TROMPA

TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 8.4

Data Management Plan v3

Grant Agreement nr	770376
Project runtime	May 2018 - April 2021
Document Reference	TR-D8.4-Data Management Plan v3
Work Package	WP8 - Project Coordination
Deliverable Type	ORDP
Dissemination Level	PU
Document due date	30 April 2021
Date of submission	10 May 2021
Leader	UPF
Contact Person	Aggelos Gkiokas (aggelos.gkiokas@upf.edu)
Authors	Aggelos Gkiokas (UPF), Alastair Porter (UPF)
Reviewers	David Weigl (MDW)

Executive Summary

This document is the 3rd and final version of the deliverable D8.4 - Data Management Plan. The aim of this deliverable is to provide information on how the FAIR (Findable, Accessible, Interoperable, Reusable) data requirement is satisfied in the data created during TROMPA.

This deliverable contains the Data Management Plan (D8.4), a deliverable that belongs to WP8 of the project, devoted to project coordination. This document contains its 3rd version (M36). We describe here the procedure used in the project to handle the data collected and generated during the project, following the Horizon 2020 online manual.

We start by summarizing the main characteristics of TROMPA data, having in mind the main goal of TROMPA, indicated in its acronym, i.e. to enrich existing online music public-domain archives. We first specify the considered music data types (audio, music scores, text and images), formats (with an emphasis on open and standard formats) and sources (external sources, data from consortium and associated partners, data generated by TROMPA technologies in WP3, crowd contributions in WP4 and additional relevant sources). We then specify the target criteria for size, reusability and the link of data with the different TROMPA use cases and user communities. This allows us to define a list of existing open repositories our project will link and contribute to.

The data generated during the TROMPA project will fulfil the FAIR standard: it will be Findable, Accessible, Interoperable and Reusable. We describe in this deliverable how the TROMPA consortium will work to fulfil this FAIR criteria, providing specific details and indications for discoverability, identifiability, naming, versioning, availability as open data, documentation, and re-usability. We also discuss licensing, software tools, quality assurance and conditions to preserve it in the future. In addition, we consider the required allocation of resources for managing that, estimating the cost, resources, responsibilities and potential value for long term preservation. Moreover, we discuss data security and the ethical aspects linked to data.

This methodology is for the moment wide and comprehensive in order to be refined during the project according to the precise definition and evolution of the use-cases and the work carried out in the different work packages.

Version Log

#	Date	Description
v1.1	10 April 2021	Initial version circulated to consortium
v1.2	21 April 2021	Consortium contributions added
v1.3	27 April 2021	Version sent for review
v1.4	3 May 2021	Review comments integrated
v2.0	10 May 2021	Final version submitted to EU

Table of Contents

1. Introduction

This deliverable is the 3rd version of the Data Management Plan (DMP) for TROMPA, and refers to the end (M36) of the project. This document belongs to WP8 (Project coordination) and outlines how the data collected and generated within TROMPA is handled during and after the end of the project. For writing this deliverable we followed the templates suggested by the Horizon 2020 online manual. For making the contents of this deliverable more accessible, it is written in the form of questions and answers. This deliverable is considered an extension of previous versions, thus the contents of this version can be identical to the previous versions (1st:M6 and 2nd:M18) in some cases. For reasons of clarity and completeness, we preferred this approach instead of providing only the differences with the previous version.

2. Data Summary

State the purpose of the data collection/generation

The main purpose of data collection and generation in our project is to enrich current public-domain classical music archives. Data is automatically generated and enriched by technologies developed in WP3 and by user communities (WP4) at various music skill levels (from music scholars to enthusiasts) by participating in the use cases and in hybrid annotation workflows. The participation in these activities involves the enrichment of existing data, creation of new data (e.g. creating music performances, generating digital scores) as well as linking between data (e.g. alignment of music scores to audio).

Explain the relation to the objectives of the project

TROMPA's main goal is to enrich public domain music archives, with a special focus on classical music. As a consequence, data collection, generation and curation that are described in this Data Management Plan is a key aspect of TROMPA and is strongly related to the objectives of the project, as defined in the DoA.

- **Objective O1. To enable (semi-)automated processing mechanisms to be effectively and adequately applied to digital and online public-domain classical music resources.** The collection of annotation data from human experts during TROMPA is important for the evaluation and improvement of technologies for automatic music description and processing (e.g. train models, develop new methods). The development of novel techniques on WP3 and WP4 allows us to automatically describe resources available in public-domain music archives, and the consequent creation of open datasets is expected to have a strong impact on the music information research community.
- **Objective O2. To establish mechanisms enabling enrichment and data quality improvement from multiple perspectives, and considering different facets of the music material**. The data collected by involving the crowd in semi-automated annotation procedures is used to enrich the TROMPA-related archives and improve the quality of the metadata for various music facets.
- **Objective O3. To make the derived knowledge and obtained enrichments sustainable and universally useful and adoptable in end-user applications.** The procedures described in this

Data Management Plan ensure the curation, sustainability,openness and usefulness of TROMPA contributions to target repositories

- **Objective O4. To demonstrate how derived knowledge and enrichments can be practically, sustainably and engagingly exploited in real-world use cases.** The collected data (crowd annotations), the data created from algorithms (automatic annotations), the data contained within the Contributor Environment and other data such as interlinks between objects (e.g. alignment of a music score to a performance recording) are used to serve the five use-cases of TROMPA.

Specify the types and formats of data generated/collected

TROMPA uses standard data formats for the various data types (audio, music scores, images, performance metadata) considered. The selected format should satisfy the following criteria:

- **Preserve data quality:** Existing data used in TROMPA as data in public-domain repositories as well as the crowd contributions and automatic annotations are stored in the formats and under the license that their authors originally provided.
- **Meet scientific standards:** Data and metadata are stored in standard formats widely accepted by the scientific community. For instance, music scores are stored in the Music Encoding Initiative (MEI) format. The output of automated tools use standard formats such as JSON or YAML. Items in the Contributor Environment are represented using standard semantic web data models and are made available via Semantic Web formats such as JSON-LD.
- **Accessible to the general public:** Data is stored in formats that can be accessed by free and open source software.

Moreover, TROMPA creates metadata for cross-modal linking between files (e.g. alignment of a score to an audio) as well as segmentation of existing files to smaller fragments based on certain attributes (e.g. music structure) or based on annotation tasks (microtasking). Detailed information about the data formats for each data type are presented in Table 2.1.

Data type	Formats for sharing, processing and preservation
Audio Files	MP3, WAV
Scanned Images	JPEG, TIFF
Digital Music Scores	MEI, MusicXML
Metadata	Structured metadata in a Neo4J graph database represented using schema.org, the Dublin Core vocabularies, and other widely used semantic Web ontologies. Structured query access using GraphQL. Linked Data (JSON-LD) export via an HTTP REST wrapper.
Performance Data, extracted features	Match files, MIDI, performance parameters (extracted from audio or symbolic data) expressed as RDF / JSON-LD (e.g. using Timeline, Segment, and Music Ontologies), JSON, JAMS
Scientific Papers	PDF, stored in open e-document repositories (Arxiv, zenodo) and institutional repositories.

Working Documents	pdf, odt, doc, docx, Google docs[1]
Annotations	Annotations represented within neo4j and Solid using the W3C Web Annotation Data Model, exported as RDF / JSON-LD
Text (lyrics, user comments, interviews etc)	These can be stored in either txt files, or included as metadata field in other formats (MEI, XML)

Table 2.1. Data types and appropriate file formats for re-using, preservation and processing.

Specify the origin of the data

The origin of data that will be used in TROMPA can be summarized as follows:

- **Data collected from external sources:** This data is collected from external resources during the project and contains audiovisual and audio recordings, images, metadata, annotations, scanned and digital scores, ontologies, user activity tracking data such expertise tracking and questionnaire data from participants.
- **Existing data from TROMPA consortium and associated partners:** This data consists of existing data from public domain archives, TROMPA partners (e.g. CDR, RCO) and associated partners (e.g. IMSLP) including audio recordings, music scores (scanned images), digital score encodings, etc.; derived images (for alignment purposes) will be stored subject to permission (if this is not obtained, modern renderings from the derived encodings will be used instead).
- **Data generated in parallel to TROMPA:** In parallel to TROMPA, several consortium members are involved in additional projects yielding data relevant to the TROMPA agenda. While this data is not formally TROMPA data, it is acquired in close connection to TROMPA's agenda and interests, and made available such that the TROMPA consortium (and the research community at large) can benefit.
- **Data generated within TROMPA:** This data was being generated during the project and contains various data types such as numerical features extracted from audio signals, datasets of annotated audio content, performance metadata including alignments and symbolic representations, cross-modal music information, documents created during the project (deliverables, reports, etc) as well as the source code of the programs created. We can summarize this data as:
  - **Automatically generated data:** This type of data is generated automatically (mostly in WP3) and can be music descriptors, such as numerical data (audio features, statistical model parameters), text descriptors (tags), synthesized audio, and alignments of music resources.
  - **User-provided contributions:** This type of data is collected during the use case scenarios. They vary from simple tasks such as single labels describing whole music pieces (e.g. Emotion annotations), up to more complex tasks such as conversion of scanned score images to MEI scores.
  - **Music performance data:** Performance data (e.g. audio recordings in choir singers use case, MIDI performance data streams in instrument players use case) was collected during performances by TROMPA users. This performance data can also be

[1] Google docs are only used for development. All public facing documents are converted in pdf.

contributed to the public archives if the musicians grant their consent to their publication.

- **Documents, datasets, source code and scientific publications**: Technical reports, scientific publications, deliverables, source code of the software developed during TROMPA and datasets that are released for scientific research[2].

Specify if existing data is being re-used (if any)

Existing data will be used for various purposes, such as:

- **Training and evaluation of algorithms:** Several methods related to music audio processing are developed in WP3 for the semi-automated crowdsourcing procedures of TROMPA. Most of these methods need to be trained on labeled (annotated) data, and TROMPA's existing data is used for this purpose. Apart from TROMPA data, other scientific data (e.g. published datasets) will be used for the same purpose.
- **Use cases:** Existing data can be used to serve the five TROMPA use cases as defined in WP6. The selection of this data is made corresponding to each individual use case, and is described in more detail in **Deliverable 3.1 - Data Resource Preparation v2**[3](M18)

A list of the existing data resources that are associated with TROMPA is presented in Table 2.2 (as in all previous versions of this deliverable) and is refined in this last version of the deliverable. The reader should note that these resources are of any of the origins mentioned in the previous section (e.g. external resources, data from associated partners) and they are not necessarily owned by the TROMPA consortium.

Repository	Volume and type of Data
IMSLP[4] Petrucci Music Library	~124,000 works, represented by ~405,000 PDF scores and ~47,000 audio recordings.
Choral Public Domain Library (CPDL)[5]	~10,000 different works in PDF, other music encoding formats and MIDI for choirs.
MuseScore[6]	~3,000,000 MuseScore-encoded scores for personal use, ~300,000 to share
EMO[7] - Early Music Online	About 32,000 page-images duplicated in various formats: b/w and grayscale TIFF, images segmented by system. Library catalogue metadata (in XML). Derived musical encodings in MEI.

[2] https://zenodo.org/communities/trompa/?page=1&size=20

[3] https://trompamusic.eu/deliverables/TR-D3.1-Data_Resource_Preparation_v2.pdf

[4] https://imslp.org/

[5] http://www.cpdl.org/wiki/

[6] https://musescore.com/

[7] https://www.royalholloway.ac.uk/research-and-teaching/departments-and-schools/music/research/research-projects-and-centres/early-music-online/

MusicBrainz[8]	Metadata for recorded music, 1.4m artists, 2m releases, 20m recordings[9], Accessible via webservice (XML, JSON) or as a Database Archive
Kunst der Fuge[10]	19,300 MIDIs
Humdrum-data Repository[11]	Over 4,100 scores encoded in Humdrum (kern) format
CDR Muziekweb catalogue[12]	Structured metadata for 600,000 music CD's, 300,000 vinyl LP's, 20,000 music DVD's, 500 cylinder recordings and more. Over 7.5 million digitised audio files in FLAC and more than 100,000 video files in MP4.
The Vienna 4x22 Piano Corpus[13]	4 pieces performed by 22 professional pianists (MIDI, audio, alignment metadata in *match* format)
Researcher-in-Residence project of Cynthia Liem at the National Library of The Netherlands[14]	Public enrichment links between the CDR Muziekweb catalogue and the Delpher historical newspaper corpus from the, including research code to be released under the GNU GPLv3 license.
BDH[15] - Biblioteca Digital Hispánica	Collection includes c100 digital facsimiles (?c.5,000 pages) of early printed music sources suitable for Optical Music Recognition (OMR) and choral singing. High-quality PDFs easily converted to TIFFs for OMR.
TLdV[16] - Tomás Luis de Victoria	About 2,000 16c Spanish vocal/choral works by Victoria, Morales, Guerrero, Vásquez and others in digital encodings (Lilypond), compiled privately by Nancho Álvarez, which can be converted to MusicXML/MEI; individual voice-parts in MIDI format. (For many works, links to video/audio performances are provided as well.)
Repositories for development purpose	Repositories that can be used for the development of the individual technologies on WP3 (training/evaluating models etc).

Table 2.2. Existing repositories considered for TROMPA.

<u>State the expected size of the data (if known)</u>

As of the end of TROMPA, we have the following disk usage requirements:

[8] https://musicbrainz.org/

[9] https://musicbrainz.org/statistics

[10] http://kunstderfuge.com/

[11] https://github.com/humdrum-tools/humdrum-data

[12] https://www.muziekweb.nl/

[13] https://repo.mdw.ac.at/projects/IWK/the_vienna_4x22_piano_corpus/index.html

[14] http://lab.kb.nl/news/introducing-kb-researchers-residence-2018

[15] http://bdh.bne.es/bnesearch/AdvancedSearch.do?showAdvanced=true

[16] https://www.uma.es/victoria/

- CE Metadata storage (Hosted by VD on Amazon AWS): 100MB
- TPL data storage (Hosted by UPF on institutional infrastructure): < 1GB
- Solid server data storage (Hosted by UPF on institutional infrastructure): < 1GB
- CPDL repertoire audio data storage (Hosted by VL on AWS): 25GB

These disk requirements are not fixed at the end of the project and may be increased, as many services remain online

Outline the data utility: to whom will it be useful

TROMPA's general goal is the enrichment of public domain classical music libraries, by the incentivisation of the crowd in five target music communities, for each of the five communities TROMPA targets a use case. These target audiences can be considered as the primary TROMPA data users:

- **Music Scholars.** Musicologists can access TROMPA data repositories to support musicological research studies by providing ways to efficiently search and analyse musical data and linked resources across different collections and modalities. Through TROMPA's scholar-facing prototype, they are also able to participate in scholarly dissemination and communication through score annotation.
- **Content Owners**. Content owners such as orchestras are an important user category for digital music resources. Digitization of orchestral scores and having them available free of charge will help orchestras survive and makes it easier to share their performances through recordings.
- **Instrument Players.** Instrument players can benefit from TROMPA data public archives that offer ways to explore music scores and corresponding performances. Performance data will be published (given suitable licensing / permissions by performer) enabling both the tracking and analysis of one's own performance characteristics over time, and pedagogical advantages (a piano teacher can get insights on their students' performance characteristics).
- **Choir Singers.** The choir singer use case will interactive feedback mechanisms surrounding rehearsals allow choir singers to practice. Similar to instrument players, performance data will be contributed.
- **Music Enthusiasts.** The music enthusiasts' use case targets to people without formal musical education, who are interested in learning more about music. In particular, since emotion is one of the main reasons why people engage with music, the use case exploits this interest in order to teach enthusiasts about the relationship between musical properties and emotions. Music enthusiasts are able to access most of the TROMPA associated data for the use case purpose.

Apart from the target audiences related to the use cases, data collected and generated in the TROMPA project data will be used by:

- **The General Public.** TROMPA data is published using open licenses where possible (pending copyright, privacy, ethical and related issues) so that it can be accessed, re-used, reproduced, re-interpreted and remixed by anyone.
- **The Scientific Community.** Research publications, open source software and datasets created during the TROMPA are accessible by scientific communities from various disciplines (musicology, computer science, music information retrieval).

3. FAIR Data

TROMPA is a member of the Open Research Data Pilot of the European Commission, which enables open access and reuse of research data generated by Horizon 2020 projects. Therefore, the data created during the TROMPA project should be Findable, Accessible, Interoperable and Reusable (FAIR) [1]. All four aspects of the FAIR data requirement will be discussed in this section. Most of the data created and curated by TROMPA is accessed via the Contributor Environment (CE). The role of CE regarding the data storage will not be to store the TROMPA data, but to store and maintain metadata of, as well as interlinks and references to the data. A detailed description of the CE and the data infrastructure of TROMPA can be found in **Deliverable 5.1 - Data Infrastructure**[17]. A journal paper providing a full overview of TROMPA's approach to FAIR data has been accepted for publication [2].

3.1 Making data findable, including provisions for metadata:

Outline the discoverability of data (metadata provision)

As stated in **Deliverable 5.1 Data Infrastructure**, the Contributor Environment's main access interface is GraphQL[18] together with a Neo4j[19] database, using Dublin Core[20] as a Metadata standard. This combination ensures performant discoverability of the public-domain data stored or referred-to in the Contributor Environment. User contributions, stored privately within each user's corresponding Solid Pod, are made discoverable using the same environment where users elect to publish their contributions to the Contributor Environment, using the publication mechanism described in [3].

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

Internally, all entities in the Contributor Environment in WP5 are uniquely identified by a UUID. Although represented internally as neo4j database nodes, each entity can also be accessed externally as an RDF (JSON-LD) representation through a unique URI mapped to each UUID. Additionally, items in the Contributor Environment that were created as the result of importing metadata from an external resource will have a link (URI) to an external source of the item being described. If we generate new data using this source data we ensure that it is also available with these identifiers, and keep semantic relationships between data by using existing identification schemes. For instance, if we generate annotations for an audio piece that has a MusicBrainz identifier, the derived annotations will be also available with this identifier. Moreover, both the audio piece and the derived annotations may be linked to the other identifiers as wikidata identifiers for referring to a music composer of the piece. Scientific publications and datasets are hosted on institutional repositories or deposited to Zenodo[21] and thus have a DOI assigned to them.

Outline naming conventions used

[17] https://trompamusic.eu/deliverables/TR-D5.1-Data_Infrastructure_v2.pdf

[18] https://graphql.org/

[19] https://neo4j.com/

[20] https://www.dublincore.org/specifications/dublin-core/dces/

[21] https://zenodo.org/communities/trompa/

We used standardized naming conventions for all the data that were created during the project. Different types of data have different naming conventions. For example for published documents such as the deliverables and the respective review documents, the reference numbers have standardized formats (see **Deliverable 8.1 - Project Handbook**, Section 10.2). For data related to an existing identifier is accessible via that identifier in an API or in its filename. For user generated data such annotations these are anonymized and the reference contains the user identification, e.g. annotation could be stored in a file with naming convention as userID_pieceID_annotationID.json.

Outline the approach towards search keywords

Regarding documents such as deliverables and reports, scientific publications, project website content and social media, all TROMPA partners have consistency in the way that they refer to the projects and components with certain keywords. All dissemination material has a certain reference to keywords related to TROMPA.

Outline the approach for clear versioning

The majority of the code written in the project (including the Contributor Environment, demonstrators, and other components) is versioned and publicly available in Github[22]. We used semantic versioning[23] project-wide and gitflow[24] standards where possible. A comprehensive list of the software used is reported in **Deliverable 8.3 - Sustainability Model**[25].

Regarding public data there is the policy to not ingest it into local storage, but rather refer to it at its native location by reference to URIs. As stated in before, all CE nodes have a local identifier (UUID). These identifiers are turned into external (URI) identifiers and provenance traces are captured, where nodes represent entities derived from other entities.

Data published in academic repositories such as Zenodo include a DOI, and subsequently released datasets will have incrementing version numbers.

Regarding annotations, depending on the use case, we keep versioning of the annotations when needed. For example, in the orchestra's use case, where the main task is to generate MEI scores from images based on microtasking, different versions of a score are stored on Github, throughout the crowdsourcing pipeline.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how.

For the Contributor Environment internal model, we selected to apply default metadata properties to all entities (nodes). These default metadata fields correspond to the schema.org and SKOS vocabularies, the Dublin Core metadata standard, and the PROV ontology, and are mapped to RDF (JSON-LD) representations via an HTTP / REST wrapper layer. Annotation metadata exposed through the TROMPA project are published using W3C recommended standards (Web Annotation Data Model) and are also made available as JSON-LD from the CE. Musical score representations generated by the project are described using representation-specific metadata standards (e.g. MEI

[22] https://github.com/trompamusic
[23] https://semver.org/
[24] https://datasift.github.io/gitflow/IntroducingGitFlow.html
[25] This deliverable is available to the consortium only

responsibility statements) and may be further described by RDF descriptions targeting these resources.

3.2 Making data openly accessible

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.

In principle, all non-personal data created in TROMPA are openly accessible. These data can contain:

- Crowd annotations (under the restriction that they are anonymous) and other metadata related to the use cases.
- Metadata that is already open and is owned by the TROMPA partners (as for example CDR metadata) will remain open.
- Outputs of algorithms such as music descriptors or synthetic voices (choirs singing use case)
- Outcomes of crowdsourcing procedures (i.e. produced digital scores from orchestras use case)
- The linking between objects from different resources will be made available.
- The scientific publications and datasets are all uploaded to Zenodo and thus are open (see next section).
- Source code of the project is considered as data and is publicly available on github.

Exceptions for keeping data closed are the following:

- If certain data can be used to trace a person or it is rights restricted, it is not openly accessible. All TROMPA partners follow the ethics guidelines of their ethics committees (or the associated ones - see **Deliverable 1.1 - H Requirement No.1**[26])
- Content that we do not own the copyright such as audio recordings, album artwork or video files. This can only be made available under specific licenses.
- Data contributed to Solid Pods (e.g., music performances; scholarly annotations) are by default closed unless consent is provided for publication.

Specify how the data will be made available

The Contributor Environment data is accessible directly through a web API by means of 4 React components developed under WP5 (Tasks 5.2-5.5). These components support web based user interfaces, including the five TROMPA pilot applications, by exposing predefined functionalities like a semantic search interface or annotation tooling, to consume and enrich the TROMPA dataset. The web API exposes the entire TROMPA dataset and all functionalities, but limits potential destructive or corruptive functionalities, or privacy sensitive data only to users granted with adequate authorization. Details on how to access the CE data is provided in Deliverable D2.3 - Technical Requirements and Integration (Section 5).

All datasets are available on Zenodo, and the source code on open access repositories (see section “Outline the approach for clear versioning”). We have created a TROMPA community on Zenodo[27].

[26] This deliverable is confidential to the consortium only

[27] https://zenodo.org/communities/trompa/?page=1&size=20

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

The Contributor Environment is accessible through http(s) protocol via RESTful and GraphQL interfaces. API docs are available for the REST interface, while the GraphQL has discoverability implemented within the interface specs. One of our selection criteria for selecting external repositories to make data available is if the repository also includes an API in order to programatically access the data. Regarding scientific publications and datasets, we use conventional data formats that require open licence software to be accessed. Details regarding guidelines to store and access data in the CE are given in **Deliverable 2.3 - Technical Requirements and Integration**[28].

Specify where the data and associated metadata, documentation and code are deposited

As a design principle, only metadata is stored within the Contributor Environment. Data from public repositories will be referred to (by URI) only, and remain where it is. User-provided data is stored privately within the user's Solid Pod, or published to the CE on request. Produced data is stored in principle by participants and made available through an URL or documented API. For instance, results of algorithms run through the TROMPA Processing Library are stored in the CE, in an S3 server maintained at the MTG, or in a Solid Pod. Other partners have other storage locations (e.g. VL stores synthetic voices in their own Amazon S3).

Source code is deposited in Github[29]. Publications hosted on institutional repositories. Documents produced by project partners such as deliverables will be deposited to project website[30] or other open domain repositories and linked through the TROMPA website.

Specify how access will be provided in case there are any restrictions

The CE has a data access layer which means that only authenticated users can submit data to it. Each partner in the consortium has an access key which allows them to authenticate to the CE. The CE has no restrictions for reading data. In the case of private data, a user provides a Solid Pod to store this data where only that user has access to their content, unless they change the permissions on the content to allow others to read it. Users have the ability to give permission to the TROMPA Processing Library to read and write private content while keeping it private to the public.

3.3 Making data interoperable

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

We reuse and build on widely used existing ontologies and data models, e.g. schema.org, Web Annotations, SKOS, and PROV-O. Every item stored and/or referenced in the CE has a unique TROMPA identifier with a corresponding URI exposed through a REST HTTP wrapper (see **Deliverable D2.3 - Technical Requirements and Integration**[31], section 5). Wherever applicable, we reference entities in external repositories by URI (e.g. MusicBrainz, Wikidata URIs).

[28] https://trompamusic.eu/deliverables/TR-D2.3-Technical_Requirements_and_Integration_v2.pdf

[29] For an extensive list see Deliverable 8.3 - Sustainability Model

[30] https://trompamusic.eu/

[31] https://trompamusic.eu/deliverables/TR-D2.3-Technical_Requirements_and_Integration_v2.pdf

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow interdisciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

The Contributor Environment facilitates and encourages the use of (multiple) ontology references for all datatypes (entities and relations) stored, supporting serialisation into RDF Linked Data via JSON-LD export, allowing interoperability (Deliverable 2.3, section 5). Where custom entities are required (e.g., in the case of specialised Web Annotation motivations), they are mapped to applicable standards using the SKOS vocabulary (e.g., skos:broader).

3.4 Increase data re-use (through clarifying licenses)

Specify how the data will be licenced to permit the widest reuse possible

We only add metadata to the CE which is already available under an open license. Items in the CE that link to external resources are available under the same license terms that the original metadata was made from. Data added to the CE by pilots (e.g. annotations) is typically made available under open data licenses such as CC-BY-4.0.

For crowd annotations we use open licences where possible. Regarding data created from partners outside the CE (e.g. datasets) the specific choice of licence will be given to each partner, but the project guidelines call for licenses that are as open as possible. Regarding source code, we suggest that all software is released under open source software licences where possible. As a consortium we recommend that the BSD 3-clause license or Apache 2.0 license are used, though partners can choose to use other similar open source licenses if required.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

The data will be available for re-use from their release date. In the general case, there will be no embargo period for this.

Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

See above.

Describe data quality assurance processes

There are several processes that are followed in order to ensure the quality of the data in many aspects. Regarding the metadata created in the CE, the data storage will be set up redundantly. We use file formats that preserve data quality as high as possible. Backup of the data is kept periodically.

Moreover, since the goal of TROMPA is to provide high quality metadata of existing music material by combining the power of algorithms and humans in annotation, TROMPA provides new high quality data (metadata, data interpretations) through the semi-automated annotation processes. There is a dedicated Work Package (WP4) whose active research is to ensure the quality of the results from the crowdsourcing.

The internal review process of the deliverables ensures the quality of these documents. Regarding research publications presenting the research outcomes of TROMPA, these are submitted in peer-review conferences/journals.

Finally, the TROMPA coordination team nominated a data expert member of the consortium, a Data Officer, whose duties include the monitoring of the data collection and processing.

Specify the length of time for which the data will remain re-usable

Metadata in the CE copied from other repositories will remain available in those source repositories as long as they stay online. VD will continue to host the CE for one year after the end of TROMPA (until 1 May 2022). At that point, we will archive the contents of the CE and make it available on Zenodo. VL will continue to host CPDL audio renderings for one year, after which they may continue to be available as part of the Cantamus platform depending on its commercial sustainability. We expect the sheet music from IMSLP that Peachnote has analyzed using OMR and aligned with audio performances, currently hosted by EGI[32] to be further used in research and integrated in commercial products and remain available as long as it is commercially sustainable. We expect that data stored in institutional repositories and Zenodo will continue to be available into the future, although by relying on these repositories we accept a small level of risk that they may become unavailable in the future[33] . Content stored at UPF (currently data storage and Solid Pods) have no short-term requirements to be removed. We expect to make this content available for at least the next 5 years.

4. Allocation of Resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

Regarding scientific publications, dataset releases, and source code, we use repositories that are free to use (see above, github, zenodo). Some of the metadata generated during the project will be contributed back to the original repositories (musicbrainz, wikipedia, CDR).

Clearly identify responsibilities for data management in your project

As mentioned above, the TROMPA coordination team nominated a data expert member of the consortium, a Data Officer, who is responsible for the data management in TROMPA. **David Weigl** from MDW, accepted this role. Within the CE, VD is responsible for managing the data of the CE. Regarding data contributed to existing repositories, the repository owners will be responsible for these data.

For scientific publications, datasets and source code, partners should follow the guidelines for uploading the data to appropriate repositories (Github, Zenodo, University open-repositories). Regarding data related to partner’s tasks, all partners should be responsible for generating and organising their data.

[32] https://documents.egi.eu/public/ShowDocument?docid=2886 - the SLA agreement between EGI and Peachnote.

[33] Zenodo indicates that they expect content to be available for at least the next 20 years: https://about.zenodo.org/policies/

Describe costs and potential value of long term preservation

Long term preservation is important in project outcomes related to Europe’s cultural heritage, as TROMPA. As specified in a previous section (*specify the length of time for which the data will remain re-usable*), the main TROMPA services and data will be hosted for one year after the end of the project.

To host public-facing web applications written in javascript, we use free hosting services such as github pages[34] and netlify[35]. While these free services currently have no restrictions that prevent us from hosting them, we are unable to guarantee that these services will remain free into the future. Some partners are opting to pay for commercial hosting services for software and data created during the project. For example, for the case of the synthesized CPDL singing scores, the server hosting and storage costs can be around 30 EUR a month, running on commercial services. Similarly, for running the CE VD foresees a cost of around 300 EUR per month.

5. Data Security

Address data recovery as well as secure storage and transfer of sensitive data

Some data that we collect may contain Personal Information, but we don’t collect any sensitive personal information. All TROMPA partners comply with relevant privacy regulations in ensuring that we only collect the minimum necessary personal information and explain to participants what we use it for. We have developed a data processing plan to ensure that personal information is only made available to researchers who require it. All partners that collect personal information have priorly got an ethics approval from their committees. The personal information we collect is:

- **Presencial Interviews**: One-to-one interviews with users. Depending on the use case these were choir conductors/singers, music enthusiasts or instrument players to gather opinion about the use case. These interviews were offline, meaning that no audio or video was recorded, only written notes will be taken.
- **Online Interviews:** Interview sessions through teleconferencing due to the COVID-19 pandemic. Some sessions were recorded for transcription purposes.
- **Workshops**: Several workshops have been organized that are related to the use cases. In these workshops we kept log files of user activities.
- **Questionnaires**: Online or printed questionnaires were given to the participants.
- **Performance data**: Data from rehearsing practice such as singers´ voices or instrument players performances and their explicit annotations, e.g. in terms of difficulty of the piece, perceived musical qualities, etc
- **Data related to the use case activities**: such as annotations (e.g. annotating induced emotion in the music enthusiasts use case) or music rehearsals (choir singers, instrument players use case).
- **Observations**: written notes, notes, pictures, audio, videos were taken in face to face environments.
- **Log files**: Log files collecting actions of pilot users.

[34] https://pages.github.com/

[35] https://www.netlify.com/

Regarding the backup of the data, we make backups of data that we generate. We can summarize the data protection mechanism as follows:

- **Data storage**: Data is stored in secure servers where only authorized people from the consortium can access to.
- **Data backup**: Data is frequently backed up in a secure storage server.
- **Data maintenance and quality**: TROMPA has assigned a member of the consortium as the Data Officer of the project. Data Officer is responsible for the data maintenance, curation and quality assurance.
- **Data anonymization**: All data collected is anonymized when possible. Participants will be given an identification number. We use standard naming conventions of the files that contain the recorded data. If needed, anonymized personal information such as age, gender or profession expertise may also be stored.
- **Data access**: Only authorized people have access to the data. Access to data is given only for research purposes.
- **Data publication**: For the purpose of data publication, i.e. scientific datasets, publications, some of the data collected was published. In the case the data was anonymized and released only if the participants provide informed consent.
- **Personal data**: No sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction) is gathered.

6. Ethical Aspects

Details about the data protection and anonymization are given in Section 5 - Data Security. All partners that collect human data during the use cases, had acquired an ethics approval from their ethics committees prior to the experiments. All participants provided informed consent before participating in an experiment. The ethical aspects regarding data collection and privacy are described in **Deliverables D1.1 - H Requirement No.1**, **D1.2 - H Requirement No.2**, **D1.3 - H Requirement No.3**. **I**n the deliverables **H Requirement No. 4** and **POPD Requirement No. 5** we submit the ethics approval from all partners and the Data Protection compliances. **POPD Requirement No.6** describes procedures for data collection, protection and destruction[36].

7. Conclusion

In this deliverable we presented the 3rd and final version of the DMP of the TROMPA project. TROMPA is by definition a project focused on open data, since it is dedicated to enrich public domain musical archives. TROMPA meets as much as possible all the FAIR requirements, by employing scientific standards for data representations, strict procedures for data maintenance, curation, and for data quality assurance. While being as open as possible, we carefully treat data protection and ethical issues. We publish public-domain data in the CE, thus being available to anyone, while we store private data into private storage (Solid Pods) under user control.

[36] All ethics related deliverables are confidential to the consortium only

8. References

8.1 Written references

[1] Wilkinson, M. D. et al (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3.

[2] Weigl, D. M, Crawford, T., Gkiokas, A., Goebl, W., Gómez, E., Gutiérrez, N. F., Liem, C. C. S., & Santos, P. (2021). FAIR Interconnection and Enrichment of Public-Domain Music Resources on the Web. Empirical Musicology Review. Accepted for publication.

[3]Weigl, D. M., Goebl, W., Hofmann, A., Crawford, T., Zubani, F., Liem, C. S., & Porter, A. (2020). Read/Write Digital Libraries for Musicology. In 7th International Conference on Digital Libraries for Musicology (pp. 48-52). ACM Digital Library.

8.2 List of abbreviations

Abbreviation	Description
DMP	Data Management Plan
DoA	Description of Action
CE	Contributor Environment
FAIR	Findable, Accessible, Interoperable, Reusable
UUID	Universal Unique Identifier
DOI	Digital Object Identifier
VoID	Vocabulary of Linked Datasets
API	Application Program Interface
UPF	University Pompeu Fabra
TUD	Technische Universiteit Delft
GOLD	Goldsmiths’ College
MDW	Universität für Musik und darstellende Kunst Wien
VD	Video Dock BV
PN	Peachnote GmbH

VL	Voctro Labs SL
RCO	Stichting Koninklijk Concertgebouworkest
CDR	Stichting Centrale Discotheek

Table 8.1. List of abbreviations