TROMPA

TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 8.4

Data Management Plan V1

Grant Agreement nr	770376
Project runtime	May 2018 - April 2021
Document Reference	TR-D8.4-Data Management Plan v1
Work Package	WP8 - Project Coordination
Deliverable Type	ORDP
Dissemination Level	PU
Document due date	31 October 2018
Date of submission	31 October 2018
Leader	UPF
Contact Person	Aggelos Gkiokas (aggelos.gkiokas@upf.edu)
Authors	Aggelos Gkiokas (UPF), Alastair Porter (UPF), Wim Klerkx (TUD)
Reviewers	Emilia Gómez (UPF), David Weigl (MDW)

Executive Summary

This document is the 1st 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 will be 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 first version (M6) and is designed as a living document that will be updated along the project, in M18 and M36 respectively. 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, video, 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 allow 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 about 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 about 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
v0.1	3 October 2018	Initial version circulated to consortium
v0.2	26 October 2018	Consortium contributions added
v0.3	29 October 2018	Comments from consortium added
v1.0	31 October 2018	Final version submitted to EU

Table of Contents

Table of Contents **4**

1. Introduction **5**

2. Data Summary **5**

3. FAIR Data **10**

3.1 Making data findable, including provisions for metadata: 10

3.2 Making data openly accessible 12

3.3 Making data interoperable 14

3.4 Increase data re-use (through clarifying licenses) 14

4. Allocation of Resources **15**

5. Data Security **16**

6. Ethical Aspects **17**

7. Conclusion **17**

8. References **17**

8.1 Written references 17

8.2 List of abbreviations 18

1. Introduction

This deliverable is the first version of the Data Management Plan (DMP) for TROMPA, and refers to M6 of the project. This document belongs to WP8 (Project coordination) and outlines how the data collected and generated within TROMPA will be handled during and after the end of the project. For writing this deliverable we followed the templates suggested by the Horizon 2020 online manual[1]. For making the contents of this deliverable more accessible, it is written in the form of questions and answers. This is a living document that will be submitted for review in M18 and M36. Thus, some of the questions are partially addressed or not detailed in the current version as some more specifications will derived from the concrete development of the use-cases and work in the WPs .

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 will automatically generated by technologies developed in WP3 and enriched by user communities (WP4) at various music skill levels (from music scholar to enthusiasts) by participating in semi-automated crowdsourcing annotation activities. These crowdsourcing annotation will involve the enrichment of existing data, creation of new data (e.g. performances, 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 will allow to automatically describe current music archives, and the creation of open datasets will also have a strong impact in the 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 will be 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

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Data Management Plan will ensure the curation, sustainability and 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 (crown annotations), the data created from algorithms (automatic annotations) and other data such interlinks between objects (e.g. alignment of a music score to an audio) will be used to serve the five use-cases of TROMPA.

Specify the types and formats of data generated/collected

TROMPA will use standard data formats for the various data types (audio, video, music scores, images). The selected format should satisfy the following criteria:

- **Preserve data quality:** Existing data as well as the crowd contributions and automatic annotations will be stored in the formats and under the license that their authors originally provided.
- **Meet scientific standards:** Data and metadata will be stored in standard formats widely accepted by the scientific community. For instance annotations will be stored using standard formats such as JSON, YAML or JAMS[2], and Semantic Web formats such as RDF and JSON-LD, music scores in Music Encoding Initiative (MEI) .
- **Space requirements:** whenever possible, data will be represented in lossless file formats (such as FLAC for audio) in order to have an archival copy of data that we collect and generate. This will increase space requirements but will ensure the integrity of our data.
- **Accessible to the general public:** Data will be stored in formats that can be accessed by free licence software.

Moreover, TROMPA will create 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	FLAC, MP3, AAC , WAV
Video	MP4
Scanned Images	JPEG, TIFF, IIIF
Digital Music Scores	MEI, MusicXML, TabCode
Metadata	Structured metadata in PICA3 standard. Dublin Core as Metadata standard for all entities stored in Contributor Environment textual/tabular library catalogue metadata (MARC or XML)
Performance Data, extracted features	Match files, MIDI, performance parameters (extracted from audio or symbolic data) expressed as RDF (e.g. using Audio Feature Ontology)

[2] https://github.com/marl/jams

Scientific Papers	PDF, stored in open e-document repositories (Arxiv, zenodo) and institutional repositories.
Working Documents	Google docs, pdf, odt, doc, docx
Score annotations	RDF (Web Annotations); MEI
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:** These data will be collected from external resources during the project and will contain 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, video, music scores (scanned images), digital scores, etc. Further image data beyond that already available for early-music Optical Music Recognition (OMR) will be acquired where possible via IIIF but not retained in highest quality after processing; 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 to be 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 will be 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 to be generated within TROMPA:** This data will be generated during the project and will contain various data types such as numerical features extracted from audio signals, datasets of annotated audio content, metadata of audio content such as 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 that will be created . We can summarize these data as:
  - **Automatic generated data:** This type of data is generated automatically (mostly in WP3) and can be music descriptors, such as numerical data (audio features, statistical models parameters), text descriptors (tags), synthesized audio, and alignments of music resources.
  - **Crowd contributions:** This type of data will be collected during the use case scenarios. They can vary from simple forms such as labeling of a whole music piece, up to more complex tasks such as transcription or music performance evaluation.

- ➢ **Music performance data:** Performance data (e.g. audiovisual recordings, sensor data streams) will be collected during performances by TROMPA users (instrument players or choir singers). This performance data will also be contributed to the public archives if the musicians grant their consent to their publication.
- ➢ **Documents, datasets and scientific publications**: Technical reports, scientific publications, deliverables and datasets that will be released for scientific research.

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

Existing data will be used for various purposes as:

- ❖ **Training and evaluation of algorithms:** Several methods related to music audio processing will 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 will be used for this purpose. Apart 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 will depend upon the cases, and will be described in more detail in Deliverable 3.1 - Data Resource Preparation (Month 10)

A preliminary list of the existing data resources that are associated with TROMPA is presented in Table 2.2 and will be refined during the project. 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 necessarily not owned by the TROMPA consortium.

Repository	Volume of Data
IMSLP Petrucci Music Library	~124,000 works, represented by ~405,000 PDF scores and ~47,000 audio recordings.
Choral Public Domain Library (CPDL)	~10,000 different works in PDF, other music encoding formats and MIDI for choirs.
Europeana Music	~319,714 music audio recordings, scanned scores and other music items
MuseScore	~3,000,000 MuseScore-encoded scores for personal use, ~300,000 to share
Répertoire International des Sources Musicales (RISM)	~1,000,000 items information 'about' sources (manuscripts, location, etc) ~30,000 composers
ECOLM - An electronic corpus of Lute music	About 2000 page-images duplicated in various formats: b/w TIFF, derived coloured TIFF. Basic metadata concerning musical contents. Derived musical encodings (from OMR) in MEI.
EMO - 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.
AcousticBrainz	Automatically extracted features for 10 million music recordings (of all types of recorded music), JSON, 400GB
MusicBrainz	Metadata for recorded music, 1.4m artists, 2m releases, 20m recordings (https://musicbrainz.org/statistics), Accessible via webservice (XML, JSON) or as a Database Archive
Kunst der Fuge	19,300 MIDIs
CDR Muziekweb catalogue	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	4 pieces performed by 22 professional pianists
NWO-KIEM project no. 314-98-122	Music rehearsal audio from. Audio considering works in the public domain, for which musicians gave explicit permission for research reuse, will be shared as research data beyond this project, also to the TROMPA consortium.
Researcher-in-Residence project of Cynthia Liem at the National Library of The Netherlands	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.
Other	Datasets that will be used for training and development of algorithms. These will be defined in later stages of the project.

Table 2.2. Existing TROMPA repositories.

State the expected size of the data (if known)

In this stage of the project the expected data size is not known.

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 will target 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.

- **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. By sharing performance annotations orchestras can benefit from other orchestras' experience and insights.
- **Instrument Players.** Instrument players can benefit from TROMPA data public archives that will offer ways to explore music scores and corresponding performances. TROMPA's annotations will offer assistance in choosing music to play that is appropriate to their level. Moreover performance data will be contributed to the public archives (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 will target to people without formal musical education, who are interested in learning more about music. They will 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:

- **General Public.** TROMPA data will be 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.
- **Scientific Community.** Research publications, open source software and datasets created during the TROMPA will be 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) (Wilkinson et al, 2016). All four aspects of the FAIR data requirement will be discussed in this section. Most of the data created and curated by TROMPA will be 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**.

3.1 Making data findable, including provisions for metadata:

Outline the discoverability of data (metadata provision)

As proposed in **Deliverable 5.1 Data Infrastructure**, the Contributor Environment's main access interface is GraphQL. Together with the intention to use Neo4j as the main database, and to use Dublin Core as a Metadata standard, this combination ensures performant discoverability of data

stored or referred-to in the Contributor Environment. Moreover, care will be taken for the discoverability of the data with respect to the language to be discovered. By employing translation mechanisms by CDR and information in MusicBrainz, natural language descriptions of musical works will at least be enabled in English, French, German and Dutch.

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, the Contributor Environment in WP5 will allow all entities and relations (nodes and edges) to be uniquely identified by UUID. If needed, the DOI identifier can be incorporated as an additional property on all nodes of Object class. Where possible we will reuse identifiers that are used by the primary sources of data that we import. If we generate new data using this source data we will ensure that it is also available with these identifiers, and will keep semantic relationships between data by using existing identification schemes. For instance, if we generate annotations for an audio piece that has a MusicBrainz identifiers, 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 as wikidata identifiers for referring to a music composer of the piece. Regarding scientific publications and datasets, will be deposited to Zenodo and thus will be referred by their DOI.

Outline naming conventions used

We will use standardized naming conventions for all the data that will be created during the project. Different types of data will 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 will be accessible via that identifier in an API or in its filename. For user generated data such annotations these will be anonymized and the the reference will contain the user identification, e.g. annotation could be stored in a file with naming convention as userID_pieceID_annotationID.json. A detailed description of the naming conventions will be provided in next versions of this deliverable.

Outline the approach towards search keywords

Regarding documents such as deliverables and reports, scientific publications, project website content and social media, all TROMPA partners will have consistency in the way that we refer to the projects and components with certain keywords. All dissemination material will have a certain reference to keywords related TROMPA. Regarding the release of data related to the music repositories, these data will be associated with keywords related to the content such as the composer and the name of the piece, as well as keywords derived from annotations, for instance keywords such as “meter change”, “chord Am” etc.

Outline the approach for clear versioning

We expect that all code written in the project (including the Contributor Environment, demonstrators, and other components) will be git versioned and publicly available in source code repositories such as Bitbucket or Github. We intend to use semantic versioning project-wide and gitflow standards where possible.

Regarding public data there is the policy to not copy and/or change it, and leave it at the original location to be referred to. Regarding data generated within the CE, here are just too many types of data coming in to say something generic about it at this point. This will be defined in a later version of this deliverable.

Data published in academic repositories such as Zenodo will include a DOI, and subsequently released datasets will have incrementing version numbers. For data we store in other repositories, we will investigate if the CE can help with this. More details about this process will be given in the next versions of this deliverable.

Regarding Annotation and Annotator Profiles' versioning, each created profile will be timestamped. This will result in multiple profiles for an annotation or for an annotator, with the most recent profile to be the profile with the most recent metadata. Moreover we will track the provenance of evolving media resources (e.g. crowdsourced editions to MEI representations of musical score) using the PROV data model.

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, our preference is to apply default metadata properties to all entities (nodes). These default metadata fields would correspond to a metadata standard like DCMI and mapped to Semantic Web representations using corresponding ontologies. Annotation metadata exposed through the TROMPA project will be published using W3C recommended standards (Web Annotation Data Model). Musical score representations generated by the project will be described using representation-specific metadata standards (e.g. MEI responsibility statements, PICA-3 standard), and may be further described by RDF descriptions targeting these resources. Metadata describing Linked (RDF) Datasets published by the TROMPA project will be published using the Vocabulary of Linked Datasets (VoID).

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 data created in TROMPA will be openly accessible. These data can contain:

- Crowd annotations (under the restriction that they are anonymous) and other metadata related to the use cases will be published.
- Metadata that is already open and is owned by the TROMPA partners (as for example CDR metadata) will remain open.
- The linking between objects from different resources will be made available.
- The scientific publications and datasets will be all uploaded to Zenodo and thus will be open.

Exceptions for keeping data close are the following:

- If certain data can be used to trace a person or it is rights restricted, it won't be openly accessible.
- 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.
- Performer data contributions will be by default closed unless consent is provided for publication.

Specify how the data will be made available

The Contributor Environment data will be accessible directly through a web API and indirectly by means of 4 React components. The components support web based user interfaces, including 5 pilot applications, by exposing predefined functionalities like a semantic search interface or an annotation tool, 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.

All datasets and scientific publications will be available on Zenodo, and the source code on open access repositories (see section "Outline the approach for clear versioning"). We have already created a TROMPA community on Zenodo[3]. Regarding other data formats described in Table 2.1, in the next version of this deliverable we will provide a detailed table similar to Table 2.1, where for each of the data formats we will provide an online repository to publish this data.

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 will be accessible through http(s) protocol via RESTful and GraphQL interfaces. API docs will be 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 (see above) will be if the repository also includes an API in order to programatically access the data. Otherwise, all online repositories must be accessible with a web browser. Regarding scientific publications and datasets, we will use conventional data formats that require open licence software to be accessed.

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

As a design principle, only metadata will be stored within the Contributor Environment. Data from public repositories will be referred to only, and remain where it is. Produced data will be stored in principle by participants and made available through an URL or documented API. Where there are concerns about performance, the produced data might be stored in a Contributor Environment managed s3 bucket.

Source code will be deposited in Bitbucket or Github. Publications will be deposited in Zenodo. Documents produced by project partners such as deliverables will be deposited to project website 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 will have a robust Access Control layer which enables management of both read and write access, on a user and user-group basis. Furthermore, the internal data model of the CE will provide entity properties to describe rights restrictions, which can then be applied either CE side or client-side.

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

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.

For its metadata, the Contributor Environment will adopt a metadata standard like DCMI. In general, we will reuse and build on widely used existing ontologies and data models, e.g. Music Ontology, Web Annotations, PROV-O. Regarding datasets and music excerpts, we will reuse identifiers from other databases instead of creating our own whenever possible (e.g. MusicBrainz, Wikidata). These standards will described in the next version of this deliverable.

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 will facilitate and encourage the use of (multiple) ontology references for all datatypes (entities and relations) stored, supporting serialisation into RDF Linked Data, allowing interoperability.

3.4 Increase data re-use (through clarifying licenses)

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

A licence property is default for all entities stored within the Contributor Environment. Responsibility to store the correct licencing information, and retrieve and apply it correctly to user applications remains with the application builder.

For crowd annotations we will 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.

A special working session is organized in the forthcoming consortium meeting in Vienna (28-29 November, 2018) to define data licensing. The current deliverable will be updated with the outcomes of this meeting in its next version (Month 18).

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. However, in the case that it is needed, a date-range can be made part of the data-licensing data.

Specify whether the data produced and/or used in the project is useable 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 will be several processes that will be 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 will use file formats that preserve data quality as high as possible. Backup of the

data will be kept periodically. The exact procedures will be reported in detail in the next version of this deliverable.

Moreover, 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 will provide 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 will ensure the quality of these documents. Regarding research publications presenting the research outcomes of TROMPA, these will be 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

Since most of the data will be published under open licence, it will be reusable forever. The exact length of time that the data will remain accessible after the project is to be determined in the next version of this deliverable.

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 will repositories repositories that are free to use (github, zenodo). Regarding other types of data, it might be possible to get support from some companies to host content (e.g. musescore for musescore.com for hosting music scores). Some of the metadata generated during the project will be contributed back to the original repositories (musicbrainz, wikipedia, CDR).

Regarding the data that is hosted on the Contributor Environment there will be an ongoing monthly cost for hosting. More details on the hosting will be discussed in the forthcoming meeting in Vienna 28-29 November. The exact costs will be reported in next versions of this deliverable.

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, whose will be responsible for the data management in TROMPA. **David Weigl** from MDW, initially accepted this role. Within the CE, VD will be 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.

Describe costs and potential value of long term preservation

Long term preservation will have a great value to Europe's cultural heritage, since we will enrich classical music public domain archives, we will comprehensively unlock our musical cultural heritage, we will provide a better insight to classical music and its interpretations, and will make classical music more accessible to citizens.

Regarding the cost for long term preservation, if we require our own infrastructure to host data that we create, then we have ongoing hosting fees. Because of this, TROMPA will focus to host our data on external services as much as possible (open repositories for specific types of data, store annotations to the existing repertoire repositories).

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 we don't expected to collect any Sensitive Personal Information. All TROMPA partners will comply with relevant privacy regulations in ensuring that we only collect the maximum necessary Personal Information and explain to participants what we use it for. We will develop a data processing plan to ensure that Personal Information is only made available to researchers who require it. In general we will collect personal information as:

- **Interviews**: One-to-one interviews with users. Depending on the use case these can choir conductors/singers, music enthusiasts or instrument players to gather opinion about the use case. These interview will be offline, meaning that no audio or video will be recorded, only written notes will be taken.
- **Questionnaires**: Online or printed questionnaires will be 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 grading of human participants answers to tests (music enthusiasts use case) or music rehearsals (choir singers, instrument players use case).
- **Observations**: written notes, notes, pictures, audio, videos may be taken if activities are conducted in face to face environments.
- **Log files**: Log files collecting actions of teachers and learners with the system.

Regarding the backup of the data, we will ensure that we make backups of data that we generate. Due to the fact that we plan to contribute data back to original repositories as we create it, we de facto create an additional backup of this data by uploading it to those repositories. We can summarize the data protection mechanism as follows:

- **Data storage**: Data will be stored in secure servers where only authorized people from the consortium can access to.
- **Data backup**: Data will be 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 a responsible for the data maintenance, curation and quality assurance.

- **Data anonymization**: All data collected will be anonymized when possible. Participants will be given an identification number. We will 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 will have access to the data. The **Data Officer** will be responsible for granting access to the data. Access to data will be given only for research purposes.
- **Data publication**: For the purpose of data publication, i.e. scientific datasets, publications, some of the data collected might be published. In the case the data will be 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) will be gathered.

6. Ethical Aspects

Details about the data protection and anonymization are given in Section 5 - Data Security. All partners that will collect data from human during the use cases, will have an ethics approval from their ethics committees. All participants will provide 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, POPD Requirement No. 5 and POPD Requirement No.6.

7. Conclusion

In this deliverable we presented a first 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 will meet 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.

In the next version of this deliverable which is to be submitted in Month 18, we will provide clearer details of our data management. We will define the exact data and metadata standards and formats, file naming and keywords conventions and data anonymization procedures. We will provide detailed information about where and how the data will be stored and backed up, the management of the users with respect to their access to the data.

8. References

8.1 Written references

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

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