TETTRIs

TETTRIs

Transforming European Taxonomy through Training, Research, and Innovations

Deliverable 9.1

Data Management Plan

WP 9/CETAF

Contributors: Ruth Lagring (RBINS), Anton Güntsch(FMG-BGBM)

Document: Deliverable
Version: v.04
Date: 26/07/2023

Funded by the European Union

TETTRIs related product

Identification	Value
Title	D9.1 - Data Management Plan
Author(s)	Marko Lovric, Ana Casino
Affiliation	CETAF
Contributors (Affiliation)	Ruth Lagring, Anton Güntsch, Marta León
Publisher	CETAF
Identifier of the publisher	
Doc. Version	v03
Resource	Report
Publication year	2023
Sensitivity	Low – Public document
Date	28/07/2023
Citation	

Abstract: This first version of the DMP establishes the initial considerations for Data Management with the information provided by the TETTRIs project patterns via an initial survey. This preliminary information is used to intrude a framework for data collection, storage and usage for the TETTRIs Project. This DMP constitutes a living document that will be updated and further elaborated as the project advances. This document will tackle the origin of the data, its storage and its usage during and after the project life.

Keywords: Data Management, Data Management Plan, Data Storage, Data Usage, Origin of Data, Open Access, Third Party Projects Data, Personal Data, Sensitive Data, Data types, etc.

Revision:

no.	Reviewer	Status	Notes
1	Wouter Addink	Needs revision	Comments regarding emphasis on management data after the end of the project, 3PP Project data, control mechanisms, FAIR measurements, provision of an overview of data.
2	Quentin Groom	Needs revision	

Settings	Value
DELIVERABLE	
Deliverable title:	Data Management Plan
Deliverable n°:	D9.1
WORK PACKAGE	
Work Package	9
Work Package Leader	RBINS
Task	T9.3
Task Leader	CETAF
Type:	Document
Lead beneficiary:	Ana Casino
Citation:	ORCID: 0000-0002-9869-6573
Due date of deliverable:	M8
Actual submission date:	26th of July 2023
Deliverable status:	Approved

Document Control Information

Settings	Value
Document Title:	9.1 Data Management Plan
Project Title (Acronym):	TETTRIs
Document Owners:	Marko Lovric, Ana Casino
Project Coordinator	Frederik Hendrickx, RBINS
Doc. Version:	v03
Sensitivity:	Public
Date:	26th of July 2023

Document Approver(s) and Reviewer(s):

NOTE: All Approvers are required. Records of each approver must be maintained. All Reviewers in the list are considered required unless explicitly listed as Optional.

Name	Role	Institution	Action	Date
Marta León	PM	CETAF	Contributor v01	7 June 2023
Ruth Lagring	Contributor	RBINS	Review v02	14 June 2023
Anton Güntsch	Pillar 2	FMB-BGBM	Review v02	14 June 2023
Wouter Addink	Reviewer	FUB-BGBM	Review v03	18 July 2023
Quentin Groom	Reviewer	Meise BG	Review v03	14 July 2023
CETAF	Approver	CETAF	Approve v04	26 July 2023

Funded by the European Union

Document history:

The Document Author is authorised to make the following types of changes to the document without requiring that the document be re-approved:

- Editorial, formatting, and spelling
- Clarification

To request a change to this document, contact the Document Author or Owner. Changes to this document are summarised in the following table in reverse chronological order (latest version first).

Revision	Date	Created by	Short Description of Changes
v01	27.05.2023	Ana Casino and Marko Lovric	
v01	6.06.2023	Marta León	Contributions
v02	26.06.2023	Anton Güntsch and Ruth Lagring	Comments on categories of data, standards, secured repositories.
v03	18.07.2023	Wouter Addink and Quentin Groom	Comments on post project DM, emphasis on 3PP Data, Control mechanisms, overview of data
v04	26.07.2023	Marta Leon	Comments addressed

Configuration Management: Document Location

The latest version of this controlled document is stored in location

Funded by the European Union

Table of contents **Error! Bookmark not defined.**

1. Preface 6
2. Introduction 6
3. Project Data 7

2.1 Origin of data 8
2.2 Collection of data 8
2.3 Data storage 9
2.5 Personal data 11
2.6 Sensitive data 11
2.7 Data Embargos 12

4. Third-Party Projects 12
5. Data usage 12
6. Accessible Data 12
7. Final remarks 13

APPENDIX 1 14

Funded by the European Union

1. Preface

This first version of the DMP establishes the initial considerations for Data Management, with the information provided by the TETTRIs project patterns via an initial survey. This preliminary information is used to lay a basis on data collection, storage and usage for the TETTRIs Project. This DMP constitutes a living document that will be updated and further elaborated as the project advances.

An inventory will be considered for partners to structurally list, for instance, the various data types per WP (digital images, expertise, taxonomy, services), including some other essential information (eg. parameter group, legal regulations (GDPR, INSPIRE,...), licences, repository for data dissemination (GenBank, ORCID, GBIF,...), storage location during the project, data formats, existing data or new data, 3PP data (Y/N), etc. This will allow a better understanding of the data management needs. While the responsibility of the data management is decentralised to each partner, discussions and meetings will be held regularly at consortium level to tackle any challenges, improve the data management, update procedures and this document etc. A dedicated space will be foreseen every GA as well at EB meetings upon partner request.

This document will tackle the origin of the data, its storage and its usage during and after the project life. The next official document update is foreseen for M18.

2. Introduction

TETTRIs is a HORIZON EUROPE (HE) funded project which aims to transform European taxonomy through, among others, facilitating knowledge sharing and improving access and use of data based on Open Science principles. These principles are one of the main drivers for the development of the project and are critical for articulating the dissemination and exploitation of results. Using the principle of "As open as possible, as closed as (legally) necessary", TETTRIs will implement a thorough and consistent Open Science approach. The project partners will make the project data available to parties outside of the TETTRIs consortium (e.g. next generation taxonomist, users in need of taxonomic knowledge, decision makers, etc.), under applicable regulatory provisions (e.g., GDPR, IPR, Open Data Directive) and recommendations (to cover e.g. Ethics, Gender and Diversity dimensions). As such, structured and regular management of data is an essential part of the transformative process for the use of taxonomic knowledge that TETTRIs aims to launch and will ensure a successful and sustainable development beyond the project lifetime.

For this purpose, this Data Management Plan (DMP) will be the main reference document for setting the framework and laying down provisions with regard to how data generated within TETTRIs will be collected, produced, stored and used. The Plan, of which the current document is the first iteration, is envisaged as a living document in accordance with the Horizon Europe Model Grant Agreement, to be updated several times during the project (specifically in M7-v2, M18-v3, and M33-v4). TETTRIs will establish the processes to ensure the DMP is well implemented in terms of responsibilities, data documentation filing and standardisation. Furthermore, it will ensure confidentiality in the context of its Third-Party Projects (3PPs) and set the conditions for archiving received proposals and their resulting outcomes.

As a baseline, the DMP follows the research data management policies of the institutions generating the data as well as the FAIR (Findability, Accessibility, Interoperability, and

Reusability) Guiding Principles for scientific data management and stewardship. Furthermore, it adheres to the stipulations under the TETTRIs Grant Agreement relating to confidentiality, security and data protection (see articles 13, 15, 16 and especially 17).

The main body of the document, divided into several sections, collates and analyses the key components related to data management based on the input from partners, collected via an online survey as the first step towards drafting the DMP. The survey was completed at Task, or WP level when so applicable, by 17 partners which covered the whole spectrum of data foreseen in TETTRIs. Each of these sections is then further complemented by in-house expert recommendations where applicable and when necessary.

3. Project Data

As mentioned in the previous section, the information on the data involved in TETTRIs presented here is obtained mostly through the online survey. Considering the complexity of the project, the data described will be interacted with in various ways depending on the goals and objectives of Tasks under which they will be. For instance, the data generated by Task 1.3 'Developing virtual reference collections' consist of digital images of European pollinators and associated metadata. These images will be used to pilot a virtual reference collection of pollinators. In a different example, Task 3.2 'Automatic mapping of taxonomic expertise' will generate a workflow (scripts) that extracts personal data from online open resources. Regardless, all TETTRIs related data will be processed in compliance with the EU GDPR Regulation (Regulation 2016/679[1]).

Moreover, many of the work packages and tasks will also depend on data sharing among them. To illustrate this, we can use the example of Task 3.1, which is about creating the marketplace for expertise, taxonomic e-services and resources. Among others, it should include data about taxonomic tools that are implemented and validated in biodiversity hotspots, so all the work packages working on those tools will need to contribute to the data. The platform will be web-based, using open source software that will be linked to the CETAF website to ensure its maintenance and sustainability. The data will be relevant for, among others, to link with platforms for international access to reference collections (T1.2, T1.3) and it may also be of some use for both, T.5.2 'Inventory of taxonomic training and roadmap for taxonomic facility contribution to university curricula', and for T.8.3 'Knowledge transfer mechanisms'.

Given the variety and the interconnected nature of the data usage under TETTRIs, it is necessary to provide an overall analysis of the different aspects of data in question as well as to provide some general good practices regarding its handling. The paragraphs below aim to do precisely that, while keeping in mind that additional information will likely be available in the later iterations of the Plan.

When it comes to the categories of data within the project, i.e. the object types collected, we are presented with a varied range of categories that are directly linked with the content and work of the Task itself. Generally, references to persons and articles will be recurrent throughout the project and during its entire lifetime. For publications, specific scientific citation rules commonly used will apply. In Tasks related to natural science collections, data

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC ('GDPR') (OJ L 119, 4.5.2016, p. 1).

will refer mainly to Taxon names, and specimens, and it will scale up depending on the granularity of the information needed. The use of standards and stable identifiers, pipelines and references will be encouraged when available for the community as agreed elements to refer to. Taxon Treatments and Sequences will be somewhat less frequent but still present mainly when molecular methods are in the scope of the work (e.g. WP6).

2.1 Origin of data

The origin of data used under the activities of TETTRIs will be a mix of publicly available datasets, private (institutional) data and self-generated data. A vast majority of partners will work with publicly available data, while many of them will also generate new data. A smaller number of tasks will also acquire private/copyrighted data for which obtaining a permit will be required. DNA sequencing will be used for generating new data, specifically for developing and validating molecular methods.

Publicly available datasets	59%
Institutional or private data (copyrighted; permit required)	35%
Self-generated data (By the project and partners)	53%
Other	12%

Figure 1: Table showing the distribution of the origin of the data according to the initial survey.

2.2 Collection of data

The types of data collected for the purposes of the research conducted under TETTRIs vary in terms of their format, size and category (object types), from Task to Task. Consequently in TETTRIs, several different data collecting processes will be employed, of which the most popular methods appear to be literature review/harvesting, interviews, surveys and participatory mechanisms in events. In certain Tasks, data will be obtained from existing natural history collections (hosted by either TETTRIs partners or other collections-based organisations), but also through field work. DNA sequencing will be used for generating new data, while specifically for developing and validating molecular methods under WP6.

TETTRIs will foster the use of stable and persistent identifiers (PIDs) at data and dataset level depending on the specific application of the data (existing and newly collected) following the already existing structures and recommendations (e.g. CETAF Identifiers Guidelines[2], EOSC PID Policy[3]) or other long used accession numbergigs for genetic sequences used by the INSDC consortium. For individuals and organisations, ORCID and/or ROR identifiers will be pursued and integrated in all references. The survey results two main

[2] https://cetafidentifiers.biowikifarm.net/wiki/
[3] https://zenodo.org/record/3780423

Funded by the European Union

types of identifiers in association with the categories of data (object types) mentioned above: PIDs and resolvable internal IDs (Internal at project and consortium level).

The implementation of Third-Party Projects (3PPs) may include additional sources and mechanisms for collecting data. Should that be the case, these additional approaches will be included in future iterations of the DMP.

2.3 Data storage

Currently, the following formats, among others, are used for the preservation of scientific data in (depending on the task) different repositories. The most common **formats** in which the data will be preserved and can be exported are CSV and XML, followed by RDF and TSV. Additionally, certain Tasks will use highly specialised formats for their specific needs, such as fasta[4], PKL[5] or tiff[6]. For documentation purposes, most partners will opt for common file formats such as pdf or docx. Proprietary file formats should be preferred only in cases where these formats have clear and proven benefits over the use of open data formats or in cases where open data formats are not available. The project will aim to implement methods for uniform and sustainable storage and publication of data according to FAIR principles, which are developed within the framework of biodiversity informatics initiatives.

Figure 2 provides a visual breakdown of the formats expected to be used by partners for storing data. As a general recommendation, partners should choose data formats for final data deposition which are lightweight, non-proprietary and widely used (e.g. CSV, XML, JSON, RDF etc.). Moreover, partners should take all the necessary steps to protect their datasets and run regular backups. Given the nature and resources of the project the data management has to be a decentralised process, an inventory of data will be developed in order to have a better understanding of the data management needs. Through this inventory partners can structurally list the various data types per WP (digital images, expertise, taxonomy, services), including some other essential information (eg. parameter group, legal regulations (GDPR, INSPIRE,...), licences, repository for data dissemination (GenBank, ORCID, GBIF,...), storage location during the project, data formats, existing data or new data, 3PP data (Y/N), etc. Revisions and discussions regarding the management of the data will be held during the GA, as well as EB meetings upon request of the partners.

[4] Text-based format for representing nucleotide sequences or amino acid sequences, using single-letter codes.
[5] A file created by pickle, a Python module that enables objects to be serialised to files on disk and deserialized back into the program at runtime. It contains a byte stream that represents the objects.
[6] A computer file used to store raster graphics and image information.

Funded by the European Union

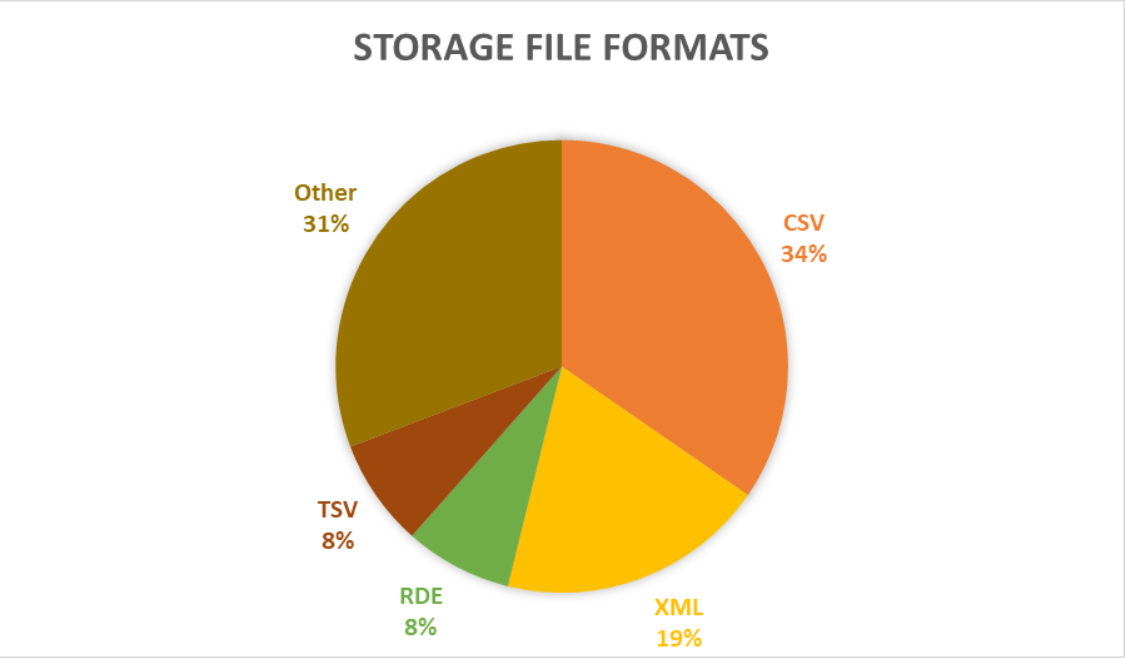

Figure 2: Formats expected to be used by partners for storing data.

As expected, the **size** of datasets handled by the different Partners will range from very small (several Megabytes) to substantially large (10-50 Gigabytes). Storage will be arranged and reported via the inventory by each task individually. It is however important to note that for many of the tasks, these numbers could still change significantly depending on various factors.

Most of the TETTRIs partners will use either their own infrastructure or web hosting (cloud) to store data, while a small number will employ a combination of both. Typically, server locations are within the EU (77%) and the non-EU servers used are largely GDPR-compliant.

2.4 Data Security

Under this section, we include security as well as sustainability of data. The data management structures for securely storing and managing data generated/processed under TETTRIs will be chosen according to the need of the particular Task. To name a few, the CETAF-DEST platform will be used for the training related data, repositories such as NCBI and EMBL for genetic data, GBIF for specimen data. Zenodo, GitHub and Wikibase will also be utilised as repositories of data when needed.

In terms of sustainable data management, TETTRIs aims to integrate its activities into existing data infrastructures as much as possible. Examples are the taxonomic information systems in WP2 (in particular EUNomen, Catalogue of Life and the GBIF Checklistbank), Citizen Science platforms such as "The Herbonauts'' and DoeDat in WP6, as well as trusted repositories for long-term presentation and curation (eg. GBIF and OBIS). The aim is to strengthen sustainable infrastructures in mutual interest and to avoid parallel development as far as possible.

Funded by the European Union

2.5 Personal data[7]

Personal data, such as contact details (e.g. name, surname, email address, location, affiliations, publications) will be collected in various Tasks of the project (at least 9 different Tasks will collect personal data according to the DMP survey conducted). Therefore, the necessary personal data protection measures will apply and EU GDPR principles will be observed (See Article 15.2 of the Grant Agreement). Personal data included in datasets will be stored occasionally, in these cases a privacy statement and consent form will be provided in advance with the support of WP10. When relevant, particularly in the publication of information the necessary amount of anonymisation will be ensured. Moreover, each institutional partner aiming to collect and store personal data will provide the necessary protection of the data and perform the administrator roles ensuring that personal data are not only secure, but also that they are not used for purposes unrelated to the TETTRIs project.

Expressed consent from individuals will be requested for the collection and use of their private personal data. Translated forms might be required in the case of 3PPs implemented in countries where English is not widely used. This will be especially important for actions in which locals are involved and full comprehension of content is instrumental (e.g in training courses).

Generally, all ethics-related data, including personal data falling under GDPR scope, will follow the principles for Ethics stated at European and, when applicable, also at national level. Specifically:

2. **HORIZON EUROPE Ethics principles** (Charter of EU Fundamental Rights)
3. **The General Data Protection Regulation (EU 2016/679)**, protecting citizens' privacy and increasing the responsibility when processing personal data

Moreover, under TETTRIs both, the direct Beneficiaries of the project as well as the approved Third-Party Projects (3PPs) leaders will have to comply with ethics requirements as stated in the **Ethics issues Table** (Link to table). This table has been prepared by the TETTRIs external Ethics Advisor and will ensure that all parties involved at any extent in TETTRIs comply with ethics when collecting, gathering, storing and sharing data.

2.6 Sensitive data[8]

There is no need foreseen for any sensitive data (in the GDPR meaning) to be gathered under the TETTRIs project. If this were to change, the appropriate handling of sensitive data is a responsibility of each institution and the research teams who gather the data. In terms of

[7] Personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data. Examples of personal data are name and surname; home address; email address such as name.surname@company.com; identification card number; location data (for example the location data function on a mobile phone); an Internet Protocol (IP) address; cookie ID*; the advertising identifier of your phone; data held by a hospital or doctor, which could be a symbol that uniquely identifies a person.

[8] Sensitive data is considered: personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs; trade-union membership; genetic data, biometric data processed solely to identify a human being; health-related data; data concerning a person's sex life or sexual orientation; data about protected species, objects of cultural heritage (e.g. Dinosaur specimens or historic, prehistoric or religious archeological sites and artefacts).

Funded by the European Union

3PPs, that responsibility remains with the leader of the approved projects in cases where those projects will include dealing with sensitive data.

2.7 Data Embargos

At this stage data is not expected to be subject to embargoes in the context of TETTRIs.

4. Third-Party Projects

As per the TETTRIs Grant Agreement, data and information gathered from the applicants submitting proposals under the 3PPs Call need to be securely stored and remain private for the sole use of their reviewing and assessment. To that end, access to that information will be restricted to members of the Community Implementation Board (CIB), the TETTRIs 3PP Administrator (Catalyze[9]), the Project Management Team (PMT), and the Project Coordinators. The latter will be responsible for the handling of the information and providing access, when necessary, to individuals from involved WPs for expert analysis. Nevertheless, once the projects have been awarded the details of the projects and the people involved will be made public (with their consent) for transparency purposes. A small summary will be made available of all the submitted proposals.

5. Data usage

Data produced by the project will be open by default and there will be no restriction to their access and re-use beyond the compliance of Creative Commons licence, data outputs should be CC-0 and publications CC-By (See Creative Commons[10]).

6. Accessible Data

TETTRIs follows the guidelines on open access to scientific publication and research data set by the Horizon Europe programme. The beneficiaries must ensure open access to peer-reviewed scientific publications, and trusted open access repositories for datasets, scripts, etc throughout and after the project's life. The same will apply to 3PPs results and publication of their results. TETTRIs results will be freely disseminated through appropriate channels including scientific publications, presentations at international conferences and workshops, whenever suitable. The publication venues will be primarily scientific highly regarded open access journals, such as the European Journal of Taxonomy (EJT), Biodiversity Data Journal (BDJ), Research Ideas and Outcomes (RIO), or others.

Presumably, all deliverables and other important project outputs that will not be published elsewhere will be published in a dedicated open access collection in RIO Journal, or archived in an open access repository (i.e. Zenodo). Furthermore, the software tools or plugins produced within TETTRIs will be available as open source code under an appropriate licence and published in the open science RIO Journal to ensure findability and

[9] An external company subcontracted via a tender to manage the administration of the Third-Party Projects.

[10] https://creativecommons.org/about/cclicenses/

reusability of all open source resources. Data publication should ensure a PID of the published dataset, citation mechanism, dissemination and usage tracking.

7. Final remarks

In the context of the TETTRIs project, it is important to differentiate between the data generated by TETTRIs partners and those produced under the 3PPs. This DMP aims to provide a framework for the governance of data while guiding the gathering, use and preservation of the data in accordance with EU regulations and following recognized recommendations in the domain throughout the project and after it has concluded. The same framework will act as a baseline for managing the data produced under the 3PPs, but those cannot yet be covered by this version of the document since the 3PPs will not be in operation until 2024. Specifically for those types of data that may be relevant, a specific clause will be inserted in the Agreements to be signed with successful 3PP applicants. In case any 3PPs include collection of certain types of data, this will need to be reported, together with the procedure envisaged to handle it, while keeping the entire responsibility on its management. Therefore, the next iteration of the DMP, due on Month 18 (May 2024), will be revised by the Project Management Team in light of the additions coming from the approved 3PPs.

Furthermore, as the project advances, more accurate and precise information will be available in terms of the data generated or used under the different Tasks in the project. If that is the case, the Task leader shall transfer that information into specific data management rules or protocols to be integrated in a new revised version of the DMP.

It is recommended that each TETTRIs partner also appoint a Data Protection Officer (DPO) to address personal data management and ethics-related topics. Their contact information will be then presented to the consortium members and the European Commission. Together with DPOs appointed by the TETTRIs partners, the PMT may decide to translate some of the important sections included in this DMP into TETTRIs Guidelines, to provide further advice and useful information on how to manage the data of the project.

APPENDIX 1

ID	Reference or Related Document	Source or Link/Location
1	Survey Results	LINK

Funded by the European Union