TERN Data Preservation Strategy

Version 1.0

tern
Ecosystem Research Infrastructure

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

Release Authorisation

File Name	TERN Data Preservation Strategy
Version	1.0
Release Authority	TERN Director

Document History

Version	Date	Author	Amendment
0.1	12.01.2022	Anusuriya Devaraju	First draft of the data preservation strategy.
0.2-0.4	12.10.2022	Siddeswara Guru & Lachlan Charles	Review, amend and finalise the document Updates in preservation strategies section
1.0	27.10.2022	Beryl Morris	Minor edits

Purpose

The Terrestrial Ecosystem Research Network (TERN) is committed to storing, preserving, and maintaining access to its digital assets (i.e., data, metadata and data products). This document provides an overview of digital preservation undertaken by TERN to support continuous and reliable access to managed digital assets. Specifically, the objectives of TERN digital preservation are to:

- Support long-term data access securely in a reusable form with appropriate metadata.
- Eliminate data loss due to the obsoleteness of software and hardware
- Protect data integrity and maintain data quality.
- Ensure data preservation and publication adhere to ethical norms, rights, and regulations.

Scope

This document applies to all TERN collected, collated, and managed digital data. However, it does not cover the preservation of administrative records, working documents, or metadata records from partnering organisations aggregated into the TERN Data Discovery Portal (TDDP).

This document applies to all TERN staff and users of digital assets offered by TERN, including data providers and any other affiliates authorised to access the data assets.

Research Data

TERN preserves datasets created or derived from the TERN observatory, collaborative projects and datasets published through TERN from Australian terrestrial ecosystems and environmental science communities. This comprises:

- Raw data – data collected from sources but not processed.
- Processed data – data derived from raw data that is processed and quality controlled.
- Derived Data – data that is not collected directly but is generated from other primary data.
- Modelled data – data derived using models with processed data as input.
- Auxiliary data – companion data associated with processed data to provide more context to data, e.g., data dictionary files, quality control flags and supplementary reports.

The general practice is to preserve all data supplied to TERN but only publish quality-controlled data and data products. A dataset encompasses data files and is published as a 'data collection' on the TERN Data Discovery Portal (TDDP).

Legal and Policy Framework

TERN adheres to the Australian Code for the Responsible Conduct of Research, which signifies the responsibilities of institutions to encourage and support responsible research conduct by establishing and maintaining good governance and management practices and responsible dissemination of research findings.

TERN functions under the University of Queensland (UQ), which is defined as the Lead Agent by the Australian Government funding body. As such, it has legal responsibility for delivering the TERN project. Therefore, the data preservation at TERN is managed in compliance with the UQ's Information Management Policy and Information Governance and Management Framework. This document should be read in conjunction with the following TERN policies and procedures:

- TERN Data Licensing Policy
- Terms of Use
- TERN Data Provider Deed

TERN respects the Intellectual Property (IP) rights, whether or not registered or registrable, of data providers and encourages all data providers to licence their data with Creative Commons Attribution 4.0 International (CC BY 4.0) for unrestricted use wherever possible.

Preservation Strategies

TERN employs several strategies to preserve and deliver adequate and uninterrupted access to TERN digital assets. However, TERN recognises that approaches taken should be flexible enough to accommodate technological evolution, changing standards and user requirements.

Roles and Responsibilities

The following vital roles contribute to the monitoring and implementation of digital preservation in TERN:

a. The TERN Advisory Board has primary responsibility for TERN's strategic intent and recommends to the Lead Agent (UQ) allocation of resources to carry out the TERN digital preservation priorities.
b. The TERN Advisory Board has a data management sub-committee that oversees TERN's data management and preservation, including quarterly implementation reviews to promote sustainable data infrastructure.
c. The TERN Data Services and Analytics' national and international responsibilities are coordinated from UQ where a team with diverse and relevant expertise performs the digital preservation activities, including:
    - Team Leadership - coordinates strategic planning, training, representation, engagement, evaluation, and project QA/QC.
    - Data curators - liaise with data providers to curate and publish datasets with appropriate metadata.
    - Business analysts - elicit and document requirements, including the changing user expectations and technological capabilities.
    - Software developers and data engineers - develop and maintain technical solutions to support continued data access and associated services.
    - System architects - administer the IT data infrastructure of TERN, monitor technological changes that may affect the infrastructure's continuous operation, and ensure the assets and services are continuously accessible and safe.
    - Data providers - supply data to TERN based on the preservation practices recommended by TERN and, if applicable, provide feedback on data, services and guidance on new requirements.

Metadata and Documentation

Metadata plays a vital role in data reusability. Therefore, appropriate metadata is required for the associated data to improve its comprehensiveness and reuse. To achieve this, TERN carries out the following activities upon data deposition:

- Determine the suitability of data and if it fits the TERN's repository data publication scope.
- Review submitted data and metadata to ensure that sufficient metadata (descriptive, provenance, technical and rights) are specified and the metadata adheres to the international standard (TERN Metadata Profile of ISO 19115-3:2016 and ISO 19157-2:2016) to aid future access.
- Check if metadata includes supplementary information (e.g., related documentation) to promote data reuse in future.
- Enrich metadata with controlled vocabularies to represent all artefacts of data and their contextual information as outlined in the TERN Vocabulary Development and Management.

Given that metadata standards evolve over time, TERN updates metadata records to the latest standard as and when required, e.g., from 19115-3 to 19115-3.2018.

TERN ensures that data is accompanied by metadata when archived and accessed from TERN data infrastructure, i.e., through self-describing formats (e.g., BagIt and NetCDF) and supplementary files (e.g., data dictionary, file-naming convention).

TERN documents internal procedures to provide guidance on metadata and data publication. For example:

- TERN SHaRED User Guide is a manual to guide the data providers to deposit data through the TERN data submission tool (SHaRED).
- TERN Data Publication Checklist is a guide intended for the data curation team to review data submissions.
- TERN NetCDF User Manual specifies conventions used by TERN to disseminate data in a NetCDF format.

For a complete list of documentation, see TERN Knowledge Base.

File Format

File formats may become obsolete over time and in some cases, data is unusable when software cannot read data files from older file formats. Therefore, TERN recommends a set of file formats that are likely to be accessible now and into the future.

To ensure that the data deposited into TERN follows recommended file formats, TERN data curators:

- Check the data format against the recommended formats at the point of data submission, e.g., (through SHaRED) and ingestion into the TERN data staging area or a specific server. If needed, TERN requests the data providers to convert files into an appropriate format (e.g., Excel to CSV).
- For data generated from the TERN observatory and collaborative projects, TERN migrates data from the original file format to a more suitable format when deemed appropriate. For example, TERN accepts raw phenology images and publishes them in JPEG. In addition, TERN

converts raster data files from GeoTIFF to Cloud Optimised GeoTIFF (COG) for efficient access in the cloud environment.

- TERN regularly assesses and updates recommended file formats that are supported within the TERN infrastructure.

Data Versioning

Procedures are in place, where feasible, to track changes to the preserved version of a data collection and its data files. For example:

- Data Submission – the TERN data submission tool (SHaRED) tracks data collections and maintains relevant versioning metadata such as date and time, editorial status and notes, and users.
- Data Staging – the data files uploaded into the TERN CloudStor staging area follows the agreed, consistent, and descriptive folder and file naming conventions. The folder structure and files are archived in file system storage for future access.
- Data Processing and Archival - the Extract, Transform and Load (ETL) pipelines will verify if data files have been altered through checksum upon transfer from the staging area to archival storage.
- Data Dissemination – TERN publishes data collections with a DOI upon request. Otherwise, all data collections are assigned a unique identifier (UUID). The metadata of a published collection includes data created, modified dates and a version number. In addition, data from the EcoPlots and EcoImages portals are disseminated in self-described data packaging format (BagIt) with checksum information for data files.

Data Withdrawal

The published data collection and associated metadata record will not be deleted. However, in an unlikely situation where a data collection is requested to be deleted, records will be marked as "retired" in the GeoNetwork metadata catalogue with an additional statement on reasons to retire the data. If any dataset is superseded with a new version of the data, the status of the original dataset will be updated to "deprecated".

Data Retention

The University Sector Retention and Disposal Schedule applies to records created in all formats in the Queensland Public Universities. Under this Schedule, TERN data is classified as Research Data (601.2/C123) of high public interest and significance to the discipline. Therefore, TERN permanently retains all the data deposited in an open and community-endorsed, standards-based format along with the contextual metadata. In addition, TERN, through the QRIScloud data centre, participated in the ARDC Data Retention Program to secure long-term storage infrastructure for all TERN data collections.

Storage

To prevent data loss in unplanned events (e.g., disaster, human error, and hard drive damage), data should be adequately backed up onsite and offsite and migrated to suitable media over time.
TERN uses storage infrastructure managed by Australian Research Data Commons (ARDC) and AARNet. Hence, TERN relies on the disaster recovery process followed by ARDC and AARNet. TERN eInfrastructure uses a combination of data storage (volume storage, object storage and file systems

managed by QRIScloud collections) at the ARDC Nectar Research Cloud to preserve TERN data, databases, and associated services:

- Volume Storage, e.g., Postgres databases.
- Swift Object Storage for large, static individual files such as images, gridded raster, and data files uploaded via the SHaRED data submission tool.
- QRIScloud collections (QCollection) are file systems that can be accessed through the Network File System (NFS) protocol. For instance, TERN remote sensing raster images and flux data are hosted as QRIS collections and available through the TERN THREDDS server. Audio files from acoustic sensors are also hosted on QCollection and available from an HTTP file server.

TERN also uses CloudStor (a public cloud-based file storage service provided by AARNET) as a data staging area.

All storage systems used by TERN support backup and replication managed by the data centre. In addition, TERN performs regular backups of databases and configuration files:

- Volume storage (databases) is backed up daily and moved to a file system storage at the Research Data Storage (RDS) of QRIScloud.
- Object Storage is replicated with copies archived at a minimum of 3 different geographical locations.
- QRIScloud collections are replicated to offline storage to protect them against media failures and significant catastrophes.
- CloudStor supports tape backups and replicates the data files at two geographically separated storage nodes.

The TERN system architect manages and administers access to the data and the copies retained at the storage systems. In addition, TERN performs a six-monthly review of sample records in the storage systems to ensure that they are readable and not corrupted.

Discovery and Access

Data preservation enables long-term access to data. Accessible datasets are those that can be discoverable and accessible for continuous reuse. Users of TERN data are obliged to attribute any data collection that they use and comply with legal and ethical obligations articulated in the TERN Data Licensing Policy.

TERN offers various applications to facilitate data discovery and access:

- TERN Data Discovery Portal (TDDP) is the central place to find all published data collections. The portal harvests metadata records of all collections from the TERN GeoNetwork Metadata Catalogue.

Since TERN manages diverse data types, the dissemination of data happens through several applications:

- For the data collections with files uploaded through SHaRED, data files are stored in the Nectar Research Cloud object store and attached with the collection metadata.
- Phenology images and plot observations are accessible through EcoImages and EcoPlots, respectively. Each of these applications offers an API for programmatic access to data.
- TERN THREDDS Data Server (TDS) is used to disseminate NetCDF and image files (tiff, GeoTiff) derived from remote sensing and models.

- HTTP server is used to publish acoustic and raster images from remote sensing and UAV.
- Spatial layers are shared through Web Mapping Service (WMS) supported by the TERN GeoServer. The TERN Landscape Data Visualiser visualises these spatial layers.

TERN has invested in curating and publishing controlled vocabularies to support standard representation of parameters, feature types and related artefacts. TERN Linked Data Services is the main site to discover the semantic model and all vocabularies used by TERN.

Security and Risk Management

TERN preserves and disseminates data following the UQ Cyber Security Policy and manages the cyber security risks of the data infrastructure according to the policy. In addition, all TERN staff undertake cyber security awareness training as part of the employees' onboarding.

TERN hosts all TERN data and services on Research cloud infrastructure developed and managed by ARDC and AARNet, which support appropriate procedures to protect resources on their respective networks. TERN has control access to folders and files at the data staging area and archival systems through read and write authorisation. Users have read-only access to all archived and published data collections to prevent unauthorised changes. Any actions on files are captured by appropriate applications, including the data submission tool (SHaRED), CloudStor and server log files of applications.

TERN undertakes all reasonable steps to safeguard, use and manage personal information supplied to the data infrastructure in compliance with the UQ Privacy Management Policy:

- The data providers' personal information (name, organisation, and email) is collected with their awareness and stored by TERN for disclosure in the metadata so that authors receive attribution and can be contacted by users of the data for further information.
- The personal information of users of selected data services is used to generate aggregated usage statistics for the funding agencies and to communicate with users in case of service disruptions and updates.
- The TERN Terms of Use inform users how personal information is collected and used.
- Data providers can lodge any concerns about how their personal information are collected and disseminated through the TERN eSupport Service.

Acknowledgement

This document was developed in consultation with the following resources:

1. Digital Preservation Handbook
2. Principles and Good Practice for Preserving Data

We at TERN acknowledge the Traditional Owners and Custodians throughout Australia, New Zealand and all nations. We honour their profound connections to land, water, biodiversity and culture and pay our respects to their Elders past, present and emerging.

TERN is enabled by NCRIS. Our work is a result of collaborative partnerships with many universities and institutions.

To find out more please go to **tern.org.au.**

tern
Ecosystem Research Infrastructure

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative