly, the use case methodology involves open cooperative work for **co-design** based on the sharing of knowledge and relevant software solutions across scientific communities and IT experts. Cooperation with Destination Earth[3] (DestinE) was initiated from M01. Consultations will be run online and open to all interested knowledge actors including Industry, SMEs, and evidence-based policy makers and civil society.

Secondly, co-design and validation are supported by making key exploitable results such as the software and the related documentation, the interoperability framework and requirements open source and open for exploitation under licenses listed as free by the Free Software Foundation and listed as "open source" by the Open-Source Initiative. All project communication and dissemination outputs will be available under an open license (Creative Commons[4]) allowing maximum uptake and reuse. Peer reviewed scientific publications will be made open access in accordance with the European Commission Open Access policy for publications: they will be self-archived upon publication in Zenodo[5] and, if possible, published in an appropriate open access journal or platform.
In addition, the Digital Twin Engine (DTE) will be openly accessible through the EOSC Portal[6] for testing purposes.

Finally, The DTE is designed to support open science with specific measures such as the capability of reusing and sharing a portfolio of base and trained models, algorithms and reference data that can be further enriched by DTE users, and the capability of reproducing research outputs such as modelling and simulation workflows thanks to provenance information.

Although interTwin will not manage all data assets directly this document establishes a set of guidelines and best practices. All parties involved, are expected to comply in their data management activities.

[3] https://stories.ecmwf.int/destination-earth/index.html
[4] https://creativecommons.org/licenses/
[5] https://zenodo.org/communities/intertwin/?page=1&size=20
[6] https://eosc-portal.eu/

1.2 Structure of the document

This document is comprised of the following chapters:

- **Section 1** presents an introduction to the project and the document.
- **Section 2** presents the purpose of data collection, type and format, origin of the data and its expected size.
- **Section 3** outlines interTwin FAIR data strategies.
- **Section 4** briefly describes the allocation of resources.
- **Sections 5**, **6** and **7** outline data security, ethical and other issues.
- **Section 8** shows the datasets per Work Package (WP).
- **Section 9** shows the datasets Use Cases.
- **Section 10** presents the next steps.
- **Section 11** concludes this deliverable.

2 Data Summary

In this chapter we describe the various types of data the interTwin consortium will handle during the project's lifetime. Some assets will be managed directly by the WPs, while others will be managed by scientific communities independently.

2.1 Origin of the project outputs reused or newly generated

The Digital Twin Engine will use a **co-design approach** taking into consideration the requirements, testing and exploitation conducted by multiple scientific communities from different domains to meet the transversal interdisciplinary DT requirements. Research communities will participate in the definition and evolution of the DTE architecture, its interoperability framework, its validation, and integration to ensure the concept and methodology address their needs. They will also contribute to the development of thematic components specialized to adjacent downstream sectors.

The role of these communities is to define an initial DTE blueprint architecture that will drive the developments of the projects and will be evolved and aligned in collaboration with external user communities and initiatives of pan-European relevance in research and industry that are contributing to the standardization of the technical aspects of DTs. External stakeholders will be engaged as follows:

- **Research communities from the long tail of science** > ***How***: consulting EOSC task forces and EOSC Portal.
- **Pan-European research infrastructures** from the ESFRI scientific domains (Social Sciences and Humanities - SSH, Environment, Life sciences, Photon and Neutron Science, Astronomy Astroparticle and Particle Physics) > ***How:*** EOSC Future engagement programme and follow-on support projects.
- **SMEs and industry** > ***How:*** EOSC Digital Innovation Hub and Platform Industrie 4.0 (manufacturing)
- **Evidence-based policy makers** > ***How:*** EuroGEO - the European regional initiative of the Group on Earth Observations. The aim of the collaboration with EuroGEO is to align the interTwin blueprint architecture with the GEO systems of systems approach and the Digital Ecosystems for developing DTs of the Earth proposed by JRC2. Additional feedback will be provided by an External Advisory Board.
- **Data space providers** > ***How:*** the Data Space Support Centre and the Alliance for Industrial Data, Edge and Cloud to ensure a common reference framework for distributed data access by DTs in a European sovereign cloud.

2.2 Existing data/software and newly generated project outputs

Software: Particular attention must be paid for integration of services into external marketplaces such as the EOSC and AI4EU as it requires seamless integration of IP from numerous sources (including, back- side- and foreground from the project beneficiaries, and third-party IP).

The role of the project is to provide support and specifications for turning this amalgamated, multi-party IP into efficient, well-managed software and services. For the reasons outlined above, software code developed by the project will be licensed under a permissive open-source license.

For outputs that are improvements to existing software, the improvement will be freely assigned to the owners of the background IP for incorporation therein and will have the permissive same open license as the software itself. Although copyright will exist in the source code generated during the project, this will not be asserted for research or future commercial use. Users are free to develop commercial applications, the flexibility for which is provided through the choice of a permissive license.

All users will be provided access to the Joinup Licensing Assistant (JLA) and the JLA compatibility checker to check inbound and outbound licensing terms in cases where applications including OS components are from different sources

2.3 Expected size of the data

The expected size of each dataset can be found in the section **8 Datasets per WP** and section **9 Datasets per Use Case**.

2.4 Formats of the project outputs

Section 8 and **9** shows the estimated datasets that interTwin is expected to produce/collect. This list is subjected to modifications, including addition or removal of datasets and their content in the later versions of DMP depending on the project developments.

2.5 Data Utility

The output of this project is, that the ***interdisciplinary Digital Twin Engine***, can be used by any scientific community, business, and civil society with the need for an open source platform based on open standards that offers the capability to integrate with application-specific Digital Twins.

3 FAIR Data

In this section we outline the interTwin policies and best practices concerning FAIR publication of research data assets. interTwin is committed to the publication of research data according to the FAIR principles.

The applicable FAIR principles are described in what follows. (Communities: High Energy Physics: [**HEP**], Radio astronomy [**RA**] Gravitational-wave Astrophysics [**GW**], Climate research and environmental monitoring [**CREM**])

3.1 Making data findable, including provisions for metadata

3.1.1 Findability of data/research outputs:

HEP: Fast simulated data and software will be published on Zenodo. Software in development will be accessible on Github[7].
RA: MeerKAT data (including metadata) are stored in the database MeasurementSet.
GW and **CREM**: Data and research outputs will be published on Zenodo.

3.1.2 Metadata

The metadata required by data repositories will be used, as outlined in table 1.
For documents, interTwin has defined a standard set of metadata that should be used, as shown in table 2.

Table 1 - Repository metadata

Element	Definition
Title	A name given to the source.
Upload type	e.g., dataset, workflow
Abstract	Describing the document contents and main conclusions.
Submitter	The person submitting the document to the repository
Authors	The people involved in writing significant portions of the document.
DOI	Provided by the resource
Publication date	The date of first publication.
Version	The version number generated by the document repository for the repository identifier. Versioning rule: • +0.1 – new version of draft • +1.0 – new version of approved document
Language	A language of the intellectual content of the resource.

[7] https://github.com/interTwin-eu

Keywords	A list of words that will support the search within the repository service
Communities	A specific community in which the upload will appear in.
License	Specifies the copyright status under which the upload will be licensed under.
Modify	The groups able to modify the document. The 'office' SSO group must be always marked.

Table 2 - Document metadata

Element	Definition
Title	A name given to the source. For milestones and deliverables as described in the Description of Work.
Lead Partner	The recognised short name of the lead partner within the interTwin project
Authors	The people involved in writing significant portions of the document.
Reviewers	The people involved in reviewing the document.
Copyright status	The material is licensed under a **Creative Commons Attribution 4.0 International License**
Document type	e.g., deliverable, report, white paper
Status	• **Draft** - the document is being prepared • **Under EC review** - the document is submitted to the EC portal and not approved yet by European Commission • **Approved by EC** - the document is approved by European Commission • **Final** - status of the document
Dissemination Level	• **Public** - can be shared without restrictions • **Confidential** - can be shared only with European Commission and project partners
Document link	The URL in the document repository that provides access to the document on **DocDB**.
Digital Object Identifier	An identification number assigned though a repository service.
Keywords	A list of words that will support the search within the Zenodo service
Abstract	Describing the document contents and main conclusions.

Metadata that will be created by the communities are currently collected and will be provided in milestone M1.3 on 31st May 2024.

3.1.3 Persistent identifiers

All research data assets produced within the project must be associated upon publication with a persistent and dereferenceable identifier. For public repositories we will adopt the identifiers provided by the resource. For workflows a DOI will be minted through Github. Outputs submitted to Zenodo, will be assigned a DOI through this service. For code we will adopt the practices of the developers community of the software we are building on.

3.1.3.1 Search keywords for discovery

Keywords will be created and then used to tag research output in Zenodo and in other registries or repositories (OpenAIRE[8] and EOSC catalogue).

3.2 Making data accessible

3.2.1 Accessibility of data/research outputs:

HEP: Fast simulated data and software will be open from the outset.
RA: Within this project, freely available radio astronomical databases are used.
GW: Curated observational gravitational-wave data are private to the collaborations, due to contractual obligations, for 18 months after the end of runs, and then released as Open Data through the Gravitational-Wave Open Science Centre portal. Raw data are published occasionally. Simulation results and trained models from this project will be released under CC BY 4.0.
CREM: Research outputs openly available and licensed under CC BY 4.0.

All research data assets produced within the interTwin project must be published in such a way that they are accessible by others. As a general rule, interTwin will consider as “accessible data” all research data assets exposed through one or more of the suitable, public community repositories.

3.2.2 Repositories

All documents, presentations and other materials that form an official output of the project (not just milestones and deliverables) are placed in the document repository[9] to provide a managed central location for all materials.

In addition, public deliverables, and publications, will be shared publicly via **Zenodo platform** to increase discoverability of the project outputs.

All profiles, specifications, configuration files, software, workflows, and code will be deposited in Zenodo and GitHub.

Therefore, the interTwin will use DocDB, Zenodo, and GitHub as their standard and main repositories.

[8] https://www.openaire.eu/
[9] https://documents.egi.eu/

3.2.3 Availability of the project outputs

As all deliverables, including documentation and guidelines necessary for (new) users after the project end, and because the bulk of the software created will be open source, access to project outputs will be ensured well beyond the project lifespan. This will ensure continuous uptake, and the possibility of creating new modifications and add-ons will be available for new and existing users and contributors even after the project ends.

As for the services (such as the DTE modules) will remain available on EOSC Marketplace. Updates and maintenance will be carried out by the interTwin Open Source Community the project intends to set up and promote. In addition, the project will leverage the Horizon Results Platform to increase visibility and potential further exploitation of results visible.

3.2.4 Standardised access protocol

All data will be accessible via URL or DOI. There will be no restrictions on the use of the research outputs, both during and after the end of this project. People accessing the data will not need to be identified and there is no need for a data access committee.

3.2.5 Metadata availability

Metadata containing information to enable users to access the data will be openly available and published together with the data, same repositories as listed under 3.2.2 Repositories. There is no time limit on metadata and data availability.

The interTwin project acknowledges the value of documentation for interoperability purposes, increased uptake by different communities, and will encourage data owners to document their research data assets. interTwin does not enforce specific provisions on documentation as long as the data asset is hosted on one of the mentioned repositories and properly curated according to the repository's best practices.

Research data itself should not be considered self-documenting and each published asset must be associated with sufficient documentation resources accessible through a public URL. Documentation must be browsable and include hypertext references to facilitate its fruition. Recommended documentation formats include markdown, HTML, and other markup languages. The inclusion of machine-readable documentation such as OpenAPI where applicable is thoroughly encouraged. If a scientific publication is tied to the research data asset, the publication itself should be referenced and/or made available as part of the documentation.

3.3 Making data interoperable

3.3.1 Interoperability of data/research outputs:

HEP: data will be released in HDF5 format.

Radio astronomy: well-established and well-documented data formats (e.g., Flexible Image Transport System, FITS; PSRCHIVE; European Pulsar Network, EPN). Other formats (e.g., HDF5, XML) are adopted where needed.
GW: FrameFile and the HDF5 formats. Interaction will be sought with the International Virtual Observatory Alliance (IVOA) for the evolution of standards.
CREM: Input and output data will mostly follow the conventions for CF (Climate and Forecast) metadata. Output data will be made available in standard formats including for example CSV, GRIB, HDF and NetCDF.

3.4 Increase data re-use

3.4.1 Reusability of data/research outputs:

HEP: Data and research output (pretrained model) will be licensed under CC BY 4.0

RA: Excellent track record in the reusability of data. Archives are routinely re-analysed, leading to the new discoveries and research fields (e.g., Fast Radio Bursts). Public data is free to use.

GW: Data and Research outputs will be openly available and will be licensed under CC BY 4.0. Some software products may be licensed under a different, Open Source Initiative (OSI)-approved licence.

CREM: Data will be licensed under CC BY 4.0, and the simulation software used in this project is open source and aligned with the open source strategy of the European Commission and the recommendation of the European Interoperability Framework.
Curation and storage/preservation costs of research outputs are activities out of the scope as the interTwin research outputs are used for validation and piloting.

3.4.2 Examples of Research Outputs:

Software [S], Data [D], Workflows [W]

HEP: [D] Lattice QCD simulated data from TB to PB scale. Datasets for the fast simulation of different HEP experiment settings. [S] open source software [D][S] Trained Deep Learning Models able to simulate different HEP experiment settings.

RA: [D] Generation of large volumes of digital-twin time series datasets with defined noise signals that can be used for training ML algorithms. [S] Existing ML libraries / tools / methods are explored in terms of their suitability for simulating noise signals. Development of scripts for integrating them into the pipelines of the TRAPUM project. [W] The workflow of the TRAPUM project is interfaced to DTE core capabilities.

GW: [D] Trained models for noise simulation and, possibly, time series of simulated noise for further development of de-noising strategies. [S]: Open-source software modules.

CREM [D]: Typical application output data and model input data are of the order of GB to TB. [S]: SFINCS (Super-Fast INundation of CoastS), Delft3D Flexible Mesh Suite and FIAT (Flood Impact Assessment Tool). Additionally, ML libraries / tools / methods will be explored. Data science Python libraries. ML models will be developed according to the state-of-the-art ML frameworks. [W]: Climate modules related to extreme storms detection and fires risk maps will be integrated with DTE core capabilities and workflows.

4 Allocation of Resources

Any expenses associated with the collection/production of FAIR data during the interTwin activities, are included in the project budget. These expenditures will be required to cover a variety of particular data processing and data management operations, ranging from data collection and documentation to storage and preservation to distribution and re-utilization.

These operations are a component of the WP that processes the relevant data, hence the needed effort will be part of the relevant WP.

The expenses of long-term data preservation are minimal, by using the EGI Online Storage and Google Drive platforms. Using Zenodo and GitHub (both free of charge) ensures that costs for long-term preservation of the data are manageable. When, applicable, a more accurate cost estimate will be provided at a later stage of the project.

4.1 Data Management responsibilities

Within the interTwin project the following roles and responsibilities are associated with Data Management, which are defined as follow:

WP leaders are in charge of organizing the data processing and quality assurance that take place inside the Work Package they are leading.
Task Leaders/Use Case leaders are in charge of the data compiled/produced throughout the operation of the task that they are responsible for. In addition to that, they also make sure that the data are properly prepared to be shared among the partners, and made publicly available, when applicable.
Data Processors are consortium partners who execute processing operations on the compiled/produced data.
Quality and Risk Manager monitors, and supports the WP leaders, and Task Leaders/Use Case leaders in keeping the DMP confluence pages up to date. In addition, reports the changes and processes via milestones and deliverables as specified in the Grand Agreement.

5 Data Security

Any gathered data will be securely handled throughout the entire duration of the interTwin project, to protect it from loss and unauthorized access. Personal data is only accessible to those who are authorized to access it.

All partners/beneficiaries responsible for processing personal data have the responsibility, to ensure that the data remains protected under all necessary security controls (including backup policies and integrity checks) and access controls (identification, authentication, authorization) within their infrastructure. In the unfortunate event of a personal data breach, the project partners will notify without delay their competent national supervisory authorities as well as the data subject(s) that may be affected by the breach. At the same time, they will document any personal data breaches and all related information.

Regarding open data, for security and for long-term preservation interTwin relies on EGI Document Repository, Zenodo, and GitHub.

As of this writing, interTwin is finalising the Data Protection Management System (DPMS[10]), further information will be provided in the Data Management Plan update.

[10] https://confluence.egi.eu/display/interTwin/Data+Protection+Management+System+DPMS – restricted to interTwin consortium members.

6 Ethical Aspects

This project includes multi- and interdisciplinary collaboration to support software development and use across several sciences. As well, "it is possible that the confidentiality and protection of personal data will necessitate specific arrangements for curation of data and handover to a follow-up activity."

The use of AI modelling and analytics could raise interesting issues and challenges. For these reasons, an ethics advisor will be useful to the project.

Moreover, based on the available documentation, interTwin has the potential for unforeseeable ethics issues related to personal data and privacy and to Artificial Intelligence, as the work plan intends to bring in unspecified use cases from SSH and Life Sciences from year 2.

Meanwhile the lack of adequate ethics management provisions is concerning, as the (unspecified) ethical management of the project rests with the administration and finance management task (T1. 2), and there is no hint that someone with expertise in ethical issues associated with such emerging technological systems will be involved.

6.1 Ethical evaluation response

To address findings from the Ethical evaluation the project:

- Set-up the project Ethics Board to ensure a proper monitoring of the ethics and data protection issues raised till the project ends.
- Development of the ethics framework due at Month 12.
    - the ethics governance processes about the use of AI modeling and analytics.
    - procedures governing the Co-design studies (WP4)
- Ensure project's compliance to the General Data Protection Regulation EU 2016/679.
    - Develop or review of the data protection procedures.
- Review the plan for use cases from an ethical perspective.
    - Develop ethics procedures for relevant use cases.
- Produce reports on the ethics issues monitoring and Horizon Europe compliance, as expected by the European Commission.
- The project identified external Ethics advisor to support the work.

A more detailed report on this approach will be provided in the upcoming deliverable D8.1 OEI – Requirements No.1, planned for submission in M12 of the project.

7 Other Issues

Within the interTwin project the following 2 Use Cases will make use of other national/funder/sectoral/departmental procedures for Data Management.

1. VIRGO Noise detector - Astrophysics DT use case (T4.4) and related thematic module (T7.3), will comply with the procedures prescribed by the Virgo collaboration.
2. Noise simulation for radio astronomy DT use case (T4.3) and related thematic module (T7.2), MPIfR and MPG, SARAO data management policies.

8 Data sets per WP

8.1 WP1 Project Coordination and Management

WP/Task	WP1
Contact	Malgorzata Krakowian (EGI.eu)
Data Summary	
Data description: Types of data	1. Project Documentation • Metrics • Risks • Procedures • Plans • Meetings agenda • Meetings participation list • Presentations • Deliverables • Mailing list archive 2. Effort and financial data
Data description: Origin of data	All the data was produced and provided by project members.
Data description: Scale of data	<1GB
Standards and metadata	plain text, .pdf, .docx, .pptx
Data sharing: Target groups	The target group is all project members and the EC Project office.
Data sharing: Scientific Impact	Not applicable
Data sharing: Approach to sharing	1. Shared within the consortium and European Commission • **Presentations**: Public presentations are made public via indico portal or external conference pages • **Deliverables**: All deliverables are shared within the consortium and also with European Commission. Public deliverables are accessible to

WP/Task	WP1
	everyone via the project website and Zenodo portal. • **Mailing list archive**: only accessible by the mailing list members. 2. Shared with the Project office and management boards to support work, as well as with the European Commission.
Archiving and preservation	Once the project is finished, all the WP1 information will be preserved by EGI Foundation for at least 5 years as well on EC funding portal.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Malgorzata Krakowian
How will long-term preservation be ensured?	Long-term preservation is not needed, except from the contractual 5 years after the project. A copy of all the documentation of the project is kept by the European Commission in the funding portal.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	To access the data shared only within the consortium, an EGI SSO account is required. Accounts and access management is the responsibility of the coordinator.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	For security and long-term preservation, interTwin relies on EGI Document Repository, Zenodo and Google Drive platforms
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	EGI Foundation will take care of the data according to the ISO 27000 standard for Information security management and GDPR.

8.2 WP2 Innovation Management and Communications

WP/Task	WP2
Contact	Xavier Salazar (EGI.eu)
Data Summary	
Data description: Types of data	Documents (e.g. meeting minutes, deliverables, publications, mailing list archive), Slides (e.g. project presentations, training material), Promotional material (e.g. printed such as flyers, posters, branding materials, etc, online - interTwin website, social media content, github, etc), audio-visual material (e.g. videos), Database (Stakeholder - including when necessary names & contact data), Feedback surveys
Data description: Origin of data	primary sources (project members) & secondary sources (external websites, documents, expert feedback, surveys etc)
Data description: Scale of data	< 1GB
Standards and metadata	plain text such as .docx, .txt, .rtf, .pdf, .pptx, xml, .xls, .html . Multimedia such as jpg/jpeg, gif, tiff, png
Data sharing: Target groups	T2.1: all project members and the EC Project office. T2.2: publicly available focusing on the target audiences of the project: including users, technology providers and infrastructure providers
Data sharing: Scientific Impact	Scientific Publications on peer reviewed journals, conferences
Data sharing: Approach to sharing	Shared within the consortium and European Commission via • **Presentations**: Public presentations are made public via

	Indico or external conference pages • **Deliverables**: All deliverables are shared within the consortium and also with European Commission. Public deliverables are accessible to everyone via the project website and Zenodo community • **Mailing list archive**: Only accessible to the mailing list members. • **Publications** will be available via the project website and interTwin community on Zenodo Repository • **Promotional and other audio-visual material** will be available via the project website Unless otherwise stated all content will be available under CC BY 4.0 license and metadata under CC0 license. Any consortium-restricted content is shared via access-protected confluence space
Archiving and preservation	Once the project is finished, all the WP2 information will be preserved by EGI Foundation for at least 5 years as well on EC funding portal. Publications will be also kept on Zenodo Community
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Xavier Salazar
How will long-term preservation be ensured?	Long-term preservation is not needed, except from the contractual 5 years after the project. A copy of all the documentation of the project is kept by the European Commission in the funding portal
Data Security	

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	To access the data shared only within the consortium, an EGI SSO account is required. Accounts and access management is the responsibility of the coordinator.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	For security and long-term preservation, interTwin relies on EGI Document Repository, Zenodo and Google Drive platforms
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	EGI Foundation will take care of the data according to the ISO 27000 standard for Information security management and GDPR.

8.3 WP3 Technical Coordination and Interoperability

WP/Task	WP3
Contact	Andrea Manzi (EGI.eu)
Data Summary	
Data description: Types of data	1. Project Documentation • Meetings agenda • Meetings participation list • Presentations • Deliverables • Mailing list archive
Data description: Origin of data	All the data was produced and provided by project members.
Data description: Scale of data	<10GB
Standards and metadata	plain text, .pdf, .docx, .pptx,
Data sharing: Target groups	The target group is all project members and the EC Project office.
Data sharing: Scientific Impact	not applicable

Data sharing: Approach to sharing	1. Shared within the consortium and European Commission - **Presentations**: Public presentations are made public via indico portal or external conference pages - **Deliverables**: All deliverables are shared within the consortium and also with European Commission. Public deliverables are accessible to everyone via the project website and Zenodo portal. - **Mailing list archive**: only accessible by the mailing list members.
Archiving and preservation	Once the project is finished, all the WP3 information will be preserved by EGI Foundation for at least 5 years as well on EC funding portal.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or re-used throughout the projects?	Software Releases artefacts, software source code and documentation will be shared via interTwin software repository and Github
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Andrea Manzi
How will long-term preservation be ensured?	Long term preservation is not needed, except from the contractual 5 years after the project. The copy of all the documentation of the project is kept by European commission in the funding portal.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure	To access the data shared only within the consortium, an EGI SSO account is required.

storage/archiving and transfer of sensitive data)?	Accounts and access management is responsibility of the coordinator.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	The data will be stored at the repositories hosted and managed by the EGI Foundation.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	EGI Foundation will take care of the data according to the ISO 27000 standard for Information security management and GDPR.

8.4 WP4 Technical co-design and validation with research communities

WP/Task	WP4
Contact	Levente Farkas (EGI.eu)
Data Summary	
Data description: Types of data	Project Documentation • Meetings agenda and minutes • Meetings participation list • Presentations • Deliverables • Mailing list archive
Data description: Origin of data	All the data was produced and provided by project members.
Data description: Scale of data	< 10GB
Standards and metadata	text, pdf, docx, xlsx, pptx
Data sharing: Target groups	Project members and the EC Project office
Data sharing: Scientific Impact	N/A
Data sharing: Approach to sharing	Shared within the consortium and European Commission

WP/Task	WP4
	• **Presentations**: Public presentations are made public via Indico or external conference pages • **Deliverables**: All deliverables are shared within the consortium and also with European Commission. Public deliverables are accessible to everyone via the project website. • **Mailing list archive**: Only accessible to the mailing list members.
Archiving and preservation	After the project's end all the WP4 information will be preserved by EGI Foundation for at least 5 years as well on EC funding portal.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Levente Farkas
How will long-term preservation be ensured?	Long term preservation (beyond the contractual 5 years of the project) is not needed. The copy of all project documentation is kept by European Commission in the funding portal.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	To access the data shared only within the consortium, an EGI SSO account is required. Accounts and access management is responsibility of the coordinator.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	The data will be stored at the repositories hosted and managed by the EGI Foundation.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	EGI Foundation handles data according to the ISO 27000 standard for Information security management and GDPR.

8.5 WP5 Digital Twin Engine Infrastructure

Information will be provided in Milestone M1.3, planned for May 2024.

8.6 WP6 Digital Twin Engine Core Modules

Information will be provided in Milestone M1.3, planned for May 2024.

8.7 WP7 Digital Twin Engine Thematic Modules

WP/Task	WP7
Contact	Charis Chatzikyriakou
Data Summary	
Data description: Types of data	• Meeting information, i.e., meeting agendas, attendees, minutes of meeting • Meeting material, i.e., presentations • Documents, i.e., deliverables • Personal data for communication purposes, i.e., WP/Task participants' lists of names and e-mail addresses
Data description: Origin of data	All the data was produced and provided by project members.
Data description: Scale of data	< 1GB
Standards and metadata	Plain text files such as .docx, .txt, .rtf, .pdf, .pptx, .xml, .xls, .html.
Data sharing: Target groups	The target group is all project members and the EC Project office.
Data sharing: Scientific Impact	Not available.
Data sharing: Approach to sharing	Shared within the consortium, the European Commission, and the public: • **Meeting information**: The meeting agendas, attendees and minutes of meeting are accessible in the project's collaboration tool (Confluence, Google Drive).

	• **Meeting material**: Public presentations are made public via Indico or external conference pages. • **Documents**: All deliverables are shared within the consortium and also with European Commission. Public deliverables are accessible to everyone via the project website and Zenodo. • **Personal data for communication purposes**: Only accessible to the mailing list administrators and members.
Archiving and preservation	Once the project is finished, all the WP7 information will be preserved by EGI Foundation for at least 5 years as well on EC funding portal.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Charis Chatzikyriakou
How will long-term preservation be ensured?	Long-term preservation is not needed, except from the contractual 5 years after the project. A copy of all the documentation of the project is kept by the European Commission in the funding portal.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	To access the data shared only within the consortium, an EGI SSO account or explicit access to documents are required. Accounts and access management is the responsibility of the coordinator.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	The data will be stored at the repositories hosted and managed by the EGI Foundation.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	*EGI Foundation will take care of the data according to the ISO 27000 standard for Information security management and GDPR.*

9 Datasets per Use case

The following overview describes the Data Management Plan for the Use Cases data that will be generated within interTwin. For each dataset, it describes the type of data and its origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.

Beneficiaries must manage responsibly the digital research data generated in the action ('data') in line with **the FAIR principles**. They should also ensure open access to research data via a trusted repository under the principle 'as open as possible, as closed as necessary. The requirements for research data management apply only to data that are generated in the course of the action. Beneficiaries should also consider **re-used data** when developing their Data Management Plans (DMPs), if they form part of their research and to the extent possible.

- *Beneficiaries must establish a DMP, addressing important aspects of Research Data Management (RDM).*
    - Beneficiaries should maintain the DMP as a living document and update it over the course of the project whenever significant changes arise. This includes, but is not limited to: the generation of new data, changes in data access provisions or curation policies, attainment of tasks (e.g. datasets deposited in a repository, etc.), changes in relevant practices (e.g. new innovation potential, the decision to file for a patent), changes in consortium composition. Beneficiaries are encouraged to encode their DMP deliverables as non-restricted, public deliverables, unless there are reasons (legitimate interests or other constraints) not to do so. In the case they are made public, it is also recommended that open access is provided under a CC BY licence to allow a broad re-use.
- *Beneficiaries must deposit the data in a trusted repository (see explanation above) and ensure open access through the repository, as soon as possible and within the deadlines set out in the DMP.*
    - Deposition of data must take place as soon as possible after data production/generation or after adequate processing and quality control have taken place, providing value and context to the data and at the latest by the end of the project. This does not entail that data must be made open, but rather that it is deposited so that metadata information is available and hence information about the data is findable. In exceptional cases in which specific constraints apply (e.g. security rules), deposition can be delayed beyond the end of the project. Data includes raw data, to the extent technically feasible, but especially if it is crucial to enable reanalysis, reproducibility and/or data reuse.

9.1 Lattice QCD Simulations - High Energy Physics use case (T4.1) and related thematic module (T7.1)

Information will be provided in Milestone M1.3, planned for May 2024.

9.2 Detector simulation - High Energy Physics use case (T4.2) and related thematic module (T7.7)

WP/Task	4/4.2, 7/7.7
Contact	Sofia Vallecorsa (CERN), Kalliopi Tsolaki (CERN), David Rousseau, Benoit Blossier (CNRS)
Established a DMP, addressing important aspects of RDM.	√ In place ○ In progress ○ Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	*Yes*
What types and formats of data will the project generate or re-use?	hdf5, ONNX, root
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	During the R&D phase, this is a representative data set for typical future applications. Later on the data sets can be updated and used for optimisation and maintenance of the DT
What is the expected size of the data that you intend to generate or re-use?	100 GB
What is the origin/provenance of the data either generated or re-used?	Monte Carlo simulation
To whom might your data be useful ('data utility') outside your project?	High Energy Physics detector design community

FAIR Data	
1.) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Yes
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Yes
	Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? Don't know yet
	Will metadata be offered in such a way that it can be harvested and indexed? Yes, depends on how the procedures for harvesting and indexing are going to be set up
2.) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Zenodo
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? This is the default choice
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Yes
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly separating legal and contractual***

	reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement. Yes
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. It's unlikely that the data produced will be subject to an embargo, the details are still being defined.
	Will the data be accessible through a free and standardised access protocol? Whatever is provided by Zenodo
	If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?
	How will the identity of the person accessing the data be ascertained? Not implemented
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** *Yes*

	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? Zenodo policies
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Yes
3.) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** HEP community
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Not known yet.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? No
4.) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** Upload on the same Zenodo record

	Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license? Yes
	Will the data produced in the project be useable by third parties, in particular after the end of the project? No
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes, HEP standard
	Describe all relevant data quality assurance processes. In progress
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects. Not known yet.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or re-used throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Software
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Kalliopi Tsolaki
How will long-term preservation be ensured?	***(costs and potential value, who decides and how what data will be kept and for how long)***

	Still under study
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	No sensitive data
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Zenodo
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).*** No
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	No need
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** No

9.3 VIRGO Noise detector - Astrophysics DT use case (T4.4) and related thematic module (T7.3)

WP/Task	4.4
Contact	Sara Vallero (INFN)
Established a DMP, addressing important aspects of RDM.	• In place √ In progress • Non

Data Summary	
Will you re-use any existing data and what will you re-use it for?	***State the reasons if re-use of any existing data has been considered but discarded.*** We will re-use Virgo data of past observing runs to study the features of transient noise and to develop the GAN architecture to be used in the DT.
What types and formats of data will the project generate or re-use?	The project will use files in *hdf5* anf *gwf* (Gravitational Wave Frame) formats. The latter is a proprietary format.[11] The project will generate data containing the trained GAN models and the DT output information. The format of these data products has not been defined yet.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	Existing detector data will be used to characterise transient noise in different readout channels of the interferometer and as input to the GAN model both in the training and inference phases. The trained models will be used in the DT operations to infer the detector response in the *strain* readout channel (sensitive to the astrophysical signal) from the response in the *auxiliary* channels (containing data from detector sensors only). The DT output will contain information about the probability of input data to be originated from an astrophysical source or from transient noise. This information needs to be propagated to downstream search pipelines (not part of the DT) and act as possible veto for further processing in case the signal has been identified as transient noise by the DT.
What is the expected size of the data that you intend to generate or re-use?	We expect a size of few tens of TB of detector data to be made available for noise characterisation studies. These

[11] https://lappweb.in2p3.fr/virgo/FrameL/VIR-067A-08.pdf

	data will also be used to define the architecture of the DT by studying a variety of readout channels (in order to identify the ones most suited for the DT) and of transient noise topologies. For the DT operations we expect a few hundred GB of data to be made available on scratch storage for periodic retraining of the GAN model. At this stage it's not possible to foresee the size of the trained models and of the output data, but we expect the latter to be orders of magnitude smaller than those required for DT operations.
What is the origin/provenance of the data either generated or re-used?	Detector data come from the European computing facilities supporting the Virgo collaboration (mainly EGO).
To whom might your data be useful ('data utility') outside your project?	Data containing the trained GAN models and the DT output can be useful for the Virgo collaboration.
FAIR Data	
1.) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** The trained GAN models will most likely be identified by a persistent identifier. At this stage it's not clear if also the DT output will be.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Metadata will be provided to allow discovery. Most likely output metadata will reflect the metadata of input detector data, namely: GPS time, read-out channels and data-quality flags.

	Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? Yes.
	Will metadata be offered in such a way that it can be harvested and indexed? Not decided yet.
2.) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Most likely yes, but the details have not been defined yet.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? No.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Not known.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** Most likely trained data will be made openly available to the Virgo collaboration. It is not clear at this stage if

	they will also be shared outside the collaboration.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. It's unlikely that the data produced will be subject to an embargo.
	Will the data be accessible through a free and standardised access protocol? Yes.
	If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? No restrictions are foreseen.
	How will the identity of the person accessing the data be ascertained? Through Virgo's federated identity providers.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why.*** Yes. ***Will metadata contain information to enable the user to access the data?*** Yes.

	How long will the data remain available and findable? Not known. ***Will metadata be guaranteed to remain available after data is no longer available?*** No.
	Will documentation or reference about any software be needed to access or read or process the data be included? Yes. ***Will it be possible to include the relevant software (e.g. in open-source code)?*** Yes, it will be made available through an open repository (i.e. GitHub).
3.) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** We will follow the prescriptions form the Virgo collaboration for the sharing within the community. We do not expect data to be interoperable across disciplines.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Not applicable.
	Will your data include qualified references to other data (e.g. other data

	from your project, or datasets from previous research)? Yes. Trained models and output data will contain references to the input data. The details of how this will be achieved are not defined yet.
4.) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** We will provide detailed documentation accompanying the code in the GitHub repository.
	Will your data be made openly available in the public domain to permit the widest re-use possible? Most likely not, but the details have not been defined yet. ***Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement?*** Yes. ***Under which license?*** Not known yet.
	Will the data produced in the project be useable by third parties, in particular after the end of the project? Data could be re-used by the Virgo collaboration and possibly also by the Einstein Telescope collaboration.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes.

	Describe all relevant data quality assurance processes. For the DT output data, the data quality assurance processes is not defined yet.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or re-used throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Not known.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	The task leader (Sara Vallero).
How will long-term preservation be ensured?	***(costs and potential value, who decides and how what data will be kept and for how long)*** Not known.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Standard data security and recovery practices will be put in place, the details have not been defined yet. There will be no handling of sensitive data.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes, the details are not known.
Ethical Aspects	

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).*** No.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not applicable.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** We will comply with the procedures prescribed by the Virgo collaboration.

9.4 Noise simulation for radio astronomy DT use case (T4.3) and related thematic module (T7.2)

WP/Task	4/4.3, 7/7.2
Contact	Yurii Pidopryhora (MPG)
Established a DMP, addressing important aspects of RDM.	• In place √ In progress • Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	***State the reasons if re-use of any existing data has been considered but discarded.*** Yes. The current dataset we are working with is constantly reused both for trying different ML approaches for classifying it and for study of its properties in order to create a reliable simulation. Same is expected for future datasets to be reused in this project.

What types and formats of data will the project generate or re-use?	Distributed Acquisition and Data Analysis (DADA), filterbank object binary file (.fil), other common radio-astronomical data formats (like FITS or CASA MeasurementSet), ascii formats (comma-separated values (CSV) or similar)
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	It is or will be re-used to develop the digital twins of radio astronomical data flow, in particular training the ML models and analysing the data for identifying key characteristics of noise, RFI and other aspects.
What is the expected size of the data that you intend to generate or re-use?	Current dataset we are re-using is ~12 Tb, the future MeerKAT datasets we are going to use are probably going to be of order 100 Tb.
What is the origin/provenance of the data either generated or re-used?	At present we use a dataset specifically created for this project, by MPIfR-operated Effelsberg 100m radio telescope observing the Crab pulsar. The MeerKAT data that we are going to use in the nearest future will come from the MeerKAT radio astronomical array operated by the South African Radio Astronomy Observatory (SARAO) in the framework of a separate scientific project lead by MPIfR scientists, we are going to re-use their raw data for our purposes.
To whom might your data be useful ('data utility') outside your project?	Radio astronomers interested in the targets observed or dealing with similar data classification issues/noise and RFI studies.
FAIR Data	
1.) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Yes, the data sets are designated in accordance with the observatory/project standards.
	Will rich metadata be provided to allow discovery? What metadata will be

	created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. The data already have headers standard for radio astronomy, identifying source, project, epoch, various parameters etc.
	Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? When archived, the necessary keywords will be added, again in accordance with the radio astronomy standards.
	Will metadata be offered in such a way that it can be harvested and indexed? The final data product will be in a standard format (like FITS) with metadata header that can be easily read. Also, the archival systems themselves usually allow for access to metadata and keywords.
2.) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Yes, the standard archive for the given type of telescope/project.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? There is no need, the standard procedure for radio astronomical data will be followed.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

	Yes.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** All the data used in this project is related to scientific projects and, as it is common in radio astronomy, will be made openly available following the standard procedures (including embargo, as clarified in the next item).
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. The standard procedure is to embargo the data for about one year after its release to the scientific team responsible for the given observational project to allow for exclusive analysis and publication. However, parts of it, especially as related to technical aspects of the data that we are working with, can be distributed with the approval of the science team PI.
	Will the data be accessible through a free and standardised access protocol? It will be kept in an archive related to the telescope/organization that performed the observations.

	If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? If the grace period is still in force for a data set in question, access has to be approved by the PI of the scientific project this data relates to.
	How will the identity of the person accessing the data be ascertained? After the data is archived (i. e. openly released), one usually has to create an account giving some basic personal info to have access to it, but it typically has minimal security. Before the data is openly released, only people with computer accounts in the given institution (in our case, MPIfR) *and* whose access to the particular machine has been cleared can have direct access to it.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No, the teams are small enough to be able to solve all the access issues by personally contacting the responsible people.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Yes
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? In principle, indefinitely, or at least on a scale of decades.

	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Processing radio-astronomical data is a difficult task that cannot be covered by a help file or instruction manual, but all the basic means for a specialist to read the data and get it ready to be processed will be provided.
3.) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** We follow standard procedures common in the whole field of radio astronomy.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? In the parts where we would go beyond the standard radio-astronomical procedures, it is unclear at this point.
	Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)? Yes, where applicable.
4.) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files***

	with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)? Where applicable, help, readme files and other materials will accompany the data and the software. We also plan to publish all the relevant finding in professional journals, providing as much explanations and additional materials (like links to files or references to repositories) as possible.
	Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license? The data and everything related will be available on the standard academic basis: free distribution and use provided proper references are given.
	Will the data produced in the project be useable by third parties, in particular after the end of the project? Yes.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes.
	Describe all relevant data quality assurance processes. Not clear at this point.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

	Academic institutions and their employees involved in this task already follow high standards in this respect.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or re-used throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Yes. The software and other materials will be available on github and similar repositories. The findings will be published in professional journals with maximum additional materials.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	In the parts directly pertaining to the interTwin: Yurii Pidopryhora and all the partners involved in the tasks. The scientific data sets in general are managed by the science team whose project it is and, later, by the relevant archive and the institution that runs it.
How will long-term preservation be ensured?	***(costs and potential value, who decides and how what data will be kept and for how long)*** We rely on the industry standards in our field.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	For the duration of the project we will keep the data in a number of copies on trusted machines in our institutions, which themselves have storage redundancies. After the project is completed, the data will be kept in standard secure archives.

Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).*** Yes, with respect to the data that has scientific value. We should be careful not to disclose any key scientific information before the science team publishes their findings. Since we are dealing with the raw data and technical issues, this should not be a significant problem, but must be remembered.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	N/A
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** MPIfR and MPG, SARAO data management policies. Any data and results of its analysis obtained within a framework of a scientific project cannot be released to (or even discussed with) outside parties without express permission of the PI of the project for a certain "grace period" (typically one year after the date of the official data release to the scientific team). All related publications must clearly reference the scientific and observation/data reduction teams of the project and all the agencies involved.

9.5 Climate Change Future Projections of Extreme Events (storms & fire) use case (T4.5) and related thematic module (T7.4)

WP/Task	T4.5 - Climate Change Future Projections of Extreme Events (storms & fire)
Contact	Donatello Elia (CMCC)
Established a DMP, addressing important aspects of RDM.	○ In place √ In progress • Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	*State the reasons if re-use of any existing data has been considered but discarded.* Data from public repositories will be used as input for the DT on climate future projection of extreme weather events: • CMIP6 climate projection data • ERA5 reanalysis data • Fire Danger Indices data • International Best Track Archive for Climate Stewardship (IBTrACS) tropycal cyclones observation data
What types and formats of data will the project generate or re-use?	• NetCDF and CSV data formats • ML model (e.g., SavedModel/HDF5 format)
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	Data will be (re-)used to develop the DTs on extreme weather events, in particular for: • Training of ML model on past/present data

	- Inference through ML models on future climate projections
What is the expected size of the data that you intend to generate or re-use?	The expected overall size is of TB order, mostly depending upon the CMIP6 data considered in the DTs
What is the origin/provenance of the data either generated or re-used?	- CMIP6 data will be downloaded from the ESGF infrastructure - ERA5 reanalysis and Fire Danger indices data from Copernicus CDS - IBTrACS observation data from NOAA
To whom might your data be useful ('data utility') outside your project?	- Scientists interested on extreme events studies - Policy and decision makers interested in what-if scenarios related to extreme events
FAIR Data	
1.) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** - Scientific data output will be identified by a single PID for the whole dataset - Trained ML models will be identified by PIDs (e.g., git commits)
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

	• Scientific data output will be in NetCDF format • There are many formats to save a trained ML model, mostly depending upon the software used. We'll use existing and widely used formats such as among others, YAML, JSON, Pickle.
	Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? • A set of keywords will be defined at the dataset level to enable search & discovery • Yes, for the trained ML models
	Will metadata be offered in such a way that it can be harvested and indexed? • Metadata can be harvested from the NetCDF files headers • We'll use standard formats and rely upon existing libraries for metadata extraction
2.) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** • Yes, the output dataset will be hosted on a trusted repo (e.g., CMCC) • For trained ML models we'll used trusted repo too (e.g., Zenodo)

	Have you explored appropriate arrangements with the identified repository where your data will be deposited? • Not yet for output dataset, but we can rely on operation setting at CMCC where other CMIP* datasets are already accessible in the wider ESGF network • Yes, in the simplest scenario it will be Zenodo
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? • Not directly. For the scientific dataset, we could define the related metadata record in Zenodo, while keeping the repository at one of the use case premises • For the ML model, Zenodo will offer the solution
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** • Yes, open access will be granted to the scientific data

	- Yes, open access will be granted to the trained ML model
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. - Not applicable - Not applicable
	Will the data be accessible through a free and standardised access protocol? - *Yes, the data be accessible through a free and standardised access protocols (HTTP, OPeNDAP (DAP protocol))* - *Yes (via the interfaces offered by Zenodo)*
	If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? - Not applicable - Not applicable
	How will the identity of the person accessing the data be ascertained? - No AuthN/AuthZ will be required for the scientific dataset and trained ML models
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

	• No in both cases (both for the scientific dataset and trained ML models)
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** • All metadata for the the scientific dataset and trained ML models will be openly available
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? • Data from the the scientific dataset and trained ML models will be available and findable for at least 10 years. • Concerning metadata they will be managed via Zenodo, so they will remain available regardless of the data availability
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? • The file format used by the scientific output dataset (i.e. NetCDF) has a very large number of free tools and software to access it.

	- For the trained ML models we'll adopt widely used format, so also in this case no need for supplementary documentation
3.) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** - For the scientific dataset, will try to follow the CF-convention (to the maximum extent possible), which is what is used in the community. - For ML model we will try to use ONNX (interoperable) format, when possible.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? - Not totally clear at this stage, but in case project-specific output would be generated, they will be certainly made openly available within the community
	Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)?

	- Yes, we will include references to the original data (see Data Summary section).
4.) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** - Via Jupyter Notebook in both cases (scientific data and trained ML models)
	Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license? - Scientific data and trained ML models will be openly available in the public domain
	Will the data produced in the project be useable by third parties, in particular after the end of the project? - Yes, both during and after
	Will the provenance of the data be thoroughly documented using the appropriate standards? - We will only provide limited provenance information, as full

	provenance is not in force yet in the climate community.
	Describe all relevant data quality assurance processes. - The scientific output will follow scientific validation procedures usually adopted in the domain. For the trained ML model, validation will be performed following ML best practices based on well-known metrics (e.g., MAE, MSE, RMSE) and according to specific use case.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or re-used throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** In addition to scientific output datasets and trained ML models, we are also considering two more categories or research output like software and workflows. In that respect we'll get inspired by FAIR principles too according to best practices and guidelines available (e.g. FAIR4RS).
Allocation of resources	
Who will be responsible for data management in your WP/Task?	All the partners involved in the Task 4.5

How will long-term preservation be ensured?	***(costs and potential value, who decides and how what data will be kept and for how long)*** Not defined yet
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Not defined yet
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Not defined yet
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).*** No
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not applicable
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** No

9.6 Early Warning for Extreme Events (floods & droughts) DT use case (T4.6) and related thematic module (T7.6)

WP/Task	T4.6

Contact	Matthias Schramm (TU Wien)
Established a DMP, addressing important aspects of RDM.	○ In place √ In progress ○ Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	***State the reasons if re-use of any existing data has been considered but discarded.*** • Several static input data will be used for detecting flood and drought anomalies that have been processed independently of the project. As a stretch goal, several of this data shall be processed dynamically, thus better adapting to the workflow and possibly allowing a higher accuracy. ○ Harmonic models: Models of theoretical raster values, if there would be no flood at the monitored time. ○ Information on land cover (CORINE Land Cover) ○ Information on backscatter values of known water bodies
What types and formats of data will the project generate or re-use?	• Raster data: NetCDF and GeoTIFF format
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	• On-the-fly processing of flood and drought information • Creating archives of flood and drought information
What is the expected size of the data that you intend to generate or re-use?	

What is the origin/provenance of the data either generated or re-used?	• Sentinel-1 time-series: ○ Origin: EC's Copernicus programme. Freely available ○ Provided by project's service providers • Static input data for detecting flood / drought anomalies (Harmonic model, CORINE Land Cover) ○ Origin: EEA, TU Wien
To whom might your data be useful ('data utility') outside your project?	• ESA, EUMETSAT, ECMWF
FAIR Data	
1.) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Yes
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Not defined yet
	Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? Yes
	Will metadata be offered in such a way that it can be harvested and indexed? Not defined yet
2.) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Not defined yet
	Have you explored appropriate arrangements with the identified

	repository where your data will be deposited? Not defined yet
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Not defined yet
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** Not defined yet
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not known yet.
	Will the data be accessible through a free and standardised access protocol? Not defined yet
	If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? Not defined yet
	How will the identity of the person accessing the data be ascertained? Not defined yet

	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? Not defined yet
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Not defined yet
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? Not defined yet
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Not defined yet
3.) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** Not defined yet
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Not defined yet

	Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)? Not defined yet
4.) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** Not defined yet
	Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license? Not defined yet
	Will the data produced in the project be useable by third parties, in particular after the end of the project? Not defined yet
	Will the provenance of the data be thoroughly documented using the appropriate standards? Not defined yet
	Describe all relevant data quality assurance processes. Not defined yet
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects. Not know yet.
Other research outputs	
In addition to the management of data, are you also considering and planning	***Such outputs can be either digital (e.g. software, workflows, protocols, models,***

for the management of other research outputs that may be generated or re-used throughout the projects?	***etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Not defined yet
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Matthias Schramm (TU Wien)
How will long-term preservation be ensured?	***(costs and potential value, who decides and how what data will be kept and for how long)*** Not known yet.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Not known yet.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Not known yet.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).***
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not known yet.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** NA

9.7 Climate Change Impacts of Extreme Events (storms, fire, floods, drought) use case (T4.7) and the related thematic module (T7.5)

WP/Task	WP4/T4.7
Contact	Christian Pagé (CERFACS)
Established a DMP, addressing important aspects of RDM.	• In place √ In progress • Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	No.
What types and formats of data will the project generate or re-use?	1- NetCDF (could also be zarr)
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	1- Create a database of extreme climate indices to support the extreme events climate change impacts DT.
What is the expected size of the data that you intend to generate or re-use?	1- On the order of a few 100s of Gb for climate indices.
What is the origin/provenance of the data either generated or re-used?	1- CMIP6 climate simulations post-processed with the icclim tool to generate climate indices.
To whom might your data be useful ('data utility') outside your project?	1- Scientific Researchers using climate data and working on the impacts of extreme events.
FAIR Data	
1.) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** 1- The climate indices database will be identified by a PID but only for the whole dataset.
	Will rich metadata be provided to allow discovery? What metadata will be

	created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. 1- NetCDF (or zarr) files will conform to the CF-Conventions with rich metadata.
	Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use? 1- No keyword for the climate indices database.
	Will metadata be offered in such a way that it can be harvested and indexed? 1- Metadata can be harvested from the NetCDF files headers (or zarr).
2.) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** *1*- Yes, the climate indices database will be stored in a trusted repository.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? 1- In the EGI infrastructure in the context of EOSC
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? 1- It is not know for the climate indices database.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted***

	access conditions), explain why clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement. 1- All the climate indices dataset will be openly accessible.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not known yet.
	Will the data be accessible through a free and standardised access protocol? 1- Climate indices database will be accessible through a free and standardised access protocol
	If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? Not defined yet
	How will the identity of the person accessing the data be ascertained? 1- Accessing the climate indices database will required login from a trustworthy AAI.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? 1- Not for the climate indices database.

c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** 1- All metadata for the climate indices database will be openly available.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? 1- Data from the climate indices database will remain available for at least 10 years.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? *1- The file format used by the climate indices database, NetCDF (or zarr), has a very large number of free tools and software to access it.*
3.) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** 1- For the climate indices database, it will only follow the CF-convention. This is what is used in the community.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you

	provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? 1- For the climate indices database, it is not known if there exist mappings or not for the CF-Convention.
	Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)? 1- The climate indices database will include references to the original CMIP data.
4.) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** 1- For the climate indices database there already exists some freely available notebooks.
	Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license? 1- The climate indices database will be in the public domain.
	Will the data produced in the project be useable by third parties, in particular after the end of the project? 1- The climate indices database will be usable by third parties after the end of the project, and even during the project.

	Will the provenance of the data be thoroughly documented using the appropriate standards? 1- The climate indices database will only provide limited provenance information, as full provenance is not in force yet in the climate community.
	Describe all relevant data quality assurance processes. 1- The climate indices database will be validated with relevant statistics.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects. Not defined yet
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or re-used throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** 1- For the DT that will be using the climate indices database to characterize the extreme events in the future climate, a semi-generic AI-based software workflow will be created and will also follow FAIR4RS principles.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Every partner producing data in T4.7.
How will long-term preservation be ensured?	***(costs and potential value, who decides and how what data will be kept and for how long)***

	1- For the climate indices database, preservation will be ensured by the ENES CDI.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	1- For the climate indices database it will be determined when the final database location will be confirmed.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	1- For the climate indices database it will be determined when the final database location will be confirmed.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).*** 1- Not for the climate indices database
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	NA
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** 1- Not for the climate indices database.

10 Next steps

10.1 DMP Change Management

This DMP is considered a living document, and it will be updated during the project to reflect the most recent developments and conclusions. In actuality, the early version of the DMP will be expanded and enhanced at least once more on M36 of the project.

Ad hoc improvements may also be deployed if deemed necessary. In general, changes need to be fully compliant with EU laws and best practices in research data management.

Updates will be entered in the changelog table that is shown on the confluence page of the concerned DMP. Analogously, the distribution of notifications on updates will be realised via the regular meetings (WPs, WP leaders, PMO, AMB).

11 Conclusion

interTwin’s DMP is a complete data management approach that complies with Horizon Europe recommendations and that aims to make data as findable, accessible, interoperable, and re-usable (FAIR) as feasible.

The DMP rely on technological solutions and standards like OpenAIRE initiative, GitHub, EGI Document Repository, Zenodo, and Google Drive for the execution of these processes. Additionally, this will ensure that the data created or compiled throughout the interTwin project, including open data and public publications, will be kept and continue to be usable once the project is completed.

The DMP is intended to safeguard the analysation of compiled/created data based on the level of their privacy and to use an alternate sharing methodology relying upon this level. Confidential information or information that raises ethical problems will not be released.

Finally, the DMP is built on guaranteeing appropriately informed consent, and protecting each participant's zone of privacy, while adhering to GDPR guidelines