Travelling Ionospheric Disturbances Forecasting System

T-FORS

Initial T-FORS standards, quality control and best practices

Version IR

The T-FORS project has received funding from the European Union's Horizon Europe Research and Innovation Actions programme under Grant Agreement No 101081835

INDEX

Index of Tables

Document Information

Milestone Number: Deliverable Title:	7 Initial T-FORS standards, quality control and best practices
Date of Delivery:	11th December 2023
Author(s):	I. Galkin (BGD), A. Belehaki (NOA), C. Cesaroni (INGV), T. Verhulst (RMI)
Work Package No.:	WP1
Work Package title:	Strategy and capabilities
Work Package leader:	BGD
Dissemination level:	Public
Nature:	Report

Abstract

The Milestone 7 document presents an initial report on existing standards and best practices for scientific data production, quality control, and archival management. It will become the foundation for the T-FORS data management plan deliverable D1.3.

Document history

Version	Date	Edited by	Reason for modification / Remarks
0.1	2023.12.04	Ivan Galkin	First draft
IR	2023.12.11	Ivan Galkin	Initial release

Disclaimer

Executive Summary

T-FORS project aims at providing new models able to interpret a broad range of observations of the solar corona, the interplanetary medium, the magnetosphere, the ionosphere and the atmosphere, and to issue forecasts and warnings for TIDs several hours ahead [AD-1]. T-FORS expects to develop prototype services with a comprehensive architectural concept and data quality management that would be used for possible future adjustments to develop a real-time operational service.

This document is the initial review of data catalogues and products required for the T-FORS development, focusing on their compliance with modern data management standards. Best practices for quality control of data products that result from TID detection methodologies are compiled, and specific techniques to assess their quality (i.e., timely availability and scientific reliability of the results) are proposed, including a monitoring system for the quality of the raw observational data. The areas for potential improvements in the TID methodologies are also identified. Based on this, an initial data management plan and the concepts for synchronized operations among the T-FORS observing facilities will be released as the D1.3 Deliverable document.

1 Purpose and Scope of the Document

This document presents the initial review of the best practices for managing T-FORS data production, quality control, and archival. It will become a prerequisite for developing the T-FORS data management plan. The following information is included in the document:

- What data will be collected, processed, and generated;
- Which standards and established best practices will be applied to data management;
- Which quality control measures will be implemented;
- How data will be shared;
- How data will be preserved, including after the end of the project.

The document is divided into six sections:

Section 1 (the current section) describes the purpose of this document and its organisation.

Section 2 lists the applicable and reference documents and contains the list of acronyms used in this document.

Section 3 describes the data catalogue, presents applicable data management standards and best practices, and reviews compliance enforcement measures.

Section 4 presents concepts of data quality control.

Section 5 discusses data sharing and archival.

Section 6 proposes areas for T-FORS enhancements resulting from best practice analysis.

2 Applicable/Reference Documents and Acronym List

2.1 Applicable Documents

The following table contains the list of applicable documents.

Table 1. List of applicable documents

AD	Document title
[AD-1]	Grant Agreement number: 101081835 — T-FORS — HORIZON-CL4-2022-SPACE-01
[AD-2]	T-FORS Deliverable 1.2: The state-of-the-art review report on standards and capabilities in TID detection procedures
[AD-3]	T-FORS Deliverable 1.1: Data Management Plan

2.2 Reference Documents

The following table contains the list of references used in this document.

Table 2. List of reference documents

RD	Document title
[RD-1]	ESA SSA Team, "Space Situational Awareness – Space Weather Customer Requirements Document", Rev.5a, SSA-SWE-RS-CRD-1001, 2011-07-28
[RD-2]	European Commission (2016), "Guidelines on FAIR Data Management in Horizon 2020", Version 3.0, 26 July 2016.
[RD-3]	"Warning and Mitigation Technologies for Travelling Ionospheric Disturbances Effects (TechTIDE)", Horizon 2020 project landing page, available at https://cordis.europa.eu/project/id/776011
[RD-4]	"Plasmasphere Ionosphere Thermosphere Integrated Research Environment and Access services: a Network of Research Facilities" (PITHIA-NRF), Horizon 2020 project landing page, https://cordis.europa.eu/project/id/101007599
[RD-5]	ESFRI (2009), "e-IRG Report on Data Management", e-Infrastructure Reflection Group, November 2009
[RD-6]	"Data Quality Flag" (2023). PITHIA-NRF Space Physics Ontology Description, Specifications available at https://esc.pithia.eu/ontology/categories/dataQualityFlag/
[RD-7]	"Metadata Quality Flag" (2023). PITHIA-NRF Space Physics Ontology Description, Specifications available at https://esc.pithia.eu/ontology/categories/metadataQualityFlag/

2.3 Acronyms

Table 3 contains the list of all acronyms used in this document.

Table 3. List of acronyms

Acronym	Definition
BGD	Borealis Global Designs EOOD
CME	Coronal Mass Ejection
DL	Data Level
DQ	Data Quality flag
ESA	European Space Agency
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HF	High Frequency
HF-INT	High Frequency-Interferometry
IEF	Intrerplanetary Electric Field
IMF	Interplanetary Magnetic Field
INGV	Istituto Nazionale di Geofisica e Vulcanologia
IPR	Intellectual Property Rights
ISO	International Standards Organisation
LSTID	Large Scale Traveling Ionospheric Disturbances
MIMO	Multiple-Input Multiple-Output
MQ	Metadata Quality flag
MSTID	Medium Scale Traveling Ionospheric Disturbances
N-RTK	Network for Real-Time Kinematic positioning service
NOA	Ethniko Asteroskopeio Athinon (National Observatory of Athens)
NOAA	National Oceanic and Atmospheric Administration
ODQ	Operational Data Quality
PITHIA	Plasmasphere Ionosphere Thermosphere Integrated Research Environment and Access services

Acronym	Definition
PITHIA-NRF	PITHIA Network of Research Facilities
RDA	Research Data Alliance
RMI	Institut Royal Meteorologique de Belgique
ROTI	Rate of Total Electron Content Index
SSA	Space Situation Awareness
T-FORS	Travelling ionospheric disturbances Forecasting System
TBD	To Be Determined
TDR	Trusted Digital Repository
TEC	Total Electron Content
TechTIDE	Warning and Mitigation Technologies for Travelling Ionospheric Disturbances Effects
TID	Travelling Ionospheric Disturbances
VTEC	Vertical TEC
WDC	World Data Centre
WP	Work-package

3 T-FORS Compliance to Data Management Standards

3.1 *General Considerations*

Although T-FORS ultimately targets the development of an operational, real-time system to issue forecasts and warnings for TID activity several hours ahead [AD-1], it will remain a *functioning prototype with recommendation for transition to operations* throughout the project's period of performance. Yet, despite the prototype status of the T-FORS systems, the data management plan and its underlying concepts of the compliance to standards, data quality control, and overall best practices of system development and management will be strictly enforced. This section outlines existing standards of data management for the infrastructures in the "glass-room" unattended operation regime.

3.2 *Background*

Traveling ionospheric disturbances (TID) are natural or man-made phenomena of quasi-periodic plasma density undulations [AD-2, Section 2.1]. The adverse effects of TIDs on operational systems that employ radio signals traversing the ionosphere are well established, though not sufficiently understood to predict or mitigate in real-time. T-FORS will approach the task differently, depending on the TID category [AD-2, Section 2.2.3]:

- ✓ Large-scale LSTID (rapid-moving, large-amplitude perturbations of the wavelength above 600 km) and
- ✓ Medium-scale MSTID (slower-moving, lower-amplitude perturbations of 50-600 km scale).

Although both types affect the radio systems in a similar way, T-FORS will employ significantly different systems of data acquisition, processing, and analysis for their forecast.

- ✓ LSTIDs are easier to detect and find contra-measures to address (e.g., by establishing a network of reference location nodes to compte and distribute augmentations to the systems). Generation of TIDs of such large spatial scale is commonly attributed to weather systems of the planetary scale that therefore admits forecast by monitoring the causality chain of event originating on the Sun.
- ✓ MSTIDs are harder to detect and not feasible to compensate for by the reference networks like N-RTK (the [augmentation] Network for Real-Time Kinematic positioning service) that would have to be built at matching spatial resolution and therefore become prohibitively expensive. Yet, MSTID impact on the operational systems can be as severe as LSTID's, warranting their short-term forecast. Though in high demand, such forecast is hard to accomplish as a wide variety of physical mechanisms that are responsible for MSTID generation (earthquakes, industrial accidents, rocket launches, volcano eruptions) defy deterministic treatment. Thus, statistical approach to MSTID prediction will be undertaken to identify their climatological behaviour as a function of local time, season, solar activity, and location.

3.3 Data Products

T-FORS will be producing four data product types:

- ✓ **LSTID activity forecasts**, released continuously;
- ✓ **MSTID activity predictions**, released continuously, and
- ✓ **TID activity warnings and alerts,** issued if prospected plasma disturbance reaches predefined levels of activity.
    - o Including Pan-European "TID in progress" alerts
- ✓ **Retrospective TID activity reports**, provided
    - o on demand for review and reanalysis, and
    - o in real-time but with delay for improved specification.

The nature of the forecast and prediction algorithms suggests different regimes of associated data analysis:

- ✓ **Training of the T-FORS models** using large historical datasets of helio- and geospace activity:

- sensor data from contributing observatories,
- activity indicators provided by external weather agencies;

✓ **Real-time T-FORS forecast and prediction messaging** using live data feeds of space weather context, and

✓ **Real-time TID-In-Progress alerts** as detected by the contributing online T-FORS sensors.

T-FORS will use a wide range of sensor data and activity indicators as its input to describe and monitor ongoing processes in the Sun-Earth system that can result in triggering TIDs and influence their propagation. Because of the *data lineage* requirement that prescribes detailed description of the data resources from their generation to processing and archival, as stated by the European Space Agency (ESA) Space Situation Awareness (SSA) Programme Requirements [RD-1], a catalogue of the T-FORS input data resources and output data products is hereby compiled and presented in the following subsections.

3.3.1 Input data: Historical helio- and geospace activity datasets for training

Model training for T-FORS LSTID forecast services, statistical studies for T-FORS MSTID prediction services, and TID-on-Progress alerts rely on availability of the retrospective data on helio- and geospace activity and a variety of sensor measurement used for TID detection. The input indicator data for T-FORS model training and statistical analyses will include, but not limited to, the following resources (Table 4):

Table 4. Summary of input sensor/indicator data resources for T-FORS model development.

Green shading: Features applicable to LSTID activity forecast; Yellow shading: Features applicable to MSTID activity prediction. White shading: Features applicable to TID-In-Progress warnings and alerts.

Name	Description	Availability	Parameters	Cadence
L S T I D F E A T U R E S				
Solar images	Halo CME catalogue	https://www.sidc.be/cactus	Principle angle, angular width and velocity estimation for each CME.	Event related
SOHO images	Solar corona images produced by SOHO	https://soho.nascom.nasa.gov	TBD	TBD
F10.7	Solar radio flux at 10.7 cm wavelength	https://www.spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-en.php	F10.7	Daily
Solar Wind at L1	Solar Wind and Interplanetary Magnetic and Electric Fields at L1 by DISCOVR and ACE	https://www.ngdc.noaa.gov/dscovr https://services.swpc.noaa.gov/products/ https://cdaweb.gsfc.nasa.gov	IMF Bx, By, Bz IMF Btot IEF Ex, Ey, Ez n_{SW}, p_{SW}, T_{SW} Newell	1 min

Name	Description	Availability	Parameters	Cadence
Electrojet activity	Disturbance Storm Time index from equatorial magnetometers	https://wdc.kugi.kyoto-u.ac.jp/	Dst	1 hr
	Electrojet indicators over European sector calculated from the IMAGE network	https://space.fmi.fi/image/www/index.php?page=il_index	IU,IL,IE	TBD
MMs	B-field components: magnetometers in the high-latitude sector	IMAGE https://space.fmi.fi/image SUPERMAG https://supermag.jhuapl.edu/	H, D, Z	1 min
ROTI	Rate of Total Electron Content Index: GNSS receivers in the high-latitude European sector	EUREF https://epncb.oma.be/ IGS https://network.igs.org/	ROTI	5 min
LSTID LABELS				
Catalogue of LSTID detections	Manually produced LSTID catalogue based on European ionosonde network (Ebro Observatory)	TBD	TBD	Event related
SPCont	Spectral content index (single ionosonde)	http://techtide.space.noa.gr	SPCont	15 min
HF EU	Automatic European Index of LSTID activity	http://techtide.space.noa.gr	HF EU	TBD
MSTID FEATURES				
VTEC	Global maps of VTEC from GNSS	https://stdb2.isee.nagoya-u.ac.jp/GPS/GPS-TEC/	VTEC 2D maps	30 sec
Ionogram images	Ionogram images showing signatures of MSTID and spread F	https://giro.uml.edu/didbase	Category Tag	30 min
Doppler Radar	MIMO Doppler radar spectrograms showing MSTID signatures		Doppler frequency timelines	1 s
TID MEASUREMENTS				
HF-INT	LSTID detections by sensing periodic variations of the ionogram-derived chars in an ionosonde network	TBD	2D velocity vector, amplitude, period	5 min

Name	Description	Availability	Parameters	Cadence
HF-TID	LSTID detections using synchronized Digisonde sounding	TechTIDE	TID velocity, amplitude, propagation direction	2.5, 5 min
Doppler Radar	MIMO Doppler radar spectrograms showing MSTID signatures	TBD	MSTID detections	1 s

3.3.2 Input data: Real-time helio- and geospace activity data feeds

The real-time indicator data for T-FORS analyses will include, but not limited to, the following resources (Table 5):

Table 5. Summary of real-time helio- and geospace activity for use in T-FORS forecast and prediction.

Name	Description	Availability	Parameters	Latency	Cadence
F10.7	Solar flux at 10.7 cm wavelength	https://www.spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-en.php		TBD	
Solar Wind at L1	Solar Wind and Interplanetary Magnetic and Electric Fields at L1 by DISCOVR and ACE	https://www.ngdc.noaa.gov/dscovr https://services.swpc.noaa.gov/products/ https://cdaweb.gsfc.nasa.gov	IMF Bx, By, Bz IMF Btot IEF Ex, Ey, Ez n_{sw}, p_{sw}, T_{sw} Newell	TBD	1 min
Electrojet activity	Disturbance Storm Time index from equatorial magnetometers	https://wdc.kugi.kyoto-u.ac.jp/	Dst	TBD	1 hr
	Electrojet indicators over European sector calculated from the IMAGE network	https://space.fmi.fi/image/www/index.php?page=il_index	IU,IL,IE	TBD	TBD
MMs	B-field components: magnetometers in the high-latitude sector	IMAGE https://space.fmi.fi/image SUPERMAG https://supermag.jhuapl.edu/	H, D, Z	TBD	1 min
ROTI	Rate of Total Electron Content Index: GNSS receivers in the high-latitude European sector	EUREF https://epncb.oma.be/ IGS https://network.igs.org/	ROTI	TBD	5 min

3.3.3 Input data: External forecast of space weather

T-FORS will leverage existing space weather services that forecast activity in the Heliophysics domain (Table 6).

Table 6. Summary of external forecast services for use in T-FORS computations.

Name	Description	Availability	Parameters	Latency	Cadence
CME tracking	SIDC OMA (CACTUS) service	TBD	Principle angle, angular width and velocity estimation for each CME.	TBD	TBD
L1 B-field forecast	EUHFORIA	TBD	IMF Bx, By, Bz IMF Btot	TBD	TBD

3.3.4 Output data: TID warning, alert, and report

Table 7. Summary of T-FORS output data products.

Name	Description	Parameters	Latency	Cadence
LSTID Forecast	Forecast of LSTID activity based on helio- and geospace activity indicators	Onset time, amplitude, period, propagation vector, duration	1 hour	15 min
MSTID prediction	Prediction of LSTID activity for given local time (empirial model)	TBD	TBD	TBD
TID-in-progress alert	Alert of TID in progress from the participating sensor detections	Onset time, amplitude, period, propagation vector, duration	1 hour	5 min

3.4 General Data Management Principles: FAIR

Compliance to the Findable-Accessible-Interoperable-Reusable (FAIR) data organisation is expected from all modern data infrastructures [RD-2]. The T-FORS project team has experience implementing the FAIR principles in two previous projects, TechTIDE [RD-3] and PITHIA-NRF [RD-4]. All T-FORS data collections will be registered in PITHIA-NRF, which will ensure their FAIRness. For brevity of this report, further description of FAIR and project compliance plan is not included here; a detailed analysis of the standard and its applicability to the relevant data sets in space weather applications can be found in our previous reports to EC Horizon 2020 [AD-3].

3.5 Data Lineage

3.5.1 Reproducibility of T-FORS Computations

Another requirement to the T-FORS data organization comes from the best practice argument of ensuring *reproducibility* of all TID warning/alert/report data product computations in subsequent reanalyses. The reproducibility is defined as the capability of different science teams arriving at the same results using the preserved artifacts (data, software codes) of the original computation. The requirement stems from the science community striving to overcome a recognized problem of failures to yield the same published research results when analysis is performed by an independent party. Dubbed “reproducibility crisis”, the problem was first noted in academic publications on pre-clinical studies of prospective medications whose outcomes could not be used by clinicians and drug developers because more than two-thirds of experiments were impossible to repeat independently.

The reproducibility is an important part of the wider-scope data *lineage* and *provenance* management practices [RD-1, RD-5] that both prescribe retaining full information how data originate and follow their life cycle. T-FORS will:

- **Preserve T-FORS computation context**: within T-FORS data infrastructure, the helio- and geospace activity context that determines the outcome of the TID forecast, prediction, warning, and alert computations must be preserved in the output data products to ensure data lineage[1].
- **Preserve T-FORS computation configuration:** all output data products will include descriptions of the software versions and algorithm parameters and configurations that were used to produce output records.

3.5.2 Retaining Life Cycle Details

Even if the context and configuration metadata are preserved for each T-FORS computation, it is still possible to lack consistency of results if the original record gets replicated and subjected to additional manipulation as a part of its life cycle. Therefore, data provenance shall be enforced to ensure description of the data origin (where and when they were generated originally) and data source (where and when they were retrieved for use). Data lineage shall be enforced to capture all details of data preparation, quality control (cleansing, validating), transformation into a suitable resolution or unit of measurement, and enhancements with additional attributes.

4 Data Quality Control

The T-FORS performance ultimately depends on the quality of the input parameters (features) provided by external and internal data resources. These resources will be generating their data in real-time when T-FORS is running its algorithms. Therefore, inputs are likely to include data noise that may cause erroneous outputs. Data quality management is a critical requirement in this scenario, suggesting several steps as outlined below.

4.1 Data Quality Categorizations

it is important to categorize quality of the T-FORS input features used for training and forecast. There are several aspects of such categorization:

4.1.1 Scientific Data Quality

Scientific data quality refers to the accuracy/precision, consistency, and relevance of the information. We will accept PITHIA ontology definition of Data Quality (DQ) flags [RD-6] to

[1] Preserving the raw measurement data used to produce various activity context metrics for T-FORS computations is out of T-FORS scope; it will remain the responsibility of the input data providers.

reflect the scientific qualities of information in terms of taken measures to clean and validate the data, as well as characterize the residual data noise. In this definition, DQ should not be confused with another common descriptor of data categories called Data Level (DL). DL does not describe the data quality but rather characterizes the amount of data processing applied to the measurements to obtain higher level data products, so it is relevant to the data lineage requirement described in Section 3.5.

The following definitions of four DQ categories are proposed for T-FORS:

Data Quality 0 (RAW): When no consideration is made to the evaluation of the data product quality, DQ flag is zero.

Data Quality 1 (CLEAN): The CLEAN flag is assigned to report the data conditioning capability of the computation process that applies *automatic, unattended* measures to exclude data noise. Example data conditioning algorithms are:

- Detection/removal of data outliers,
- Filtering to exclude data jitter,
- Content sanity checks against physical criteria (e.g., exclusion of negative density or altitude values, or other comparisons against established threshold values).

Data Quality 2 (EVALUATED): The EVALUATED flag is assigned to those data products that provide *confidence* and/or *uncertainty* metrics evaluated automatically. The metrics may be related to computed statistical properties of the measurement such as precision (as expressed by the standard deviation of repeated measurements of the same quantity), presumed performance as indicated by previous statistical error analyses against validated datasets, and inter-comparisons within an ensemble of alternative computations.

The DQ2 flag is commonly assigned to data products that report uncertainty bars for the provided values.

Data Quality 3 (VERIFIED-CLEAN): The VERIFIED-CLEAN flag is given to data collections and their computation steps that involve human experts to ensure the removal of data noise and various artifacts.

A typical example of data with DQ3 flag would be manual re-scaling of ionograms to remove the autoscaling errors.

Data Quality 4 (VALIDATED): Scientific data of the best quality that are additionally validated by comparisons against independent measurements or models.

Data with DQ4 designation are usually part of the consortium of models and measurements used collectively for analysis and confirmed to agree in their descriptions. Typical examples arise in event studies involving multiple instruments and models, or specific calibration-and-validation (CalVal) campaigns undertaken to validate novel sensor instrumentation. For example, joint analysis of the peak density height measurements in the ionosphere as observed by ionosonde, incoherent radar, and radio occultation networks can result in a dataset of the VALIDATED quality flag.

An important part of designating the VALIDATED flag DQ4 to an input data stream is evaluation of its *consistency*. The same TID event can be specified by multiple observational means: for example, LSTID in progress is visible to GNSS and ionosonde instrument suites in their different regimes of measurement and associated analysis techniques. Consistency of the ensemble description will be one of the challenging requirements to alert subsystem that detects "TID in progress" events. We will have to harmonize observed features across the sensor networks for a unified consortium specification of the TID detections.

4.1.2 Operational Data Quality

The operational data quality (ODQ) describes technical requirements to support requested scientific data quality. We will recognize the following requirements to ODQ:

Operational Data Timeliness: Given the requirements of the space *weather* service with nowcast and forecast capability, timeliness of the input parameters for T-FORS computations becomes an important quality consideration and an operational requirement. Data resources that cannot provide required promptness of their data streams will have to be reconsidered for implementation. The initial assessments of the expected latency of data availability for T-FORS analysis are provided in Tables 1-3. Further attention will be given to the timeliness of input parameters as the T-FORS techniques mature.

Operational Data Completeness: Completeness of the live input data reports to T-FORS is one of the significant design challenges: missing input means deterioration of the forecast capability. Although operational control over participating sensor networks is beyond the project scope, the ability to recognize gaps and pauses in the sensor coverage will be an important design factor.

4.1.3 Metadata Quality

High quality of metadata that describe the research data resources and services are another important requirement in the modern data management systems. We will comply with the PITHIA-NRF project guidelines on metadata quality (MQ) that, in turn, are based on the Quality Data Management recommendations of the European Strategy Forum on Research Infrastructures (ESFRI) [RD-3]. For brevity of this report, we will omit the MQ component descriptions that can be found in [RD-7].

4.2 Data Quality Monitoring

Real-time monitoring of data quality metrics is a desirable function of the operational systems, albeit not trivial to develop and execute, especially that with automatic failure alerts. However difficult it may be, the outcome is important and even critical for eventual transition to operations. The following additional components of the monitoring system are instrumental to successful implementation:

Data Documentation: Comprehensive documentation includes information about the sensors, data collection processes, and associated metadata. It serves as a reference point for operators to understand the data and ensure its quality. Sensor specifications, calibration procedures, measurement units, and relevant contextual details are commonly included.

Regular Audits: Periodic audits of the infrastructure are conducted to assess data quality compliance. These audits help identify gaps, inefficiencies, or issues in the quality control process. Adjustments and improvements can be made based on the audit findings.

Feedback Loop is established with data consumers and stakeholders. Feedback collection ensures that the data meets the requirements and expectations of the end-users.

4.3 Quality of Data Resources

Quality of data resources concerns quality of data generation, quality of data repository and quality of data usage.

Quality of Data Generation: When generating research data, PITHIA-NRF data providers should examine the data format and documentation.

> **DGQ1. Data format**. Preferred formats are formats designated by a data repository for which it guarantees that they can be converted into data formats that will remain readable and usable. The preferred formats should be *de facto* standards employed by T-FORS communities.
>
> **DGQ2. Documentation**. Providing metadata describing any kind of research resources (data, models, services) is an urgent requirement for T-FORS. The Research Data Alliance (RDA) Recommendation provides guidelines for documenting data by describing the why, who, what, when, where, and how of the data.

Quality of Data Repository: The data repository is responsible for access and preservation of digital research data on the long term. PITHIA-NRF data providers should consider two factors that determine the quality of the data repository:

> **DRQ1. Organisation and processes.** Organisations that play a role in digital archiving and are establishing a Trusted Digital Repository (TDR) minimally possess a sound financial, organisational and legal basis on the long term. Depending on the task assigned to an organisation, a TDR may distinguish itself qualitatively by carrying out research and by cooperating with other organisations in the realm of data archiving and data infrastructure. The outcomes of such research are shared, both nationally and internationally. In addition, these organisations will also share physical infrastructures, software and other knowledge among each other, where possible.
>
> **DRQ2. Technical Infrastructure.** The technical infrastructure constitutes the foundation of a Trusted Digital Repository. The OASIS Open reference model, an ISO standard, is the de facto standard for using digital archiving terminology and defining the functions that a data repository fulfils.

Quality of Data Usage: The quality of the use of research data is determined by the degree to which the data can be used without limitation for scientific research by the various target groups, while complying with certain rules of conduct. The open and free use of research data takes place within the legal frameworks and the policy guidelines as determined by the relevant (national) authorities.

5 T-FORS Data Sharing and Archival

5.1 Making data openly accessible

5.1.1 Openly available data

The new products, generated in the frame of the T-FORS project, will be made openly available subject to the following IPR, which are adopted:

- Background IP (data, models, computer codes, products etc.) must be provided to other project partners, if needed for the project.
- Foreground IP (data, models, computer codes, products etc.) is owned by the partner which generated the IP.
- Foreground IP will be openly available during the project for use within the project, but its wider dissemination should be reasonably protected (e.g. CCBYNCSA) and can be widely disseminated with the Executive Board permission.
- If possible and reasonable, commercially exploitable Foreground IP will be protected through registration and patenting.
- The consortium members are expected to disseminate their project results, subject to protection of the IP, and to do so in open access journals.

The software codes will be released with open access through Zenodo as soon as its operational capacity is confirmed, and a manual is prepared.

The TID detection parameters and the TID drivers will be accessible as soon as the T-FORS database is ready to support the models' execution. The access will be provided through an API, and the files will have a JSON format.

The results from the forecasting models will be provided with open access through the T-FORS API.

5.1.2 Closed and restricted data

Closed and/or restricted data derive from the IPR, and relevant rules being set up in the Grant Agreement and in the Consortium Agreement.

According to the Grant Agreement (Article 16) Background is defined as "data, know-how or information … that is needed to implement the action or exploit the results". Background IPR has been documented in the Consortium Agreement and will be taken into account concerning the data accessibility.

5.2 Making data persistent

Modern data management systems are built for posterity, with a clear plan how data will be curated and preserved, including after the end of the project. This quality of the data collections is often referred to as *persistence* – a characteristic property of data collections that outlive the system that created it originally. In practical terms, it's the process of storing data in a medium that exists beyond the life cycle of the original project. But it is not only about permanent storage; the core persistence principles and practices include:

Storage: Data persistence indeed begins with storing data in a medium that outlasts the lifespan of the running system or application. This persistent storage can take various forms—disk drives, solid state drives, or even cloud-based storage.

Durability: The stored data needs to be able to withstand system crashes, power outages, and other digital mishaps.

Retrieval: In data persistence, the ability to retrieve stored data efficiently is as important as storing it in the first place. This involves data indexing and search mechanisms that can locate and access data with ease. T-FORS will rely on PITHIA-NRF infrastructure that allows advanced modes of data search and retrieval, including search by content.

Consistency refers to the aspect of data persistence that ensures data remains unchanged, regardless of real-life challenges such as power losses and system malfunctions.

Concurrency Control: Multiple processes of data generation, submission, search, and retrieval often need access to the same data simultaneously in the space weather operational configuration. Data persistence includes techniques that ensure concurrent access without losses of data integrity.

At the time of this report, data persistence requirement for T-FORS remains an unanswered concern to be further clarified in time for the D1.3 report on data management plan.

6 T-FORS enhancement areas resulting from best practice analysis

The best practice review conducted for this Milestone 7 report indicates areas for T-FORS improvements and enhancements, especially given that at the end of the project the system prototype will be considered for transitioning to operations. The following topics of T-FORS organization will require additional attention and effort:

Harmonization of alternative methods of TID forecast. The space weather monitoring systems, unlike their terrestrial weather counterparts, face the challenge of "Rich Physics, Little Data" dilemma: available remote and *in-situ* sensor instrumentation is sophisticated and therefore scarce, providing only fragmentary views of the geosystem that needs to be understood and evaluated to predict its behaviour. Given that sensor networks and computational models available in the project will provide only bits and pieces of the system state, the harmonization and fusion of their views will be a hard task that will necessitate data quality assessments in order to blend their streams into a consistent view of the TID phenomena.

Data quality control for TID sensor networks. The TID-in-progress alerts are generated based on internal resources providing real-time sensor data. Most of sensor provide images (ionograms, spectrograms, animated maps) that require intelligent system to extract relevant visual features. These intelligent systems need to be (1) protected against excessively noisy or corrupt data input and (2) supplied with uncertainty metrics to avoid use of unreliable records at face value. The need for additional quality control measures is identified for all of the contributors to T-FORS alerts.