interTwin

D3.3 Interoperability protocols for data, metadata and workflow semantics across disciplines and research infrastructures report

Status: APPROVED BY THE EC

Dissemination Level: public

Funded by the European Union

Abstract

Key Words	Interoperability, data, metadata, workflows, semantics

This document provides a first evaluation of the data, metadata and workflow semantics used by the disciplines and research infrastructures in the project.

The focus is on what the use cases are developing and planning to develop in WP4, covering data formats, metadata and workflows needed for the development of Digital Twins Applications.

Document Description

D3.3 Interoperability protocols for data, metadata and workflow semantics across disciplines and research infrastructures report			
Work Package number 3			
Document type	Deliverable		
Document status	APPROVED BY THE EC	**Version**	1
Dissemination Level	Public		
Copyright Status	 This material by Parties of the interTwin Consortium is licensed under a **Creative Commons Attribution 4.0 International License**.		
Lead Partner	EURAC		
Document link	**https://documents.egi.eu/document/3932**		
DOI	**https://zenodo.org/records/14974592**		
Author(s)	• Federica Legger (INFN) • Sara Vallero (INFN) • Christian Page (CERFACS) • Sandro Fiore (UNITN) • Alexander Jacob (EURAC) • Brian Pondi (WWU) • Kalliopi Tsolaki (CERN) • Yurii Pidopryhora (MPG) • Andrea Manzi (EGI) • Raul Bardaji (EGI)		

Reviewers	• Ivan Rodero (EGI) • Isabel Campos (CSIC)
Moderated by:	• Sjomara Specht (EGI)
Approved by	• Thomas Geenen (ECMWF) on behalf of TCB

Revision History			
Version	**Date**	**Description**	**Contributors**
V0.1	27/11/2023	ToC	Alexander Jacob (EURAC) Andrea Manzi (EGI)
V0.2	14/02/2024	Included first draft of section 2	Federica Legger (INFN) Sara Vallero (INFN) Christian Page (CERFACS) Sandro Fiore (UNITN) Alexander Jacob (EURAC) Brian Pondi (WWU) Kalliopi Tsolaki (CERN) Yurii Pidopryhora (MPG) Andrea Manzi (EGI)
v0.3	21/02/2024	Added Section 1, Section 2 completed, added first version of Section 3	Andrea Manzi (EGI) Raul Bardaji (EGI)
v0.4	21/03/2024	Version ready for internal review	Alexander Jacob (EURAC)
v0.5	28/03/2024	Version reviewed	Isabel Campos (CSIC) Ivan Rodero (EGI)
v0.6	28/03/2024	Version ready for TCB review	Andrea Manzi (EGI)
v0.7	04/04/2024	Version reviewed by TCB	Thomas Geenen (ECMWF)
V1.0	05/04/2024	**Final**	

Terminology / Acronyms

Term/Acronym	Definition
DT	Digital Twin
IPCC	Intergovernmental Panel on Climate Change
WMO	World meteorological organisation

Table of Contents

Executive summary

This document provides a first evaluation of the data, metadata and workflow semantics used by the disciplines and research infrastructures in the project. The focus is on what the use cases are developing and planning to develop in WP4, so covering data formats. metadata and workflows involved in the development of Digital Twins applications.

Data interoperability may become a key factor to extract knowledge when it comes to analysing a given physical phenomenon from several perspectives, for example involving different observational instruments (satellites, networks of sensors, detectors, etc) measuring the occurrence of storms, or floods, or cosmic information in general from astrophysics experiments.

Despite the adoption of common popular data formats (such as HFD5 or NETCDF, or Python style files), effective interoperability remains a pending challenge. The lack of a standardised format for metadata, within a given domain with known exceptions like the meteorological community with standards enforced by WMO and implemented in the GRIB format), further complicates the integration and joint analysis of data from different sources. Intermediate layers to present the metadata in a more homogeneous manner, using a (reduced) community-agreed set of parameters, is an effort to be pursued at the community level. Automation of the transformation of the metadata between different implementations is also another venue to explore.

The construction of workflows in an area in which more work is needed in order to streamline the process of describing the setup of workflows. This is important to achieve a Digital Twin Engine that allows for any kind of application domain to be deployed effectively. The existing tools work on very different scopes and scales, some being specific to a set of predefined tools or functions like openEO or ECFlow, some tailored to specific types of machine learning, like KubeFlow. Some focus on calling software from the command line, while others encode the flow in an abstract set of function or process calls, expressed as nodes on a graph. In general, it is extremely case dependent. The usage of CWL as a common layer is a first approach to interoperability, however it comes often at the expense of less functionalities. This is a trade-off to be decided by the Digital Twin designer depending on the specific needs.

The way towards high interoperability goes through the development and adoption of more unified metadata standards that can be applied across datasets. These standards must be flexible enough to cover the specific needs of each domain while providing a common framework that facilitates the exchange and integration of data.

1 Introduction

This document presents the work in WP3 to collect information from the different use cases for DTs and domains and to analyse possible interoperability aspects.

The interTwin project involves mainly two macro domains DTs, physics, and environment, which internally have more subdomains as follows:

- Physics
    - High Energy Physics
    - Radio Astronomy
    - Gravitational Wave Astrophysics
- Environment
    - Climate
    - Earth Observation

The work has been done in collaboration with WP4 experts, which reported the various aspects collected for data formats, metadata, and workflow semantics.

1.1 Scope

The document is a first report of the interoperability activities in the interTwin project and could be used for further work in the project to define the interoperability paradigms between use case domains.

1.2 Document structure

This document is structured as follows:

1. **Introduction**: Sets the stage for the document.
2. **Domains descriptions**: Lists the various data, metadata, and workflows semantics areas for the project.
3. **Interoperability Analysis:** Analyses the possible interoperability aspects between domains
4. **Conclusions**: Summarises the key findings and implications of the work.

2 Domains description

2.1 Climate

The European Network for Earth System Modelling, ENES, was launched in 2001. It gathers the community modelling of the Earth's climate system with the aim to accelerate progress in this field. This community is strongly involved in the assessments of the Intergovernmental Panel on Climate Change (IPCC) and provides the predictions upon which EU mitigation and adaptation policies are based. The ENES Research Infrastructure (ENES-RI) provides the European contribution to the Earth System Grid Federation which maintains a global system of federated data centres enabling access to the largest archive of model climate data worldwide.

In interTwin, the DT applications related to ENES are the following:

- Changes in tropical storms in response to climate change
- Wildfire risk assessment in response to climate change
- Alpine droughts early warning
- Changes in extreme weather events, including rainfall, temperature, and wind, in response to climate change

2.1.1 Data formats

Table 1 - Climate data formats

Name	Description	Standards used	Example (link)
NetCDF	NetCDF refers to a collection of software libraries and self-describing, machine-independent data formats that facilitate the creation, access, and sharing of array-oriented scientific data.	CF Conventions NetCDF Classic and 64-bit Offset Format are an international standard of the Open Geospatial Consortium	http://esg1.umr-cnrm.fr/thredds/fileServer/CMIP6_CNRM/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp585/r1i1p1f2/day/tasmax/gr/v20190219/tasmax_day_CNRM-CM6-1_ssp585_r1i1p1f2_gr_20150101-21001231.nc

Zarr	ZARR stands as a file storage format tailored for N-dimensional arrays, employing chunking and compression techniques to optimise data storage.	Zarr V2 (community standard)	CMIP6 data in the cloud can be found in both Google Cloud and AWS S3 storage buckets: gs://cmip6 (part of Google Cloud Public Datasets) s3://cmip6-pds (part of the AWS Open Data Sponsorship Program)
GRIB	GRIB, short for "General Regularly distributed Information in Binary form," serves as a data format primarily utilised for encoding outputs generated by meteorological models. This table-driven, binary format is specifically crafted for efficient transmission, storage, and computer processing. Over time, GRIB has undergone modifications and enhancements to	n/a	https://climatedataguide.ucar.edu/climate-tools/common-climate-data-formats-overview

	accommodate the evolving requirements of the meteorological and oceanological communities.		

2.1.2 Metadata

Table 2 - Climate Metadata

Name	Description	Standards used	Example (link)
CF Conventions	Metadata conventions for climate model data	CF Conventions	**https://cfconventions.org**

2.1.3 Workflow Semantics

Table 3 - Climate Workflow Semantics

Name	Description	Standards used	Example (link)
Jupyter Notebook	No workflow system, just scripts or jupyter notebooks for user interaction	n/a	**https://gitlab.com/is-enes-cdi-c4i/notebooks**
Cylc	Cylc is a general purpose workflow engine that also orchestrates cycling systems very efficiently	ISO 8601 (Data elements and interchange formats - Information interchange - Representation of dates and times)	**https://cylc.github.io/**

ecFlow	ecFlow is a client/server workflow package that enables users to run a large number of programs in a controlled environment.	n/a	https://ecflow.readthedocs.io/en/latest/

2.2 Earth Observation

Earth Observation (EO) involves the systematic collection, analysis, and interpretation of data about Earth's physical, chemical, and biological systems using remote sensing technologies, such as satellites and aerial sensors. This field plays a crucial role in enhancing our understanding of Earth's processes and dynamics, contributing significantly to environmental monitoring, natural resource management, and disaster risk reduction. EO technologies offer invaluable insights into various aspects of the Earth system, such as climate patterns, land use changes, water resources management, and ecosystem health. By providing critical data and information, EO supports informed decision-making for sustainable development, climate change mitigation and adaptation strategies, and emergency response to natural and anthropogenic disasters.

In interTwin, the DT applications related to EO domain are the following:

- Early warning system for droughts and floods
- Climate change impacts of storms, fire, floods, and drought

2.2.1 Data formats

Table 4 - Earth Observation Data Formats

Name	Description	Standards used	Example (link)
NetCDF	NetCDF refers to a collection of software libraries and self-describing, machine-independent data formats that facilitate the creation, access,	CF Conventions https://cfconventions.org/	https://www.unidata.ucar.edu/software/netcdf/

	and sharing of array-oriented scientific data.		
GeoTIFF	A GeoTIFF file extension incorporates geographic metadata, providing details about the spatial location represented by each pixel in the image. When generating a GeoTIFF file, spatial information is integrated into the .tif file through embedded tags.	OGC GeoTIFF	https://www.ogc.org/standard/geotiff/
ZARR	ZARR stands as a file storage format tailored for N-dimensional arrays, employing chunking and compression techniques to optimise data storage.	Zarr V2 (community standard)	https://portal.ogc.org/files/100727
Cloud Optimized GeoTIFF (COG)	A Cloud Optimized GeoTIFF (COG) is a standard GeoTIFF file designed for hosting on an HTTP file server. Its internal structure is optimised to enhance cloud-based workflows by capitalising on clients' capability to make precise HTTP GET range requests, allowing them to request only the specific portions of the file they require.	https://docs.ogc.org/is/21-026/21-026.html	https://docs.ogc.org/is/21-026/21-026.html
HDF5	Hierarchical Data Format (HDF) is a versatile file format designed to facilitate the storage and manipulation of scientific data across diverse operating systems and machines. A library of callable routines and a suite of utility	n/a	https://portal.hdfgroup.org/display/HDF5/Design+Specifications

	programs and tools for creating and using HDF files were introduced. HDF accommodates various data types, including scientific data arrays, tables, and text annotations, as well as multiple types of raster images and their associated colour palettes.		
JPEG2000	JPEG200 is an image compression standard and coding system.	ISO/IEC 15444-1	https://jpeg.org/jpeg2000/

2.2.2 Metadata

Table 5 - Earth Observation Metadata

Name	Description	Standards used	Example (link)
STAC	The SpatioTemporal Asset Catalog (STAC) specification offers a standardised framework for describing and cataloguing spatiotemporal assets.	in approval state at OGC as community standard	https://stacspec.org/en/about/stac-spec/
INSPIRE	INSPIRE, an initiative by the European Commission, establishes a European spatial data infrastructure for a unified environmental policy. Enforced by Directive 2007/2/EC since May 15, 2007, it requires Member States to progressively provide pre-existing digital geospatial data on 34 topics initially in a compliant format and later through	n/a	https://inspire.ec.europa.eu/

	interoperable network services. No new recording of analogous geodata is mandated.		
ISO	ISO 19115-1:2014 establishes the schema essential for detailing geographic information and services through metadata. This standard furnishes details regarding the identification, extent, quality, spatial and temporal aspects, content, spatial reference, portrayal, distribution, and additional properties of digital geographic data and services.	https://www.iso.org/standard/53798.html	n/a

2.2.3 Workflow Semantics

Table 6 - Earth Observation Workflow Semantics

Name	Description	Standards used	Example (link)
openEO API process graphs	The openEO API employs the idea of a JSON Process Graph, functioning as a symbolic expression tree stored in JSON format as a directed acyclic graph. This graph delineates processing tasks and operations, specifying input data, operations, and their associated parameters. This flexibility allows users to construct processing workflows of diverse complexities. Once these graphs are transmitted to	in submission state at OGC as a community standard	https://openeo.org/documentation/1.0/developers/api/reference.html https://github.com/Open-EO/PSC/blob/main/documents/ogc-submission.adoc

	an openEO-compliant backend service, the service executes the specified computational steps and delivers the desired output.		
Application Package with CWL	The Core Standard of OGC API—Processes (Part 1) facilitates the encapsulation of computational tasks into executable processes, which servers can present through a Web API for invocation by client applications. This standard defines a processing interface for communication via a RESTful protocol, employing JavaScript Object Notation (JSON) encodings.	OGC API processes	https://www.ogc.org/standard/ogcapi-processes/

2.3 High Energy Physics

High Energy Physics experiments have long been using simulation tools for different tasks, ranging from the initial phases of detector design to the final analysis step in order to compare the experimental data to theoretical models. LHC allowed the discovery of the Higgs boson, which was awarded the Nobel prize in Physics in 2013.

In order to understand what the new physics might be beyond what is currently captured in the Standard Model (SM), we need to be able to make highly accurate theoretical predictions and confront them with the outcomes of the ongoing experimental efforts. Lattice QCD provides, inter alia, a mathematically well-defined approach to simulate the SM at high temperatures and densities. This can be used to develop a theoretical understanding of matter in the plasma phase and eventually explain results of recent heavy ion collision experiments at BNL and CERN, as well as prospective results from the FAIR facility in Darmstadt. Therefore, Lattice QCD simulations are always a reference when it comes to comparing the experimental results with the theoretical predictions. Lattice simulations are also a part of the game in simulations related to Astrophysics observations, such as dark matter (e.g. SKA experiment), or Physics of neutron stars, which is key to understanding the phenomena behind the onset of Gravitational waves. In such cases a combination of effective models and Lattice QCD have already been used by several collaborations to do pioneering explorations. In that sense Lattice QCD simulations are of interdisciplinary nature.

In interTwin, the DT applications related to High Energy Physics domain are the following:

- Lattice QCD Simulations
- Detector Simulations

2.3.1 Data formats

Table 7 – High Energy Physics Data formats

Name	Description	Standards used	Example (link)
ROOT	Data analysis framework used by HEP. ROOT files are saved during GEANT4 simulations.	n/a	https://root.cern/ https://root.cern/manual/trees/
HDF5	Hierarchical Data Format (HDF) is a versatile file format designed to facilitate the storage and manipulation of scientific data across diverse operating systems and machines. A library of callable routines and a suite of utility programs and tools for creating and using HDF files were introduced. HDF accommodates various data types, including scientific data arrays, tables, and text annotations, as well as multiple types of raster images and their associated colour palettes.	n/a	https://www.hdfgroup.org/solutions/hdf5/
.pkl	Format used in Python for serialising and deserializing Python object structures.	n/a	https://docs.python.org/3/library/pickle.html
.pth	Format used to store PyTorch models, tensors, or state dictionaries.	n/a	https://pytorch.org/tutorials/beginner/saving_loading_models.html

.tfrecords	Format used by TensorFlow for storing a sequence of binary records. Useful for handling large datasets and commonly used in TensorFlow for input data pipeline optimizations.	n/a	**https://www.tensorflow.org/tutorials/load_data/tfrecord**

2.3.2 Metadata

Table 8 - High Energy Physics Metadata

Name	Description	Standards used	Example (link)
ML Metadata (MLM)	In the context of ML processes and pipelines, metadata is crucial for understanding, managing, and tracking the ML lifecycle. Key types: Data/Model/Training metadata, Performance Metrics, Experiment/Operational metadata, Provenance and Lineage, Annotations and Labels, User and Usage Metadata	n/a	There are different libraries for tracking ML metadata, depending on the ML framework. E.g.: **https://github.com/google/ml-metadata** **https://www.tensorflow.org/tfx/guide/mlmd**
QCDml (ILDG)	The International Lattice Data Grid (ILDG) was started in 2002 with the aim of making the basic data sets from **Lattice QCD** simulations available to the international scientific community.	They consist of a set of metadata and binary files. The metadata is available as XML documents which conform to the XML schema developed by the	**https://hpc.desy.de/ildg/specifications/** **https://www2.ccs.tsukuba.ac.jp/ILDG/** **Example ensemble**

	The data sets stored are ensembles of gauge field configurations.	ILDG metadata working group.	**metadata**

2.3.3 Workflow Semantics

Table 9 - High Energy Physics Workflow Semantics

Name	Description	Standards used	Example (link)
Kubeflow	Deployment of machine learning (ML) workflows on Kubernetes.	n/a	**https://www.kubeflow.org/**

2.4 Radio Astronomy

The MeerKAT radio telescope[1] is a precursor to the world's largest radio telescope, the Square Kilometre Array (SKA), which is being constructed in South Africa and Australia. Radio telescopes can be used to observe transient signals, such as Fast Radio Bursts, and take images of the universe, such as supermassive black holes in the centre of galaxies. Noise simulation of a radio telescope (MeerKAT) should be performed during the initial signal processing step ("beamforming"), where data streams are recorded as multi-frequency time-series. Astrophysical signals are identified by looking for periodicities using several algorithms. Providing DTs to simulate the noise background of radio telescopes will support the identification of rare astrophysical signals in (near-)real time. The result will contribute to a realisation of "dynamic filtering" (i.e. steering the control system of telescopes/sensors in real-time).

In interTwin, the DT application related to Radio Astronomy domain is the following:

- Noise simulation for radio astronomy

[1] https://www.sarao.ac.za/science/meerkat/about-meerkat/

2.4.1 Data formats

Table 10 - Radio Astronomy Data formats

Name	Description	Standards used	Example (link)
filterbank file (.fil)	A binary data storage format to store time series of power spectra, common in radio astronomical studies of pulsars. Raw data recorded by a radio telescope is typically converted to this format for storage and analysis.	Filterbank format is a standard. Description of work with this type of files can be found in the following documents/links: **https://sigproc.sourceforge.net/sigproc.pdf** **https://notebook.community/UCBerkeleySETI/breakthrough/GBT/filterbank_tutorial/Filterbank%20Tutorial%20(public)** **https://github.com/UCBerkeleySETI/breakthrough/tree/master/GBT/filterbank_tutorial**	**https://github.com/UCBerkeleySETI/breakthrough/blob/master/GBT/filterbank_tutorial/blc04_guppi_57563_69862_HIP35136_0011.gpuspec.0002.fil**
FITS file (.fits)	The most common binary data format in radio astronomy and astronomy in general.	Flexible Image Transport System (FITS) **https://www.aanda.org/articles/aa/pdf/2010/16/aa15362-10.pdf**	**https://fits.gsfc.nasa.gov/fits_samples.html**

MeasurementSet (MS)	A format for storing radio astronomical data, used by CASA software package (Common Astronomy Software Applications), currently the main software package to handle many types of radio astronomical observations. Each MeasurementSet is a directory tree, containing many "tables", which can be either in binary or ASCII format. See the following for details: **https://casadocs.readthedocs.io/en/stable/notebooks/casa-fundamentals.html**	MeasurementSet **https://casa.nrao.edu/Memos/229.html** **https://casacore.github.io/casacore-notes/264.html**	**https://casaguides.nrao.edu/index.php/Measurement_Set_Contents**

2.4.2 Metadata

Table 11 - Radio Astronomy Metadata

Name	Description	Standards used	Example (link)
filterbank file (.fil)	this data standard includes an ASCII header for storing metadata	See **Table 10**	See **Table 10**
FITS file (.fits)	this data standard includes an ASCII header for storing metadata	See **Table 10**	See **Table 10**

MeasurementSet (MS)	In addition to binary "tables" with data, this data structure also includes ASCII "tables" with metadata.	See **Table 10**	See **Table 10**

2.4.3 Workflow Semantics

Table 12 - Radio Astronomy Workflow Semantics

Name	Description	Standards used	Example (link)
Jupyter notebooks	No workflow system, just scripts or jupyter notebooks for user interaction	n/a	n/a

2.5 GW Astrophysics

The discovery of gravitational waves was one of the main scientific results in recent years and was awarded the Nobel Prize in 2017, and many other events have then been observed by Advanced Virgo and Advanced LIGO. The observation of a binary neutron star merger in 2017, that was also observed by several ground- and space- based EM observatories, marked a new era in the study of the cosmos and paved the way to multi-messenger astronomy. Besides getting ready for their next observation runs, the gravitational-wave community is designing a next-generation observatory, the Einstein Telescope, which was recently included in the EU ESFRI roadmap.

In interTwin, the DT application related to Radio Astronomy domain is the following:

- VIRGO Noise Detector DT

2.5.1 Data formats

Table 13 -GW Astrophysics Data formats

Name	Description	Standards used	Example (link)
HDF5	Hierarchical Data Format (HDF) is a versatile file format designed to facilitate the storage and manipulation of scientific data across diverse operating systems and machines. A library of callable routines and a suite of utility programs and tools for creating and using HDF files were introduced. HDF accommodates various data types, including scientific data arrays, tables, and text annotations, as well as multiple types of raster images and their associated colour palettes.	n/a	https://docs.hdfgroup.org/hdf5/develop/_t_b_l.html
GWF	Gravitational Wave Frame file custom format containing both data and metadata	n/a	https://lappweb.in2p3.fr/virgo/FrameL/VIR-067A-08.pdf
PNG	Portable Network Graphics is a raster-graphics file format that supports lossless data compression.	https://w3.org/TR/2003/REC-PNG-20031110/	https://www.w3.org/TR/2003/REC-PNG-20031110/

Pickle	Serialised Pytorch models	n/a	https://formats.kaitai.io/python_pickle/

2.5.2 Metadata

Table 14 - GW Astrophysics Metadata

Name	Description	Standards used	Example (link)
ML Metadata (MLMD)	Metadata associated with ML pipelines	n/a	https://github.com/google/ml-metadata
GWF	Gravitational Wave Frame file custom format containing both data and metadata	n/a	https://lappweb.in2p3.fr/virgo/FrameL/VIR-067A-08.pdf

2.5.3 Workflow Semantics

Table 15 - GW Astrophysics Workflow Semantics

Name	Description	Standards used	Example (link)
KubeFlow Pipelines (KFP) SDK	A set of Python packages to specify and run ML workflows with KubeFlow.	n/a	https://v0-6.kubeflow.org/docs/pipelines/sdk/sdk-overview/

3 Interoperability analysis

Domain knowledge relies often on the possibility to interoperate data and their related metadata. Chapter 2 provides a detailed overview of the data and metadata formats used in domains such as climatology, Earth observation, high energy physics, radio astronomy, and gravitational wave astrophysics. Although common file formats such as HDF5, NetCDF, and specific serialisation formats like .pkl for Python are identified, the actual interoperability among these use cases remains limited, since these formats represent highly configurable data structures and hence different configurations result in a completely different organisation of data within each file.

3.1 Common File Formats

HDF5 and NetCDF formats are widely recognized and utilised across various domains for storing and managing large sets of scientific data. These formats offer a hierarchical structure and the capability to store metadata alongside the data, which facilitates access to and organisation of information.

In addition to HDF5 and NetCDF, .pkl (Pickle) files are commonly used in the Python programming realm for serialising and deserializing objects. This functionality allows for the saving of Python objects, such as data structures or machine learning models, in files that can be transported and read later. However, this type of format presents significant limitations in terms of interoperability. .pkl files are entirely dependent on the Python development or execution environment that created the file. This means a .pkl file generated in an environment with certain library versions may not be compatible with another environment where these library versions differ. This dependency on the environment and library versions introduces an additional challenge for interoperability, as it does not guarantee the consistency or usability of the data across different research or development settings.

Despite the adoption of these formats, effective interoperability between disciplines remains a pending challenge. The lack of a standardised format for metadata further complicates the integration and joint analysis of data from different sources. Moreover, the specific objective of each use case introduces differences in how data should be interpreted and handled, limiting the ability to work in an integrated and cohesive manner across different fields of study.

3.2 Common Workflow Descriptions

In the different domains and also areas of application different tools are used for describing a scientific workflow. Many specific solutions exist ranging from process graphs as expressed in openEO, over ECFlow for climate data processing to specific tools for machine learning like kubeFlow. Additionally, the common workflow language CWL allows expressing the chaining of command line tools at a very generic level. OGC

Application packages seek to give a more standardised structure to the organisation of each containerized application resource and its according CWL descriptor document, like CF convention does for netCDF files.

3.3 Interoperability Challenges

There are two main challenges of interoperability, the first being the ability to link the plethora of software tools through a unified workflow and the second regards the flow of data starting from a multitude of differently organised files.

The main obstacle of the latter to interoperability is the lack of a standardised format for metadata. Although file formats may be common, the content and structure of metadata vary significantly among different use cases, as already mentioned initially in the context of netCDF files. This variability complicates the integration and joint analysis of data from diverse sources.

There is in the different domains lots of experience on how to build workflows, but most of the concepts are very specific and not generic enough, to serve all domains and use cases. Workflows also need to be linked effectively to data in order to trigger processing when new data arrives in order to allow for continued updates in digital twins, which is paramount in e.g. early warning systems.

Furthermore, the specific objective of each use case introduces differences in how data should be interpreted and managed. For instance, the data used in High Energy Physics may require a completely different analytical approach than data used in climatology, even if both are stored in an HDF5 format. For this purpose, the project foresees and implements Thematic Modules which allow for different types of analysis based on the requirements of each domain and use case specifically.

3.4 Towards Higher Interoperability

To overcome these challenges and move towards higher interoperability, it is necessary to develop and adopt more unified metadata standards that can be applied across different disciplines. These standards must be flexible enough to cover the specific needs of each domain while providing a common framework that facilitates the exchange and integration of data.

Furthermore, it is necessary to encourage collaboration among research communities to share best practices and tools that facilitate interoperability. Developing platforms and services that can efficiently handle different data and metadata formats, and interpreting their semantics coherently, will be a significant step forward in this effort.

Additionally, in cases where data input formats cannot be harmonised due to reliance on external sources, an effort should be made to create an intermediate layer with some harmonisation so that the input data can be used for multiple case studies without making too many changes in the data acquisition process. The output format of the case studies should be agreed upon and as similar as possible for greater interoperability. This

approach ensures that even when direct standardisation of data formats is not feasible, the data can still be used effectively across different studies, enhancing the overall interoperability of the research ecosystem.

Instead of relying on file internal metadata this can be worked around by building metadata external to the file itself in forms of well-organised catalogues following a common specification. An example to this would be STAC.

4 Conclusions

This document describes a first analysis of the existing landscape of data and its metadata descriptors in the context of the interTwin consortium members and the related scientific domains and how is organised in form of scientific workflows including one or multiple tools for data processing, analysis and possibly visualisation of the final results.

It is clear from this initial comparison that interoperability is a key challenge and that especially on the workflow end more work is needed in order to allow for unification of the process of describing the setup of such a workflow, in order to be able to achieve a Digital Twin Engine that allows for any kind of application domain to be deployed effectively. The existing tools work on very different scopes and scales, some being specific to a set of predefined tools or functions like openEO or ECFlow, some tailored to specific types of machine learning, like KubeFlow. Some focus on calling software from the command line while others encode the flow in an abstract set of function or process calls, expressed as nodes on a graph. For the final iteration of the implementation of the core components for workflow management a solution must be found to be able to link the different thematic modules and core modules in a unified way.

The diversity on the data level can more easily be solved, since more mature concepts of abstract meta data models exist, that allow for exact description and organisation of files in form of catalogues. More work here is still needed though on the standardisation and exchange of refined data in the form of machine learning models and how to store and refer to those trained models and make them reusable in the context of inference, retraining, domain adaptation etc.

As there is no further deliverable to report updates on the WP3 interoperability activity, a final report will be included in the final project technical report.