



interTwin

D1.4 Final Data Management Plan

Status: Under EC Review
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, metadata, re-usable data

This final version of the Data Management Plan (DMP) outlines the approach adopted for the collection, processing, monitoring, and cataloguing of research data, in accordance with the FAIR principles. Building upon the first version (D1.2), this updated deliverable reflects the evolution of data practices throughout the project lifecycle and consolidates the final strategies and tools implemented for data management.



Document Description			
D1.4 Final Data Management Plan			
Work Package 1			
Document type	Deliverable		
Document status	Under EC Review	Version	1
Dissemination Level	Public		
Copyright Status	 <p>This material by Parties of the interTwin Consortium is licensed under a Creative Commons Attribution 4.0 International License.</p>		
Lead Partner	EGI		
Document link	https://documents.egi.eu/document/3920		
DOI	https://zenodo.org/records/17099293		
Author(s)	<ul style="list-style-type: none"> Andrea Anzanello (EGI) 		
Reviewers	<ul style="list-style-type: none"> Patricia Ruiz (EGI) Matteo Agati (EGI) 		
Moderated by:	<ul style="list-style-type: none"> Andrea Anzanello (EGI) 		
Approved by	Malgorzata Krakowian (EGI) on behalf of AMB		

Revision History			
Version	Date	Description	Contributors
V0.1	30/07/2025	Review and update of the previous version	Andrea Anzanello (EGI)
V0.2	21/08/2025	Structure review and corrections	Patricia Ruiz (EGI), Matteo Agati (EGI)
V0.3	05/09/2025	AMB review	Malgorzata Krakowian (EGI)
V0.4	10/09/2025	Final revision	Andrea Anzanello (EGI)
V1.0		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

1	Introduction	8
1.1	Purpose and scope of the document.....	8
1.2	Structure of the document.....	9
2	Data Summary	10
2.1	Origin of the project outputs reused or newly generated	10
2.2	Existing data/software and newly generated project outputs	11
2.3	Expected size of the data	11
2.4	Formats of the project outputs.....	11
2.5	Data Utility	11
3	FAIR Data	12
3.1	Making data findable, including provisions for metadata.....	12
3.1.1	Findability of data/research outputs:	12
3.1.2	Metadata	12
3.1.3	Persistent identifiers.....	13
3.2	Making data accessible.....	14
3.2.1	Accessibility of data/research outputs:	14
3.2.2	Repositories	14
3.2.3	Availability of the project outputs.....	15
3.2.4	Standardised access protocol	15
3.2.5	Metadata availability	15
3.3	Making data interoperable	15
3.3.1	Interoperability of data/research outputs:	15
3.4	Increase data re-use.....	16
3.4.1	Reusability of data/research outputs:	16
3.4.2	Examples of Research Outputs:	16
4	Allocation of Resources.....	18
4.1	Data Management responsibilities	18
5	Data Security	19
6	Ethical Aspects.....	20
6.1	Ethical evaluation response	20
7	Other Issues	21
8	Data sets per WP	22
8.1	WP1 Project Coordination and Management	22
8.2	WP2 Innovation Management and Communications.....	24
8.3	WP3 Technical Coordination and Interoperability.....	26
8.4	WP4 Technical co-design and validation with research communities.....	28
8.5	WP5 Digital Twin Engine Infrastructure.....	30



8.6	WP6 Digital Twin Engine Core Modules	32
8.7	WP7 Digital Twin Engine Thematic Modules.....	34
9	Datasets per Use case	37
9.1	Lattice QCD Simulations - High Energy Physics use case (T4.1) and related thematic module (T7.1)	38
9.2	Detector simulation - High Energy Physics use case (T4.2) and related thematic module (T7.7)	44
9.3	VIRGO Noise detector - Astrophysics DT use case (T4.4) and related thematic module (T7.3).....	49
9.4	Noise simulation for radio astronomy DT use case (T4.3) and related thematic module (T7.2).....	57
9.5	Climate Change Future Projections of Extreme Events (storms & fire) use case (T4.5) and related thematic module (T7.4)	66
9.6	Early Warning for Extreme Events (floods & droughts) DT use case (T4.6) and related thematic module (T7.6)	74
9.7	Climate Change Impacts of Extreme Events (storms, fire, floods, drought) use case (T4.7) and the related thematic module (T7.5)	86
10	Conclusion.....	93

List of Tables

Table 1 Repository metadata	12
Table 2 Document metadata	13



Executive summary

This deliverable (D1.4), final version of the previous D1.2, outlines the framework through which interTwin has created, managed, and collected data over the three years of the project. It also details how this data has been made available for verification and reuse, as well as the curation and long-term storage practices adopted after the project's completion.

The Data Management Plan (DMP) has guided the design and implementation of data practices aligned with the FAIR principles (Findable, Accessible, Interoperable and Reusable), defined data security measures, and addressed ethical considerations related to data collection and generation.

The interTwin project has fully adhered to the Horizon Europe Open Science and FAIR principles, following the approach of making data “as open as possible, as closed as necessary.” Project beneficiaries have shared data in a way that maximises its value to partners and external stakeholders, while safeguarding the privacy and rights of third parties involved in data generation.

Throughout the project, data management has been carried out in full compliance with the EU General Data Protection Regulation (GDPR). Each dataset has been evaluated in terms of sensitivity, privacy, and security before making decisions on public accessibility. This final deliverable consolidates all updates and final practices, providing a comprehensive view of the data lifecycle and management choices adopted by the interTwin consortium.

1 Introduction

The final version of the DMP provides a comprehensive overview of the standards effectively used, the storage and publication mechanisms implemented, and the measures taken to ensure data verification and reuse.

D1.4 consolidates the data management frameworks, building upon the initial structure and preliminary guidelines. The Data Management Plan has been progressively updated to reflect the evolution of data handling within interTwin.

1.1 Purpose and scope of the document

Open Science practices have been consistently implemented throughout the interTwin project, in line with Horizon Europe principles. This section summarises the key actions and outcomes achieved over the course of the project.

Firstly, the use case methodology promoted open, collaborative work through co-design activities, based on the sharing of knowledge and software solutions across scientific communities and IT experts. Cooperation with Destination Earth¹ (DestinE) was established from Month 1, and open consultations were conducted online, targeting a broad audience including industry, SMEs, policymakers, and civil society.

Secondly, co-design and validation efforts were supported by releasing key exploitable results such as software, documentation, interoperability frameworks and requirements under licenses recognised as free by the Free Software Foundation and as open source by the Open Source Initiative. All communication and dissemination materials were published under open licenses (Creative Commons²) to maximise uptake and reuse. Peer-reviewed scientific publications were made openly accessible in line with the European Commission's Open Access policy, via self-archiving in Zenodo³ and, when possible, publication in open-access journals or platforms.

In addition, the Digital Twin Engine (DTE) was made openly accessible through the EOSC Portal⁴ for testing. It has been specifically designed to support Open Science through features such as the reuse and sharing of base/trained models, algorithms, and reference datasets, which can be enriched by users. The DTE also ensures reproducibility of research outputs (e.g., modelling and simulation workflows) through integrated provenance tracking.

¹ <https://stories.ecmwf.int/destination-earth/index.html>

² <https://creativecommons.org/licenses/>

³ <https://zenodo.org/communities/intertwin/?page=1&size=20>

⁴ <https://eosc-portal.eu/>



While interTwin has not directly managed all data assets, this document establishes a set of shared guidelines and best practices, which have been adopted by all parties involved in data management activities.

1.2 Structure of the document

This document comprises 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** concludes this deliverable.

2 Data Summary

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

2.1 Origin of the project outputs reused or newly generated

The Digital Twin Engine used 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 Digital Twin (DT) requirements. Research communities participated 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 also contributed to the development of thematic components specialised to adjacent downstream sectors.

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

- **Research communities from the long tail of science** > **How:** consulting EOSC task forces and the 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 collaboration with EuroGEO aims 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 was paid to the integration of services into external marketplaces such as the EOSC and AI4EU, as it required 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 was to provide support and specifications for turning this amalgamated, multi-party IP into efficient, well-managed software and services. For the reasons outlined above, the software code developed by the project was licensed under a permissive open-source license.

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

All users were 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, were from different sources.

2.3 Expected size of the data

The expected size of each dataset can be found in sections 8 and 9.

2.4 Formats of the project outputs

Sections 8 and 9 show the estimated datasets that interTwin produced/collected. This list was subjected to modifications, including the 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 were published on Zenodo. Software in development will be accessible on GitHub ⁵.

RA: MeerKAT data (including metadata) are stored in the database MeasurementSet.

GW and **CREM:** Data and research outputs were published on Zenodo.

3.1.2 Metadata

The metadata required by data repositories was 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: <ul style="list-style-type: none">• +0.1 – new version of draft• +1.0 – new version of approved document
Language	A language of the intellectual content of the resource.

⁵ <https://github.com/interTwin-eu>



D1.4 Final Data Management Plan

Keywords	A list of words that support the search within the repository service
Communities	A specific community in which the upload will appear.
License	Specifies the copyright status under which the upload is 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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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 through a repository service.
Keywords	A list of words that support the search within the Zenodo service
Abstract	Describing the document contents and main conclusions.

Metadata that was created by the communities is 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



D1.4 Final Data Management Plan

we adopted the identifiers provided by the resource. For workflows a DOI was minted through GitHub. Outputs submitted to Zenodo were assigned a DOI through this service. For code, we adopted the practices of the developer community for the software we are building on.

3.1.3.1 Search keywords for discovery

Keywords were created and then used to tag research output in Zenodo and in other registries or repositories (OpenAIRE⁶ and the EOSC catalogue).

3.2 Making data accessible

3.2.1 Accessibility of data/research outputs:

HEP: Fast simulated data and software were 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 are 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 to others. As a general rule, interTwin 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 repository⁷ to provide a managed central location for all materials.

In addition, public deliverables and publications are shared publicly via the [Zenodo platform](#) to increase the discoverability of the project outputs.

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

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

⁶ <https://www.openaire.eu/>

⁷ <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 was open source, access to project outputs was ensured well beyond the project lifespan. This ensured continuous uptake, and the possibility of creating new modifications and add-ons remained available for new and existing users and contributors even after the project ended.

As for the services (such as the DTE modules), they remained available on the EOSC Marketplace. Updates and maintenance were carried out by the interTwin Open Source Community that the project had set up and promoted. In addition, the project leveraged the Horizon Results Platform to increase visibility and potentially further exploit the results.

3.2.4 Standardised access protocol

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

3.2.5 Metadata availability

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

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

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

D1.4 Final Data Management Plan

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 was sought with the International Virtual Observatory Alliance (IVOA) for the evolution of standards.

CREM: Input and output data mostly follow the conventions for CF (Climate and Forecast) metadata. Output data are 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 new discoveries and research fields (e.g., Fast Radio Bursts). Public data is free to use.

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

CREM: Data are 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.



D1.4 Final Data Management Plan

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 were included in the project budget. These expenditures were required to cover a variety of specific data processing and data management operations, ranging from data collection and documentation to storage, preservation, distribution, and re-utilisation.

These operations were part of the Work Packages that processed the relevant data; therefore, the required effort was included within the respective WPs.

The expenses for long-term data preservation were minimal, as EGI Online Storage and Google Drive platforms were used. The use of Zenodo and GitHub (both free of charge) ensured that long-term data preservation costs remained manageable. When applicable, more accurate cost estimates were provided at later stages of the project.

4.1 Data Management responsibilities

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

WP leaders were in charge of organising the data processing and quality assurance that takes place inside the Work Package they are leading.

Task Leaders/Use Case leaders were in charge of the data compiled/produced throughout the operation of the task that they are responsible for. In addition to that, they also had to make sure that the data were properly prepared to be shared among the partners, and made publicly available, when applicable.

Data Processors were consortium partners who executed processing operations on the compiled/produced data.

Quality and Risk Manager monitored and supported the WP leaders, and Task Leaders/Use Case leaders in keeping the DMP confluence pages up to date. In addition, he reported the changes and processes via milestones and deliverables as specified in the Grant Agreement.

5 Data Security

All data gathered during the interTwin project was securely handled to ensure protection against loss and unauthorised access. Access to personal data was strictly limited to authorised individuals only.

All partners and beneficiaries responsible for processing personal data ensured that appropriate security measures were in place throughout the project. These included infrastructure-level controls such as backup policies, integrity checks, and access control mechanisms (identification, authentication, authorisation). In the event of a personal data breach, partners promptly notified their respective national supervisory authorities and any affected data subjects. All incidents were documented in accordance with regulatory requirements.

For open data management and to ensure both security and long-term preservation, interTwin relied on the EGI Document Repository, Zenodo, and GitHub.

Finally, the project finalised the Data Protection Management System (DPMS), with relevant information and procedures consolidated in the final version of the Data Management Plan.

6 Ethical Aspects

The project involved multi- and interdisciplinary collaboration to support software development and its application across a range of scientific domains. From the outset, it was acknowledged that the confidentiality and protection of personal data could require specific arrangements for data curation and handover to any follow-up activities.

The use of Artificial Intelligence (AI) modelling and analytics raises relevant ethical considerations. To address these, an ethics advisor was appointed to the project.

Based on the documentation available during the early stages, interTwin was identified as potentially involving unforeseen ethical issues, particularly related to personal data, privacy, and the use of AI. This was especially relevant as the project work plan anticipated the inclusion of new use cases from the Social Sciences, Humanities, and Life Sciences from Year 2 onwards.

Initially, the lack of a dedicated ethics management framework raised concerns, as ethical oversight was integrated under the administration and finance management task (T1.2), without clear involvement of experts in the ethics of emerging technologies. However, the appointment of the ethics advisor helped mitigate this gap and provided guidance on addressing ethical challenges as the project evolved.

6.1 Ethical evaluation response

To address findings from the Ethical evaluation of the project:

- Set-up the project Ethics Board to ensure 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 GDPR EU 2016/679.
 - Develop or review 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 an external Ethics advisor to support the work.

A more detailed report on this approach was provided in the deliverable D8.1 OEI – Requirements No.1, submitted in M12 of the project.

7 Other Issues

Within the interTwin project the following 2 Use Cases made 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, SARA0 data management policies.

8 Data sets per WP

8.1 WP1 Project Coordination and Management

WP/Task	WP1
Contact	Malgorzata Krakowian (EGL.eu)
Data Summary	
Data description: Types of data	<ol style="list-style-type: none"> 1. Project Documentation <ul style="list-style-type: none"> • 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	<ol style="list-style-type: none"> 1. Shared within the consortium and European Commission <ul style="list-style-type: none"> • 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

D1.4 Final Data Management Plan

WP/Task	WP1
	<p>everyone via the project website and Zenodo portal.</p> <ul style="list-style-type: none"> • Mailing list archive: only accessible by the mailing list members. <p>2. Shared with the Project Office and management boards to support work, as well as with the European Commission.</p>
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 the 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	
<i>Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?</i>	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



	<ul style="list-style-type: none"> • 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 <p>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</p>
Archiving and preservation	<p>Once the project is finished, all the WP2 information will be preserved by EGI Foundation for at least 5 years as well on the EC funding portal.</p> <p>Publications will be also kept on Zenodo Community</p>
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 <ul style="list-style-type: none"> 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	<ol style="list-style-type: none"> 1. Shared within the consortium and European Commission <ul style="list-style-type: none"> • 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 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	
<i>Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?</i>	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 <ul style="list-style-type: none"> • 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



WP/Task	WP4
Data sharing: Approach to sharing	<p>Shared within the consortium and European Commission</p> <ul style="list-style-type: none"> • 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	
<i>Do you, or will you, make use of other national/funder/sectorial/departmental</i>	EGI Foundation handles data according to the ISO 27000 standard for Information security management and GDPR.



WP/Task	WP4
<i>procedures for data management? If yes, which ones?</i>	

8.5 WP5 Digital Twin Engine Infrastructure

WP/Task	WP5
Contact	Daniele Spiga
Data Summary	
Data description: Types of data	<p>Project Documentation</p> <ul style="list-style-type: none"> • Meetings agenda and minutes • Meetings participation list • Presentations • Deliverables • Mailing list archive <p>Service logs generated by used WP5 services</p>
Data description: Origin of data	All the data was produced and provided by project members. This includes members that uses the services
Data description: Scale of data	< 1TB
Standards and metadata	txt, pdf, docsl, xlsx, pptx, json
Data sharing: Target groups	Project members and the EC Project office
Data sharing: Scientific Impact	N/A

D1.4 Final Data Management Plan

Data sharing: Approach to sharing	<p>Shared within the consortium and European Commission</p> <ul style="list-style-type: none"> • 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. <p>Service monitoring. Public information are made available via WP5 service web interface</p>
Archiving and preservation	N/A
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Daniele Spiga
How will long-term preservation be ensured?	Long term preservation beyond the project lifetime is not needed
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)?	Usual policies in the data centers contributing to the infrastructure
Will the data be safely stored in trusted repositories for long-term preservation and curation?	No
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departamental procedures for data management? If yes, which ones?	No



8.6 WP6 Digital Twin Engine Core Modules

WP/Task	WP6
Contact	Isabel Campos
Data Summary	
Data description: Types of data	<ul style="list-style-type: none"> • 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	Test data generated by the different tools under consideration
Data description: Scale of data	order of a few Gigabytes
Standards and metadata	N/A
Data sharing: Target groups	The target group is all project members and the EC Project office.
Data sharing: Scientific Impact	The value of this data is practically zero as it is mock data to demonstrate the tools

D1.4 Final Data Management Plan

Data sharing: Approach to sharing	<p>Shared within the consortium, the European Commission, and the public.</p> <ul style="list-style-type: none"> • Meeting information: The meeting agendas, attendees and minutes of the 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	No
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Isabel Campos
How will long-term preservation be ensured?	Long-term preservation is not needed, except for 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)?	Usual policies in the data centers contributing to the infrastructure
Will the data be safely stored in trusted repositories for long-term preservation and curation?	No
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departement	No



al procedures for data management? If yes, which ones?	
--	--

8.7 WP7 Digital Twin Engine Thematic Modules

WP/Task	WP7
Contact	Charis Chatzikyriakou
Data Summary	
Data description: Types of data	<ul style="list-style-type: none"> • 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.

D1.4 Final Data Management Plan

Data sharing: Approach to sharing	<p>Shared within the consortium, the European Commission, and the public:</p> <ul style="list-style-type: none"> • Meeting information: The meeting agendas, attendees and minutes of the 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 as on the 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 for 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 is required. Accounts and access management are 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	



D1.4 Final Data Management Plan

<p>Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?</p>	<p>EGI Foundation will take care of the data according to the ISO 27000 standard for Information security management and GDPR.</p>
--	--



9 Datasets per Use case

This final overview presents the DMP for the Use Case datasets generated within the interTwin project. For each dataset, it summarises the type and origin of the data, the metadata standards adopted, the data sharing approach and target audiences, as well as the strategies for long-term archival and preservation.

Throughout the project, beneficiaries have managed digital research data responsibly, in alignment with the **FAIR principles**, ensuring open access through trusted repositories, in accordance with the principle ‘as open as possible, as closed as necessary’. These research data management practices were applied to all data generated during the project, and were relevant to re-use data that contributed to research activities. This final version of the DMP reflects the concrete implementation of these principles and the established procedures adopted by the consortium.

During the Project, beneficiaries had to follow the main rules to maintain the DMP up to date.

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 broad re-use.

Beneficiaries must deposit the data in a trusted repository 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)

WP/Task	WP4/T4.1 WP7/T7.1
Contact	Gaurav Sinha Ray (CSIC)
<i>Established a DMP, addressing important aspects of RDM.</i>	<input checked="" type="checkbox"/> In place <input type="checkbox"/> In progress <input type="checkbox"/> None
Data Summary	
Will you re-use any existing data and what will you re-use it for?	Yes, previously generated lattice configurations were copied over to the data lake to validate its capacity and performance (see here for more details).
What types and formats of data will the project generate or re-use?	Binary files: Lattice field configurations. Text files: Metadata such as the values of the input parameters used to generate the configurations. Log files are text as well.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The sample data was chosen to be representative of a typical lattice project. If the data lake project is successful in deploying a prototype framework within which this sample data can be moved around with relative ease, then its potential utility will extend to the broader lattice (and HEP) community.
What is the expected size of the data that you intend to generate or re-use?	O(100)GB maximum.
What is the origin/provenance of the data either generated or re-used?	Monte Carlo simulations in HPC systems (LUMI in particular).
To whom might your data be useful ('data utility') outside your project?	Other researchers in the Lattice QCD community.



FAIR Data	
1.) Making data findable, including provisions for metadata	<p><i>Will data be identified by a persistent identifier?</i></p> <p>Yes</p>
	<p><i>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.</i></p> <p>At the end of the project we will have added metadata. The ILDG initiative has published standards regarding metadata for lattice data. We hope to see, and are working towards, the inclusion in the data lake prototype of the ability to search by reference to the ILDG metadata standard.</p>
	<p><i>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</i></p> <p>Yes</p>
	<p><i>Will metadata be offered in such a way that it can be harvested and indexed?</i></p> <p>Yes</p>
2.) Making data openly accessible	
a) Repository:	<p><i>Will the data be deposited in a trusted repository?</i></p> <p>Yes</p>
	<p><i>Have you explored appropriate arrangements with the identified repository where your data will be deposited?</i></p> <p>Yes</p>
	<p><i>Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?</i></p>



	Yes
b) Data:	<p><i>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.</i></p> <p>Yes</p>
	<p><i>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.</i></p> <p>N/A</p>
	<p><i>Will the data be accessible through a free and standardised access protocol?</i></p> <p>Yes</p>
	<p><i>If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?</i></p> <p>No</p>
	<p><i>How will the identity of the person accessing the data be ascertained?</i></p> <p>Using an OpenID Connect based AAI.</p>
	<p><i>Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?</i></p> <p>No</p>
c) Metadata:	<p><i>Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant</i></p>



	<p>Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?</p> <p>Yes</p>
	<p>How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?</p> <p>Hopefully indefinitely, contingent on sufficient funding for the ILDG initiative. Yes, the metadata should outlast the data residing on the data lake.</p>
	<p>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)?</p> <p>Yes, the openQxD repo is publically available and open source.</p>
3.) Making data interoperable	<p>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?</p> <p>No interdisciplinary applications.</p>
	<p>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?</p> <p>Yes</p>
	<p>Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?</p> <p>Yes</p>

<p>4.) Increase data re-use</p>	<p><i>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.)?</i></p> <p>README files</p>
	<p><i>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?</i></p> <p>The data will be open but there is no wide interest in the public domain to reuse it as it is either mock data or it is only useful to a small number of other lattice theorists.</p>
	<p><i>Will the data produced in the project be useable by third parties, in particular after the end of the project?</i></p> <p>Yes</p>
	<p><i>Will the provenance of the data be thoroughly documented using the appropriate standards?</i></p> <p>Yes</p>
	<p><i>Describe all relevant data quality assurance processes.</i></p> <p>We will use the FAIR evaluator as a validation service.</p>
	<p><i>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.</i></p>
<p>Other research outputs</p>	
<p>In addition to the management of data, are you also considering and planning</p>	<p><i>Such outputs can be either digital (e.g. software, workflows, protocols, models,</i></p>



D1.4 Final Data Management Plan

for the management of other research outputs that may be generated or re-used throughout the projects?	<i>etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?</i> No
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Gaurav Sinha Ray
How will long-term preservation be ensured?	<i>(costs and potential value, who decides and how what data will be kept and for how long)</i> No data preservation
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?	No
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	No
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	No
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	No



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.	<input checked="" type="checkbox"/> In place <input type="checkbox"/> In progress <input type="checkbox"/> 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



	<p><i>discipline, please outline what type of metadata will be created and how.</i></p> <p>Yes</p>
	<p><i>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</i></p> <p>For 3DGAN case: no, in CaloINN case: we re-use an existing public dataset.</p>
	<p><i>Will metadata be offered in such a way that it can be harvested and indexed?</i></p> <p>Yes, depends on how the procedures for harvesting and indexing are going to be set up</p>
2.) Making data openly accessible	
a) Repository:	<p><i>Will the data be deposited in a trusted repository?</i></p> <p>Zenodo</p>
	<p><i>Have you explored appropriate arrangements with the identified repository where your data will be deposited?</i></p> <p>This is the default choice</p>
	<p><i>Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?</i></p> <p>Yes</p>
b) Data:	<p><i>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.</i></p> <p>Yes</p>



	<p><i>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.</i></p> <p>It's unlikely that the data produced will be subject to an embargo, the details are still being defined.</p>
	<p><i>Will the data be accessible through a free and standardised access protocol?</i></p> <p>Whatever is provided by Zenodo</p>
	<p><i>If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?</i></p>
	<p><i>How will the identity of the person accessing the data be ascertained?</i></p> <p>Not implemented</p>
	<p><i>Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?</i></p> <p>No</p>
c) Metadata:	<p><i>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?</i></p> <p>Yes</p>
	<p><i>How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?</i></p> <p>Zenodo policies</p>
	<p><i>Will documentation or reference about any software be needed to access or read or process the data be included?</i></p>



	<p><i>Will it be possible to include the relevant software (e.g. in open-source code)?</i></p> <p>Yes</p>
3.) Making data interoperable	<p><i>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?</i></p> <p>HEP community</p>
	<p><i>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?</i></p>
	<p><i>Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?</i></p> <p>No</p>
4.) Increase data re-use	<p><i>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.)?</i></p> <p>Upload to the same Zenodo record</p>
	<p><i>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?</i></p> <p>Yes</p>

	<p><i>Will the data produced in the project be useable by third parties, in particular after the end of the project?</i></p> <p>No</p>
	<p><i>Will the provenance of the data be thoroughly documented using the appropriate standards?</i></p> <p>Yes, HEP standard</p>
	<p><i>Describe all relevant data quality assurance processes.</i></p> <p>For 3DGAN: the dataset produced with reliable software (GEANT4), for CaloINN: we re-use an existing dataset known as for banchmarking.</p>
	<p><i>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.</i></p>
Other research outputs	
<p>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?</p>	<p><i>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?</i></p> <p>Software</p>
Allocation of resources	
<p>Who will be responsible for data management in your WP/Task?</p>	<p>Vera Maiboroda</p>
<p>How will long-term preservation be ensured?</p>	<p><i>(costs and potential value, who decides and how what data will be kept and for how long)</i></p> <p>Datasets are available on Zenodo, without an expiration date.</p>
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?	<i>Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).</i> 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/departamental procedures for data management? If yes, which ones?	<i>Please list and briefly describe them.</i> 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)
<i>Established a DMP, addressing important aspects of RDM.</i>	<ul style="list-style-type: none"> In place ✓ In progress Non
Data Summary	

<p>Will you re-use any existing data and what will you re-use it for?</p>	<p><i>State the reasons if re-use of any existing data has been considered but discarded.</i></p> <p>We will re-use Virgo data from past observing runs to study the features of transient noise and to develop the GAN architecture to be used in the DT.</p>
<p>What types and formats of data will the project generate or re-use?</p>	<p>The project will use files in hdf5 and gwf (Gravitational Wave Frame) formats. The latter is a proprietary format, the specification can be found here. The project will generate data containing the trained GenNN models (.pt files) and the DT output metadata (csv, json, png).</p>
<p>What is the purpose of the data generation or re-use and its relation to the objectives of the project?</p>	<p>Existing detector data will be used to characterise transient noise in different readout channels of the interferometer and as input to the GenNN 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 <i>strain</i> readout channel (sensitive to the astrophysical signal) from the response in the <i>auxiliary</i> channels (containing data from detector sensors only). The DT output will contain:</p> <ul style="list-style-type: none"> • The generated strain readout channel's spectrogram • The subtraction of the generated from the real strain readout channel's spectrogram • Veto flags (whether the aforementioned subtraction is free of transient signals above a specified threshold or not) • Metadata on the subtraction, such as maximum intensity value (signal to noise ratio) and number of pixels above threshold in the spectrogram • Accuracy metrics <p>Veto flags are intended to be passed to downstream detection and parameter estimation pipelines (not part of the DT),</p>



	while Metadata, generated and subtracted spectrograms are intended for expert validation purposes.
What is the expected size of the data that you intend to generate or re-use?	We expect a size of a few TB of detector data to be made available for noise characterisation studies. These 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 GenNN model. We expect a few GB of data for DT output and GenNN model weights.
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 GenNN models and the DT output can be useful for the Virgo collaboration.
FAIR Data	
1.) Making data findable, including provisions for metadata	<i>Will data be identified by a persistent identifier?</i> Both the trained GenNN models and the DT outputs are identified by persistent timestamps identifiers on MLFlow and TensorBoard.
	<i>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.</i> Metadata will be provided to allow discovery. Output metadata will reflect the metadata of input detector data, namely:GPS time and read-out channels and the DT output data, namely veto



	<p>flags, maximum SNR of cleaned data, Area (in pixels) of uncleaned data.</p> <p><i>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</i></p> <p>Yes, including network and preprocessing parameters.</p> <p><i>Will metadata be offered in such a way that it can be harvested and indexed?</i></p> <p>Metadata is indexed based on GPS time</p>
2.) Making data openly accessible	
<p>a) Repository:</p>	<p><i>Will the data be deposited in a trusted repository?</i></p> <p>Input data is stored in the Virgo RSE at the Vega EuroHPC. The RSE is part of a private Virgo Virtual Organisation (virgo.intertwin.eu), created to restrict data access to only authorised people who are part of the Virgo community.</p> <p>Output data will be stored in Zenodo.</p> <p><i>Have you explored appropriate arrangements with the identified repository where your data will be deposited?</i></p> <p>No.</p> <p><i>Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?</i></p> <p>No.</p>
<p>b) Data:</p>	<p><i>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</i></p>



	<p><i>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.</i></p> <p>Input data is available to Virgo collaboration members via INFN CNAF. Data access to non-members will not be provided.</p>
	<p><i>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.</i></p> <p>It is unlikely that the data produced will be subject to an embargo.</p>
	<p><i>Will the data be accessible through a free and standardised access protocol?</i></p> <p>Yes.</p>
	<p><i>If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?</i></p> <p>No restrictions are foreseen.</p>
	<p><i>How will the identity of the person accessing the data be ascertained?</i></p> <p>Through Virgo's federated identity providers.</p>
	<p><i>Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?</i></p> <p>No.</p>
c) Metadata:	<p><i>Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why.</i></p>



	<p>Yes.</p> <p><i>Will metadata contain information to enable the user to access the data?</i></p> <p>Yes.</p> <p><i>How long will the data remain available and findable?</i></p> <p>For the duration of the project.</p> <p><i>Will metadata be guaranteed to remain available after data is no longer available?</i></p> <p>No.</p> <p><i>Will documentation or reference about any software be needed to access or read or process the data be included?</i></p> <p>Yes.</p> <p><i>Will it be possible to include the relevant software (e.g. in open-source code)?</i></p> <p>Yes, both software and documentation are available at https://github.com/interTwin-eu/DT-Virgo-dags/tree/main/Final_Release</p>
<p>3.) Making data interoperable</p>	<p><i>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?</i></p> <p>We follow the prescriptions from the Virgo collaboration for the sharing within the community. We do not expect data to be interoperable across disciplines.</p> <p><i>In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly</i></p>

	<p><i>used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them?</i></p> <p>Not applicable.</p> <p><i>Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?</i></p> <p>Yes. Trained models and output data will contain references to the input data, such as GPS time and data preprocessing parameters (whitening, frequency range, time-frequency resolution, duration).</p>
<p>4.) Increase data re-use</p>	<p><i>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.)?</i></p> <p>We provide detailed documentation accompanying the code in the GitHub repository https://github.com/interTwin-eu/DT-Virgo-dags/tree/main/Final_Release</p> <p><i>Will your data be made openly available in the public domain to permit the widest re-use possible?</i></p> <p>Only to Virgo collaboration members.</p> <p><i>Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement?</i></p> <p>Yes.</p> <p><i>Under which license?</i></p> <p>Virgo data are released publicly under CC BY 4.0 license. Output data, like the</p>

	trained ML models, will be released with MIT license.
	<p><i>Will the data produced in the project be usable by third parties, in particular after the end of the project?</i></p> <p>Data could be re-used by the Virgo collaboration and possibly also by the Einstein Telescope collaboration.</p>
	<p><i>Will the provenance of the data be thoroughly documented using the appropriate standards?</i></p> <p>Yes.</p>
	<p><i>Describe all relevant data quality assurance processes.</i></p> <p>For the trained ML model, validation will be performed following ML best practices based on well-known metrics (e.g., MAE, MSE, RMSE).</p>
	<p><i>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.</i></p>
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?	<p><i>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?</i></p> <p>Not applicable.</p>
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?	<i>(costs and potential value, who decides and how what data will be kept and for how long)</i> There will be no long term data preservation.
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?	No.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	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/departamental procedures for data management? If yes, which ones?	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)

D1.4 Final Data Management Plan

<i>Established a DMP, addressing important aspects of RDM.</i>	✓ In place • In progress • Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	The datasets we are working with are constantly reused both for trying different ML approaches for classification and for studies of their properties in order to create reliable simulations.
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?	Most of the current testing is done with datasets of sizes of tens of GB, however the raw datasets from the telescopes we use have sizes of tens of TB.
What is the origin/provenance of the data either generated or re-used?	"Real life" datasets originate from two telescopes, MPIfR-operated Effelsberg 100m radio telescope and the MeerKAT radio astronomical array operated by the South African Radio Astronomy Observatory (SARAO). Simulated data is generated by our own digital twins.
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	<p><i>Will data be identified by a persistent identifier?</i></p> <p>Yes, the data sets are designated in accordance with the observatory/project standards.</p>
	<p><i>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.</i></p> <p>The data already have headers standard for radio astronomy, identifying source, project, epoch, various parameters etc.</p>
	<p><i>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</i></p> <p>When archived, the necessary keywords will be added, again in accordance with the radio astronomy standards.</p>
	<p><i>Will metadata be offered in such a way that it can be harvested and indexed?</i></p> <p>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.</p>
2.) Making data openly accessible	
a) Repository:	<p><i>Will the data be deposited in a trusted repository?</i></p> <p>Yes, the standard archive for the given type of telescope/project.</p>
	<p><i>Have you explored appropriate arrangements with the identified</i></p>

	<p><i>repository where your data will be deposited?</i></p> <p>There is no need, the standard procedure for radio astronomical data will be followed.</p> <p><i>Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?</i></p> <p>Yes.</p>
<p>b) Data:</p>	<p><i>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.</i></p> <p>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).</p> <p><i>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.</i></p> <p>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</p>

	<p>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.</p>
	<p><i>Will the data be accessible through a free and standardised access protocol?</i></p> <p>It will be kept in an archive related to the telescope/organization that performed the observations.</p>
	<p><i>If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?</i></p> <p>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.</p>
	<p><i>How will the identity of the person accessing the data be ascertained?</i></p> <p>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) <i>and</i> whose access to the particular machine has been cleared can have direct access to it.</p>
	<p><i>Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?</i></p> <p>No, the teams are small enough to be able to solve all the access issues by personally contacting the responsible people.</p>



<p>c) Metadata:</p>	<p><i>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?</i></p> <p>Yes</p> <p><i>How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?</i></p> <p>In principle, indefinitely, or at least on a scale of decades.</p> <p><i>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)?</i></p> <p>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.</p>
<p>3.) Making data interoperable</p>	<p><i>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?</i></p> <p>We follow standard procedures common in the whole field of radio astronomy.</p> <p><i>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</i></p>

	<p><i>vocabularies to allow reusing, refining, or extending them?</i></p> <p>In the parts where we would go beyond the standard radio-astronomical procedures, it is unclear at this point.</p> <p><i>Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?</i></p> <p>Yes, where applicable.</p>
<p>4.) Increase data re-use</p>	<p><i>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.)?</i></p> <p>Where applicable, help, readme files and other materials will accompany the data and the software. We also plan to publish all the relevant findings in professional journals, providing as much explanations and additional materials (like links to files or references to repositories) as possible.</p> <p><i>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?</i></p> <p>The data and everything related will be available on the standard academic basis: free distribution and use provided proper references are given.</p> <p><i>Will the data produced in the project be usable by third parties, in particular after the end of the project?</i></p> <p>Yes.</p>

	<p><i>Will the provenance of the data be thoroughly documented using the appropriate standards?</i></p> <p>Yes.</p>
	<p><i>Describe all relevant data quality assurance processes.</i></p> <p>Standard QA procedures apply to all radio-astronomical data, they are a part of the data acquisition process at the observatories. QA for the synthetic data is performed as a part of its generation by the DT.</p>
	<p><i>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.</i></p> <p>Academic institutions and their employees involved in this task already follow high standards in this respect.</p>
Other research outputs	
<p>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?</p>	<p><i>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?</i></p> <p>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.</p>
Allocation of resources	
<p>Who will be responsible for data management in your WP/Task?</p>	<p>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</p>

D1.4 Final Data Management Plan

	project it is and, later, by the relevant archive and the institution that runs it.
How will long-term preservation be ensured?	<i>(costs and potential value, who decides and how what data will be kept and for how long)</i> 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 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?	<i>Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).</i> 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	<i>Please list and briefly describe them.</i>



procedures for data management? If yes, which ones?	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)
<i>Established a DMP, addressing important aspects of RDM.</i>	✓ In place <ul style="list-style-type: none"> • In progress • Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	<p><i>State the reasons if re-use of any existing data has been considered but discarded.</i></p> <p>Data from public repositories will be used as input for the DT on climate future projection of extreme weather events:</p> <ul style="list-style-type: none"> • CMIP6 climate projection data • ERA5 reanalysis data • Fire Danger Indices data

	<ul style="list-style-type: none"> International Best Track Archive for Climate Stewardship (IBTrACS) tropical cyclones observation data
What types and formats of data will the project generate or re-use?	<ul style="list-style-type: none"> 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?	<p>Data will be (re-)used to develop the DTs on extreme weather events, in particular for:</p> <ul style="list-style-type: none"> 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
What is the origin/provenance of the data either generated or re-used?	<ul style="list-style-type: none"> 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?	<ul style="list-style-type: none"> Scientists interested on extreme events studies
FAIR Data	
1.) Making data findable, including provisions for metadata	<p><i>Will data be identified by a persistent identifier?</i></p> <p>Input to the applications are identified by PID in the related repositories. Scientific results do not have a PID because these are generated on-demand</p>
	<p><i>Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case</i></p>



	<p><i>metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.</i></p> <p>Scientific data is in NetCDF format with related metadata defined according to CF-Conventions stored in the header.</p>
	<p><i>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</i></p> <p>NetCDF data conform to the CF-Conventions for metadata</p>
	<p><i>Will metadata be offered in such a way that it can be harvested and indexed?</i></p> <p>Metadata can be harvested from the NetCDF files' headers. Input to the applications are indexed in the respective catalogues (e.g., ESGF, Copernicus, etc.)</p>
2.) Making data openly accessible	
a) Repository:	<p><i>Will the data be deposited in a trusted repository?</i></p> <p>Input data is available from trusted repositories. Data resulting from the applications are not deposited.</p>
	<p><i>Have you explored appropriate arrangements with the identified repository where your data will be deposited?</i></p> <p>Not needed</p>
	<p><i>Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?</i></p> <p>Yes</p>
b) Data:	<p><i>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</i></p>



	<p><i>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.</i></p> <p>All data for the demonstrators is openly available.</p>
	<p><i>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.</i></p> <p>Not applicable</p>
	<p><i>Will the data be accessible through a free and standardised access protocol?</i></p> <p>Yes</p>
	<p><i>If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?</i></p> <p>Input data is openly available from the related repositories.</p>
	<p><i>How will the identity of the person accessing the data be ascertained?</i></p> <p>For data from Copernicus, such as ERA5, an account on C3S is required. Other data is openly accessible.</p>
c) Metadata:	<p><i>Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?</i></p> <p>No</p>
	<p><i>Will metadata be made openly available and licensed under a public domain dedication CC0, as per the</i></p>



	<p><i>Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?</i></p> <p>All metadata for the scientific dataset is openly available and described in related documentation (e.g., CF convention for CMIP6)</p> <p><i>How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?</i></p> <ul style="list-style-type: none"> • Data is expected to be available on the respective repositories in the long term • Metadata is expected to be available on the respective indexes in the long term <p><i>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)?</i></p> <p>Yes. Libraries of the software packages developed in the project are described in deliverables and are publicly available on GitHub. All the dependencies for handling the data are also available as open source.</p>
<p>3.) Making data interoperable</p>	<p><i>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?</i></p> <p>Data follows the CF-convention (to the maximum extent possible) for the metadata and is stored as NetCDF format (both standard-de-facto in the community).</p>

	<p><i>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?</i></p> <p>Not applicable</p>
	<p><i>Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?</i></p> <p>Not applicable</p>
4.) Increase data re-use	<p><i>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.)?</i></p> <p>Via Jupyter Notebook and readme files openly available on GitHub</p>
	<p><i>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?</i></p> <p>The main project outputs are software packages made available on Github under different FOSS licenses. Data generated using these packages can be public domain according to each user</p>
	<p><i>Will the data produced in the project be usable by third parties, in particular after the end of the project?</i></p> <p>The software and notebooks will remain openly available on Github after the project ends.</p>



	<p><i>Will the provenance of the data be thoroughly documented using the appropriate standards?</i> Provenance information is provided according to W3C PROV standards.</p> <p><i>Describe all relevant data quality assurance processes.</i> The scientific output will follow validation procedures usually adopted in the domain. For the ML models, validation will be performed following ML best practices based on well-known metrics (e.g., MAE, MSE, RMSE) and according to specific use cases.</p> <p><i>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.</i></p>
Other research outputs	
<p>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?</p>	<p><i>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?</i></p> <p>We include in this category trained ML models, software, workflows and provenance documents. We tried to address FAIR principles to the best extent possible (e.g., software and workflows history tracked on GitHub, ML models training tracked on MLflow, provenance documents on yProv). With respect to ML models, publications are ongoing and we expect to publish on open repositories (e.g., Zenodo) the ML models used in the study and based on the project's software.</p>
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?	<p><i>(costs and potential value, who decides and how what data will be kept and for how long)</i></p> <p>Results are generated on-demand by the applications so this does not apply.</p>
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)?	Results are generated on-demand by the applications so this does not apply.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Results are generated on-demand by the applications so this does not apply.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	<p><i>Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).</i></p> <p>No</p>
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/departamental procedures for data management? If yes, which ones?	<p><i>Please list and briefly describe them.</i></p> <p>No</p>

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	Bjorn Backeberg (deltares)
<i>Established a DMP, addressing important aspects of RDM.</i>	✓ In place <ul style="list-style-type: none"> • In progress • Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	<p>Yes, a variety of existing data will be used to set up and run the models, including:</p> <ul style="list-style-type: none"> • Topographic data • Hydrological data • Meteorological data • Land use and land cover data • Hydraulic data (river and channel cross-section, infrastructure data) • Geospatial data • Sentinel-1 σ₀ backscatter data (Equi-7 Grid, 20m pixel spacing) for microwave backscatter based flood mapping • PLIA values of those Sentinel-1 data • Parameters of an harmonic model that fits the backscatter time series over land of the corresponding pixel • Socio-economic data • Historical event and damage reports • Climate data • Geotechnical data (e.g. soil properties) • Infrastructure and building data

<p>What types and formats of data will the project generate or re-use?</p>	<p>Data types include Raster and Vector data.</p> <p>Formats include:</p> <ul style="list-style-type: none"> • GeoTIFF • NetCDF • GRIB • HDF and HDF5 • Shapefile • GeoJSON • CSV • Zarr
<p>What is the purpose of the data generation or re-use and its relation to the objectives of the project?</p>	<p>Terrain and Hydrology Analysis: Digital Elevation Models (DEMs), contour maps, and soil moisture data are used to understand the landscape, water flow paths, and infiltration rates, which are essential for identifying flood-prone areas and modeling flood extents.</p> <p>Weather and Climate Monitoring: Rainfall data, weather forecasts, and climate models are utilized to predict precipitation, assess future weather conditions, and understand long-term climate impacts on flood risks.</p> <p>Land and Water Body Management: Land use maps, vegetation cover data, and river cross-sections provide insights into how land cover and water body characteristics influence runoff, infiltration, and flood dynamics, helping in flood mitigation planning.</p> <p>Infrastructure and Impact Assessment: Infrastructure data, building footprints, and critical infrastructure information are crucial for assessing the impact of floods on buildings and essential services, aiding in the development of protection and emergency response strategies.</p>



	<p>Historical and Real-Time Flood Analysis: Historical flood data and satellite imagery are used to re-analyze past flood events and thus to validate flood models, and to monitor real-time flood situations, which are key for accurate flood prediction and response.</p> <p>Socioeconomic Evaluation: Population density maps and economic data help in understanding the potential human and economic impacts of flooding, supporting risk assessment and resource allocation for flood defence and recovery efforts.</p>
What is the expected size of the data that you intend to generate or re-use?	The total data size could range from approximately 200 GB to 1 TB, assuming a moderately large region with detailed, high-resolution data. For a more specific, localized study area or lower resolution data, the size could be on the lower end of this range. Conversely, for a larger region or higher resolution data, the size could exceed 1 TB.
What is the origin/provenance of the data either generated or re-used?	<ul style="list-style-type: none"> • European Space Agency (ESA) • European Environment Agency (EEA) • European Centre for Medium-Range Weather Forecasts (ECMWF) • European Commission Joint Research Centre (JRC) • European Soil Data Centre (ESDAC) • European Forest Institute (EFI) • European Drought Observatory (EDO) • European Environment Information and Observation Network (EIONET) • Copernicus Programme (including Sentinel satellites)

	<ul style="list-style-type: none"> • European Flood Awareness System (EFAS) • Alpine Drought Observatory (ADO) • National Mapping Agencies (e.g., Ordnance Survey - UK, IGN - France, Instituto Geográfico Nacional - Spain) • National Meteorological and Hydrological Services (e.g., Met Office - UK, Météo-France - France, AEMET - Spain) • National Statistical Offices (e.g., Eurostat for pan-European data, INSEE - France, INE - Spain) • National Environmental Agencies (e.g., Environment Agency - UK, Environment Agency Austria, Environment Agency - Germany) • Hydrological Survey Agencies (e.g., Bundesanstalt für Gewässerkunde - Germany, Institut français de recherche pour l'exploitation de la mer - France, Centro de Estudos e Investigação Científica - Portugal) • Universities and Research Institutions across Europe • TU Wien / EODC • IPCC (Intergovernmental Panel on Climate Change) • OpenStreetMap • National land use and planning agencies • Local and national government agencies • Utilities and infrastructure companies
<p>To whom might your data be useful ('data utility') outside your project?</p>	<p>Government Agencies:</p> <ul style="list-style-type: none"> • National and local government agencies responsible for disaster management, emergency response, land use planning, environmental protection, and

	<p>climate change mitigation policies.</p> <ul style="list-style-type: none"> • Water resource management authorities responsible for managing rivers, dams, and flood control infrastructure. • Environmental agencies concerned with preserving natural habitats and ecosystems affected by floods. • Copernicus Emergency Management Service (CEMS) <p>Emergency Services:</p> <ul style="list-style-type: none"> • Fire departments, police, and medical services involved in emergency response and evacuation procedures during flood events. <p>Infrastructure Operators:</p> <ul style="list-style-type: none"> • Utilities such as water, electricity, and telecommunications providers that need to safeguard critical infrastructure from flood damage. <p>Insurance Companies:</p> <ul style="list-style-type: none"> • Insurers interested in assessing and underwriting flood risks to properties and infrastructure. <p>Urban Planners and Developers:</p> <ul style="list-style-type: none"> • City planners, architects, and developers seeking to incorporate flood risk considerations into urban development projects and land use plans. <p>Community Organizations:</p> <ul style="list-style-type: none"> • Non-governmental organizations (NGOs), community groups, and volunteer organizations involved in community resilience building,
--	---

	<p>disaster preparedness, and response efforts.</p> <p>Businesses and Industries:</p> <ul style="list-style-type: none"> • Businesses operating in flood-prone areas, such as agriculture, manufacturing, and tourism, interested in assessing and mitigating their flood risks. <p>Researchers and Academia:</p> <ul style="list-style-type: none"> • Scientists, researchers, and academic institutions studying flood dynamics, climate change impacts, and resilience strategies. <p>International Organizations:</p> <ul style="list-style-type: none"> • Organizations like the United Nations, World Bank, and European Union interested in supporting global efforts to mitigate flood risks, especially in vulnerable regions. <p>General Public:</p> <ul style="list-style-type: none"> • Residents living in flood-prone areas interested in understanding their flood risk, accessing flood warnings, and participating in community resilience initiatives.
FAIR Data	
<p>1.) Making data findable, including provisions for metadata</p>	<p><i>Will data be identified by a persistent identifier?</i></p> <p>No - data generated from what-if scenarios will be for local decision makers, e.g. city planners.</p> <p><i>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.</i></p>

	<ul style="list-style-type: none"> • NetCDF files will conform to the CF-Conventions with rich metadata. • STAC compliant metadata has been created for the Sentinel-1 based input data to make them findable at the used EODC's backend.
	<p><i>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</i></p> <p>Yes, for NetCDF output. Also, using the STAC standard will provide keywords.</p>
	<p><i>Will metadata be offered in such a way that it can be harvested and indexed?</i></p> <p>Metadata can be harvested from the NetCDF files headers and from the STAC API.</p>
2.) Making data openly accessible	
a) Repository:	<p><i>Will the data be deposited in a trusted repository?</i></p> <p>No</p>
	<p><i>Have you explored appropriate arrangements with the identified repository where your data will be deposited?</i></p> <p>No</p>
	<p><i>Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?</i></p> <p>Not applicable</p>
b) Data:	<p><i>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</i></p>

	<p><i>to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.</i></p> <p>All data of the demonstrators will be openly available. They are for demonstration purposes only, not for decision-making.</p>
	<p><i>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.</i></p> <p>Not applicable</p>
	<p><i>Will the data be accessible through a free and standardised access protocol?</i></p> <p>Yes</p>
	<p><i>If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?</i></p> <p>Accessing an open dataset will require a login from a trustworthy AAI. The user requires a quota to process data at the backend.</p>
	<p><i>How will the identity of the person accessing the data be ascertained?</i></p> <p>Accessing an open dataset will require login from a trustworthy AAI.</p>
c) Metadata:	<p><i>Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?</i></p> <p>No</p>
	<p><i>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?</i></p>



	<p>Yes</p> <p><i>How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?</i></p> <ul style="list-style-type: none"> • When a new version of a data collection is available, the latest version will only be made available via the metadata catalogue. • Collections where data is no longer available will be labeled as such, item level metadata will be deleted in the catalogue. • End user request to restore the data depends on the individual data management strategy chosen. However, deletion of an entire data collection is not common, but if it is deleted, the data will be removed from all storage tiers including any backups. <p><i>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)?</i></p> <p>Yes. GitHub libraries of the developed software packages are listed in the project's deliverables and are made publicly available.</p>
<p>3.) Making data interoperable</p>	<p><i>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?</i></p> <p>For the output in NetCDF, it will only follow the CF-convention. This is what is used in the community.</p>

	<p><i>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?</i></p> <ul style="list-style-type: none"> • For the NetCDF output, it is not known if there exists mappings or not for the CF-Convention. • The stored metadata of the input geodata follows the STAC protocol. • The flood monitoring workflows follow the openEO syntax⁸ for allowing the workflow to be reusable at 3rd-party backends. <p><i>Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?</i></p> <p>The NetCDF output will include references to the original input data.</p>
<p>4.) Increase data re-use</p>	<p><i>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.)?</i></p> <p>We will provide documentation via Jupyter Notebooks and documentation on the code repositories. GitHub libraries of the software are listed in the project's deliverables and made publicly available.</p> <p><i>Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be</i></p>

⁸ <https://doi.org/10.3390/rs13061125>

	<p><i>licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?</i></p> <p>The main project output are software packages made available on Github under various open-source licenses. Data generated using these packages can be public domain according to each user.</p>
	<p><i>Will the data produced in the project be usable by third parties, in particular after the end of the project?</i></p> <p>The code on Github will remain available after project end and any data generated by it will be usable. Note that any data is for demonstration purposes only.</p>
	<p><i>Will the provenance of the data be thoroughly documented using the appropriate standards?</i></p> <p>The NetCDF output will only provide limited provenance information, as full provenance is not in force yet in the community. Any additional provenance is not in scope for this use case.</p>
	<p><i>Describe all relevant data quality assurance processes.</i></p> <p>The objective of this demonstrator is to demonstrate that users can easily set up flood risk models anywhere on Earth, quality assurance of the output is out of scope for this work.</p>
	<p><i>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.</i></p>
Other research outputs	
<p>In addition to the management of data, are you also considering and planning for the management of other research</p>	<p><i>Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials,</i></p>



outputs that may be generated or re-used throughout the projects?	<i>antibodies, reagents, samples, etc.) Are those also following FAIR principles?</i> The workflows that are developed for the use cases are publicly available on GitHub. We are not considering and planning for the management of other research outputs in this demonstrator after the project's funding period.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Every partner producing data in T4.6
How will long-term preservation be ensured?	<i>(costs and potential value, who decides and how what data will be kept and for how long)</i> Only demonstration data will be generated, there is no plan to ensure preservation.
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)?	3rd-party input data is transferred from Copernicus services and can be re-transferred again in case of data loss.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Not applicable.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	<i>Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).</i> Yes, the data generated are for demonstration purposes only and should not be used for decision-making. Furthermore, the use cases do not cover real-time information that potentially could be harmful to individuals.
Will informed consent for data sharing and long-term preservation be included	Not applicable.



in questionnaires dealing with personal data?	
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departamental procedures for data management? If yes, which ones?	<i>Please list and briefly describe them.</i> N/A

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)
<i>Established a DMP, addressing important aspects of RDM.</i>	<input checked="" type="checkbox"/> In place <ul style="list-style-type: none"> <input type="checkbox"/> In progress <input type="checkbox"/> Non
Data Summary	
Will you re-use any existing data and what will you re-use it for?	Yes, CMIP6 data for several climate models and scenarios. This is an existing dataset that will be used as input.
What types and formats of data will the project generate or re-use?	NetCDF
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	<ul style="list-style-type: none"> CMIP6 data is needed to run the extreme events climate change impacts DT. Output dataset is the results of the calculations of the characteristics of climate extremes.
What is the expected size of the data that you intend to generate or re-use?	<ul style="list-style-type: none"> On the order of 300 Gb for CMIP6 input data. On the order of 10 Mb for output data.



What is the origin/provenance of the data either generated or re-used?	<ul style="list-style-type: none"> • CMIP6 climate simulations from ESGF data infrastructure. • Output dataset is created by the DT application.
To whom might your data be useful ('data utility') outside your project?	Scientific Researchers using climate data and working on the impacts of extreme events.
FAIR Data	
1.) Making data findable, including provisions for metadata	<i>Will data be identified by a persistent identifier?</i> No, because the output dataset is on-demand.
	<i>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.</i> NetCDF files will conform to the CF-Conventions with rich metadata.
	<i>Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?</i> Yes, for NetCDF output.
	<i>Will metadata be offered in such a way that it can be harvested and indexed?</i> Metadata can be harvested from the NetCDF files headers.
2.) Making data openly accessible	
a) Repository:	<i>Will the data be deposited in a trusted repository?</i> No.
	<i>Have you explored appropriate arrangements with the identified</i>



	<p><i>repository where your data will be deposited?</i></p> <p>No repository is needed for output data.</p>
	<p><i>Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?</i></p> <p>No temporary datasets will be generated by users.</p>
b) Data:	<p><i>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.</i></p> <p>All data can be openly accessible, it will be the decision of the application users.</p>
	<p><i>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.</i></p> <p>N/A</p>
	<p><i>Will the data be accessible through a free and standardised access protocol?</i></p> <p>Yes</p>
	<p><i>If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?</i></p> <p>No</p>

	<p><i>How will the identity of the person accessing the data be ascertained?</i></p> <p>Accessing open dataset will require login from a trustworthy AAI, on platforms such as EUDAT B2DROP.</p> <p><i>Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?</i></p> <p>No</p>
<p>c) Metadata:</p>	<p><i>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?</i></p> <p>All metadata will be openly available.</p> <p><i>How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?</i></p> <p>It will depend on the decision of users.</p> <p><i>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)?</i></p> <ul style="list-style-type: none"> • The CMIP6 input data is in NetCDF, and a very large number of free tools and software are available to access it. • The output data is in xarray and can be read by several python libraries.
<p>3.) Making data interoperable</p>	<p><i>What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data</i></p>



	<p><i>interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?</i></p> <p>For the climate indices database, it will only follow the CF-convention. This is what is used in the community.</p> <p><i>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?</i></p> <p>For the climate indices database, it is not known if there exist mappings or not for the CF-Convention.</p> <p><i>Will your data include qualified references¹ to other data (e.g. other data from your project, or datasets from previous research)?</i></p> <p>The climate indices database will include references to the original CMIP data.</p>
<p>4.) Increase data re-use</p>	<p><i>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.)?</i></p> <p>For the NetCDF output, there already exist some freely available notebooks.</p> <p><i>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</i></p>

	<p>out in the Grant Agreement? Under which license?</p> <p>The NetCDF output can be in the public domain according to each user.</p>
	<p>Will the data produced in the project be usable by third parties, in particular after the end of the project?</p> <p>The NetCDF output will be usable by third parties after the end of the project, and even during the project, according to each user.</p>
	<p>Will the provenance of the data be thoroughly documented using the appropriate standards?</p> <p>The NetCDF output will only provide limited provenance information, as full provenance is not in force yet in the climate community.</p>
	<p>Describe all relevant data quality assurance processes.</p> <p>The NetCDF output will be generated by the DT that has been validated.</p>
	<p>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.</p>
Other research outputs	
<p>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?</p>	<p>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?</p> <p>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</p>

D1.4 Final Data Management Plan

	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?	<i>(costs and potential value, who decides and how what data will be kept and for how long)</i> Output data is on-demand, so this will not apply.
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)?	Output data is on-demand, so this will not apply
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Output data is on-demand, so this will not apply
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	<i>Yes or No. (If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action DoA).</i> NA
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/departamental procedures for data management? If yes, which ones?	<i>Please list and briefly describe them.</i> NA



10 Conclusion

The Data Management Plan has been conceived as a living document, updated throughout the project to reflect the most recent developments and decisions. The document was updated at M21 and finalised at Month 36, as planned. All changes were made in full compliance with EU legislation and established best practices in research data management.

Updates were documented in the changelog table available on the Confluence page dedicated to the DMP. Notifications regarding updates were shared through regular project meetings at the Work Package (WP), WP leader, PMO, and AMB levels.

interTwin's DMP represents a comprehensive and structured approach to data management, fully aligned with Horizon Europe guidelines and the FAIR principles, ensuring data are Findable, Accessible, Interoperable, and Reusable. The plan was supported by robust technical solutions and standards, including the OpenAIRE initiative, GitHub, the EGI Document Repository, Zenodo, and Google Drive. These platforms were used to store, preserve, and share data generated or compiled throughout the project and will help ensure continued accessibility beyond the project's end.

The DMP ensured that all data were assessed according to their sensitivity and privacy level. Based on this evaluation, appropriate data sharing strategies were applied, with confidential or ethically sensitive data withheld from public release. All data management processes followed the principles of informed consent, privacy protection, and full compliance with the EU General Data Protection Regulation (GDPR).

In conclusion, this final version provides a solid foundation for the long-term preservation, sharing, and reuse of the project's research outputs, while safeguarding ethical and legal responsibilities. The DMP stands as a key component of the project's commitment to Open Science, transparency, and responsible data stewardship.