Tethys RDR Preservation Plan

Authored by

Francisco Porras-Bernardez
Viktoria Haider
Christian Linsberger
Werner Stöckl
Arno Kaimbacher
Thomas Brus

Contents

December 2023; Version 1.1

1. Introduction

Tethys RDR is a digital geoscientific Research Data Repository (RDR) of GeoSphere Austria, formerly of the Geological Survey of Austria (GBA).

The content and purpose of the repository is based on digital, georeferenced research data generated at or on behalf of GeoSphere Austria should be published and archived. The georeferenced data publications provided by the repository and the associated metadata can be freely viewed and - in accordance with the legal regulations - used.

In principle, all data publications are visible, with the exception of those that have been subject to an embargo. This data is only available at a defined point in time. However, the associated metadata is publicly available.

The Tethys RDR is currently provided in German and partly in English. A full English version is planned in the near future.

The data publication guidelines of the repository as RDR Policy contain basic information for handling research data and ensure a uniform quality of the data. If they are observed, a solid, transparent and sustainable data publication can be guaranteed.

In the Tethys RDR Handbook[1] it is regulated which data can be published, how, in which form and to what extent. It also provides information on rights and obligations.

This repository is based on two principles that are known and recognized in the scientific world and enable the sustainable archiving and publication of reusable scientifically relevant data: Open Science Principles and FAIR Guiding Principles[2].

The main purpose of Tethys RDR is the long-term digital preservation of current and future scientific outputs produced by GeoSphere Austria. This preservation plan describes the repository's approach for defining and implementing preservation actions. The document summarises deposition requirements (metadata schemas and data formats), legal aspects, management of assets, management of the threat of obsolescence and technical preservation arrangements.

[1] https://www.tethys.at/docs/HandbuchTethys.pdf

[2] WILKINSON et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. – Sci. Data 3: 160018. http://dx.doi.org/10.1038/sdata.2016.18

2. Deposition requirements

2.1. Metadata schemas

Three widely used metadata standards are used in Tethys RDR: Dublin Core[3], ISO 19139[4], and DataCite[5]. Tethys not only publishes supplementary research data for text publications, but also other data types. Hence, another purely descriptive grouping element ("dataset type") was introduced. This includes the seven types[6]: analysis, measurements, monitoring, remote sensing, geographic information system, models and mixed type. These categories are used to classify the datasets and are intended to facilitate filtering by common characteristics, especially in the case of a growing database.

2.2. File formats

There is a group of formats that meet the criteria of "openness", "security" and "sustainability". Tethys RDR currently supports the following formats for the data and its supplements: ascii-grid, csv, gpkg, jpeg, pdf (PDF/A), png, shp, txt, xlsx and netCDF.

To ensure that data remains legible and therefore usable even after 10 years and much later, it is necessary to publish them in the appropriate formats. Formats that can only be specifically read by one or a few selected proprietary software types and that lose their validity and thus also their validity at the latest with the software updates are to be excluded as a matter of principle lose unrestricted legibility. This would also strongly contradict the FAIR Guiding Principles.

For this reason, formats that implement open standards and are not dependent on proprietary software should be used. Self-made or encrypted formats are also categorically excluded as data publication for Tethys RDR.

The currently supported formats[7] as well as the planned formats are listed in the Tethys RDR handbook.

3. Legal aspects

According to Open Data Principles, only some[8] Creative Commons (CC) licenses are available for publication in Tethys (currently version 4.0). Once the research data publication has been successfully published, licenses that have been granted can no longer be changed.

If, among other things, third-party data is processed when the data is created, submitters must check which copyrights or rights of use are bound to the third-party data. If the third-party data have already been provided with CC licenses, the licenses can be used to determine whether the data may be used or under which license the newly created data can be published.

[3] https://www.dublincore.org/ (Accessed on 19.06.2023)
[4] https://www.iso.org/standard/32557.html (Accessed on 19.06.2023)
[5] https://datacite.org/ (Accessed on 19.06.2023)
[6] https://www.tethys.at/docs/HandbuchTethys.pdf#page=10
[7] https://www.tethys.at/docs/HandbuchTethys.pdf#page=21
[8] https://www.tethys.at/docs/HandbuchTethys.pdf#page=20

If the third-party data has individual copyrights or rights of use, submitters have to check carefully whether the data may be processed and with which CC license these rights are compatible. In case of doubt, submitters should clarify this with the help of a legal expert. If multiple third-party data with different licenses have been used, all terms of use must be considered in their entirety in order to correctly license the newly created data.

The authors of the publications keep the intellectual property rights to their data. The submitters have to grant copy, transformation, storage and distribution rights to the repository by agreeing to both the Data Policy and the Terms & Conditions. This is done by the submitters when uploading the publication through their frontend. The CC licensing allows that no separate right of use has to be formulated and the user does not have to obtain permission to process the work every time. The existing right of use must allow at least one further processing.

The submitters endeavour to observe the copyrights of all graphics and texts used in all publications, to use graphics and texts they have created themselves or to use license-free graphics and texts. All brands and trademarks mentioned on the Tethys website and registered by third parties are subject without restriction to the provisions of the applicable trademark law and the property rights of the respective registered owner. The conclusion that trademarks are not subject to the rights of third parties should not be drawn solely on the basis of the mere mention. The content on the repository's website is licensed under a Creative Commons Attribution 4.0 license

A non-compliance from an individual user or organization at deposit, during curation/preservation, and during access and reuse could trigger legal actions from GeoSphere Austria.

4. Long-term preservation

All the objects in the repository are subjected to the same level of responsibility for preservation.

4.1. Object re-appraisal

Re-appraisal of objects will be performed whenever long-term preservation formats are deemed as obsolete. In such case, the affected objects will be migrated to the new formats by the Tethys RDR team.

During the deposit process, the team manually evaluates compliance with the rights management policy affecting every object. During curation and preservation, any detected non-compliance by the repository's staff or its users might trigger a removal of assets (see 4.3.). The number of visits and downloads is monitored, and this serves as an indirect estimation of the access and reuse of the digital objects. However, a proper monitoring of rights compliance during these phases can only rely on a passive approach consisting on offering communication channels to users or submitters. Then, both groups could report about any detected violation of the licensing conditions of a publication.

Tethys RDR understands its duty to keep track of any developments that could impact the sustainability and long-term availability of its data. As a result, we consistently monitor advancements in technology and science, and we incorporate them into our technical and organizational strategies. Our planning includes

a biannual revision of the accepted data formats, the data submission forms and the Handbook, among other documents (see section 7).

4.2. Minimum stated retention periods

The guaranteed minimum retention period for the digital objects published in the repository is **10 years**. For this purpose, several elements are expected for each publication:

- A clear data explanation containing the description of the published datasets in such a way that can be understood and used in the future according to scientific "best practices".
- Data formats appropriate for long-term preservation and reuse.

4.3. Removal of assets

Once the data has been published, it can no longer be changed, supplemented or deleted. In exceptional cases, access to datasets can be blocked, but the metadata are retained without exception, since citability and scientific traceability should be guaranteed at all times.

Data published in Tethys RDR is treated like print publications. Once the data has been published, neither content corrections nor major corrections to the metadata can be made. If minor corrections to the metadata are required due to spelling errors, they will be corrected without creating a new version of the data publication. It is therefore of great importance to prepare the data publication carefully and to think carefully about how the datasets are published and under what license conditions.

The Tethys repository assigns a DOI for each published dataset, which means that once the data is published, it cannot be changed or deleted. In exceptional cases (violation of legal rights or subject of a justified complaint), access to the datasets in question may be blocked so that only the DOI can be cited (fig. 1). In this case, the DOI redirects to a landing page that explains why the dataset had to be removed. This ensures that the citability and scientific traceability of the dataset are maintained even if access to the dataset itself is restricted. See also the general guideline for publishing research data in the handbook[9].

The deletion of a record initiates the following process:

[1] Conduct an investigation: The repository may conduct an investigation to determine whether the dataset or metadata record is in fact in violation of legal rights or is the subject of a legitimate complaint.
[2] Notify the submitter: If the dataset or metadata record is found to be in violation, the repository may attempt to notify the depositor of the issue and the reason for removal.
[3] Remove the dataset or metadata record: If the violation is confirmed, the repository administrator removes the dataset or metadata record from its collection.
[4] Document the removal: The repository administrator documents the removal of the dataset or metadata record, including the reason for removal and any communications with the submitter.
[5] Review and revise policies: The repository may review and revise its policies and procedures to prevent similar violations from occurring in the future.

[9] https://www.tethys.at/docs/HandbuchTethys.pdf#page=18

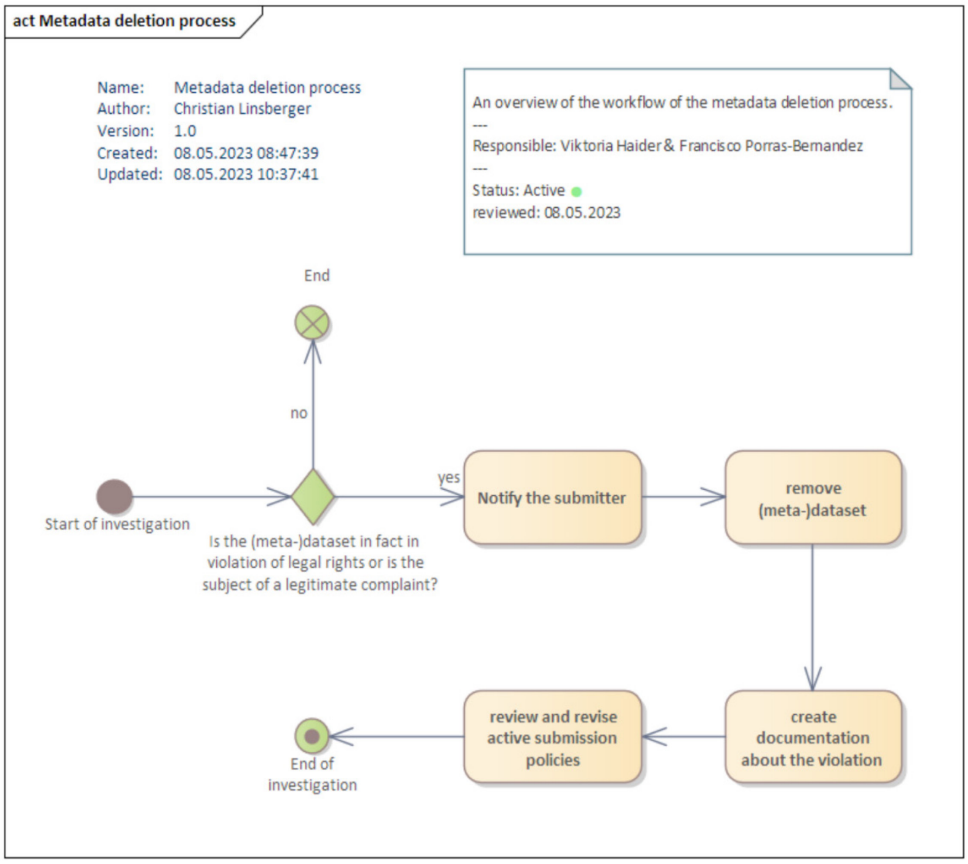

Figure 1: (Meta)data deletion process

In order to assure a Long-term Availability (LTA) of formats, data publications are only allowed in sustainable and standardised formats. If it is necessary to accept the data in an alternative format because (1) there is no satisfactory alternative and (2) there is general acceptance by the intended community, additional data must be provided in a sustainable format, even if this could involve a loss of information. To ensure the LTA of a truly referenceable data publication, the repository will not alter the data publication. If an update is required, Tethys RDR will favour versioning the data publication and would publish the data in a new version.

5. Management of the threat of obsolescence

Software development

There are several processes that are used to monitor and manage the need of technical changes in software development of Tethys.

Change management: This involves establishing a formal process for requesting, reviewing, approving, and implementing changes to Tethys software development. Changes may include new features, bug fixes, or modifications to existing functionality.

Version control: This enables the developer team to keep track of changes to software code and ensure that everyone is working on the most up-to-date version. This is done using Git version control.

Continuous integration/continuous delivery (CI/CD): These are practices that involve automating the building, testing, and deployment of the software. This helps catch errors and ensure that changes are released into production environments as quickly and reliably as possible.

Agile methodologies: Agile development focuses on iterative development cycles, with frequent feedback and collaboration between developers, stakeholders, and end-users. This approach helps to ensure that software development stays aligned with changing requirements and priorities.

Formats monitoring

Well established metadata schemas and data formats for long-term preservation are enforced on the submitters. A periodical evaluation of the adequacy of schemas and formats is done and the repository would adapt fast to any change.

In case that a migration of the digital objects to new formats is necessary, a migration plan will be developed specifically targeting the affected objects.

Continuity of services

The 2022 Austrian Federal Law on GeoSphere Austria (GeoSphere Austria Act - GSAG[10]) determines that the organization has the mandate to establish and operate a central data infrastructure as a service for science, businesses, public administration and society. This implies being a sustainable and reliable operator and our repository is a fundamental tool to achieve this objective for the geoscientific data generated within the organization.

Tethys RDR is designed in such a way that the database behind it can be easily transferred to another standardized digital asset management system (DAMS). In the unexpected case that Tethys RDR as such can no longer be actively continued, the existing data publications can be transferred with their DOIs into another DAMS software.

[10] https://ris.bka.gv.at/geltendefassung.wxe?abfrage=bundesnormen&gesetzesnummer=20011885&ShowPrintPreview=True (Accessed on 19.06.2023)

6. Technical preservation arrangements

Tethys RDR aims to provide secure, sustainable, and long-term storage of research data. This means that the repository is designed to ensure that research data is stored in a way that ensures its integrity, security, and accessibility for the long-term, even as technology and data formats change over time. In order to achieve this, various conditions must be met. The following processes and documents describe how data is stored securely, backed up regularly, and preserved for the long-term:

Custody transfer

Prior to the creator's data upload onto Tethys for publication, it undergoes a review process conducted by Tethys editors. The resulting drafts of the publication are then stored or archived within a dedicated, restricted-access directory exclusively provided for Tethys. Once the data and metadata are deemed publication-ready and uploaded onto Tethys, they are stored in the Digital Asset Management System, ensuring their integrity and preventing any unauthorized alterations or replacements. It is important to note that any data rejected by the editor during the final upload process will not be archived.

Preservation information metadata

After publication, submitters cannot update their data files or metadata without the involvement of the Tethys RDR staff. The curation process ensures that any updates made to the dataset are properly reviewed, approved, and documented, thereby maintaining the accuracy and integrity of the repository's contents. Major changes to the file(s), title, or authors result in a new version that cannot be published without undergoing another curation process. The new and the previous dataset are cross-referenced in their respective descriptive metadata. Changes made to any other metadata fields (e.g. keywords, description, item type) will not result in the creation of a new version. These minor changes to data or metadata that are not publicly available to users are tracked by assigning a new internal version number in the database.

Disaster recovery plan

The disaster recovery plan for Tethys specifies how the repository will be recovered from a data loss or system failure. The plan includes procedures for restoring data files from backups, recovering the database and restoring the access to the repository. In the recovery plan there are scenarios for recovering specific data files (if checksum test fails), whole folders for all files of a specific dataset and restoring a file in all available backup versions. In addition, the restoration of the entire IT system using Docker containers is described in detail.

Data Architecture Diagram

With the help of the Data Architecture Diagram[11] the whole repository staff has a clear understanding of the management off all storage locations. Tethys RDR collects and manages scanned geological maps,

[11] https://gitea.geologie.ac.at/geolba/tethys.backend/wiki/DataArchitectureDiagram

spatial data sets, and other types of research outputs. The diagram is providing information on how to prepare data sources for deposit, like licenses, correct use of keywords, file formats and file upload limits.

Backup strategy

IBM Spectrum Protect (formerly known as Tivoli Storage Manager or TSM) is used to protect the data stored in the Tethys Repository. Tivoli Spectrum Protect provides a comprehensive backup and recovery solution for research data, which can help ensure data availability, integrity, and recoverability in case of a disaster or data loss. For the repository up to 90 incremental versions of each data file will be backed up and are available for recovery. This means that there are 90 possible instances of the data. These versions are typically created based on the backup schedule and retention policies defined by the computer centre of *GeoSphere Austria*. The digital objects stored on Tethys RDR are backed up daily.

Risk management techniques

For a research data repository, they involve identifying potential risks to the data, assessing the likelihood and impact of those risks, and implementing strategies to mitigate or manage those risks. The following risk management techniques are used for the Tethys RDR:

- Access Controls to protect against unauthorized access: The access to the administrative backend is limited to authorized users only. To provide a secure connection, only HTTPS is allowed. Fail2ban [12]protects the repository server from brute-force attacks, denial-of-service attacks, and other malicious activities. It works by monitoring log files and dynamically updating firewall rules to block IP addresses that are exhibiting suspicious behaviour.
- Back up data regularly: up to 10 incremental backups of the data are maintained to ensure that it can be recovered in case of data loss or corruption.
- Encrypt sensitive data: Sensitive data, such as personally identifiable information and passwords are encrypted inside the PostgreSQL database.
- Monitor activity logs: Activity logs of the webserver are monitored via fail2ban to detect suspicious activities, such as unauthorized access attempts or data exfiltration.
- Implement data retention policies: Policies for the data retention and data deletion process are implemented.
- Conduct regular security assessments: The security of the repository is regularly assessed by the repository staff to identify potential vulnerabilities and implements strategies to address them.

The potential disruptive physical threats, which can occur at any time and affect the normal business process, are listed below:

- Fire: Fire suppression systems are installed, there are fire and smoke detectors on all floors and there are also fire compartments.
- Electric power failure: Redundant UPS systems with standby generators are available. (Monitoring: 24/7)
- Communication Network loss: Unfortunately, there is no redundant repository sever in case of network loss. By monitoring the network in real-time and receiving alerts when network loss is

[12] http://www.fail2ban.org/ (Accessed on 19.06.2023)

detected, the IT department can quickly investigate and resolve issues before they impact end-users.
- Flood: All critical equipment is located on a second floor.
- Sabotage: Only authorized IT personal have access to the server room.

Handling and monitoring of storage media deterioration

Tethys RDR calculates internal checksums during the ingestion workflow. These checksums ensure that ingested data has not been altered or damaged. If even a single bit of the data changes, the checksum will also change, indicating that the data has been changed unintentionally on the file store. During the file upload, Tethys calculates and stores MD5 and SHA512-checksums for each file.

Data Integrity

For internal fixity checks, Tethys Repository operates an automated cron job that routinely tests all the MD5 and SHA512-checksums for data stored by the Tethys Repository and produces a regular report providing appropriate warnings if a silent data corruption is detected in the storage layer. Corresponding code of the cron job can be downloaded via Tethys Code repository. If the web backend is in production mode, the logger writes the error messages as a mail to the administrator. Depending on these warnings, the administrators will investigate the cause of the changes and the corrupted files will be restored from the backup (IBM Spectrum Protect).

7. Update of the plan

This plan will be reviewed every two years and the revision process will be initiated by the repository management. The following table shows the planned schedule for the revision of the individual aspects affected by the plan:

Process	Frequency	Responsible
Revision of the Preservation Plan	Biannually	Tethys managers
Revision of preferred data formats for long-term preservation	Biannually	Tethys team
Changes in legal or regulatory aspects	Biannually	Tethys managers
Hardware failures	Daily	IT department
Workflows	Annually or when needed	Tethys team
Revision of the Handbook	Biannually	Tethys managers
Revision of data submission forms	Biannually	Tethys managers
Renewal of certification	Every 3 years	Tethys team