TROMPA

TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 3.1

Data Resource Preparation

Grant Agreement nr	770376
Project runtime	May 2018 - April 2021
Document Reference	TR-D3.1-Data Resource Preparation v2
Work Package	WP3 - Automated Music Data Processing and Linking
Deliverable Type	Report
Dissemination Level	PU- Public
Document due date	31 October 2019
Date of submission	31 October 2019
Leader	UPF
Contact Person	Aggelos Gkiokas (aggelos.gkiokas@upf.edu)
Authors	Aggelos Gkiokas, Emilia Gomez, Alastair Porter, Helena Cuesta (UPF), Álvaro Sarasúa (VL), David Weigl, Werner Goebl (MDW), Tim Crawford (GOLD), Cynthia Liem (TUD), Marcel van Tilburg (RCO),

	Ingmar Vroomen (CDR)
Reviewers	David Weigl (MDW), Wim Klerkx (VD)

Executive Summary

This document is the 2nd version and the final version of the deliverable D3.1 Data Resource Preparation. It is submitted in the scope of Work Package 3 and it aims to provide information about the target repertoires that will be used in TROMPA use cases, as well as how these repertoires will be imported in the TROMPA ecosystem. To an extent, the contents of this version are identical to the previous version, but for reasons of clarity and completeness, since this is the final version of the deliverable, we preferred this approach instead of providing only the differences with the previous version. The elements of the selected repertoires are connected to existing online repositories, and will be exploited and enriched in the TROMPA use cases.

At first, we present an overview of the most common Public Domain music repositories considered in the TROMPA project for the different use cases: what volumes of data are contained, the corresponding data licence, an overview of the represented musical styles as well as potential uses for these repositories. Next we provide an overview of the five TROMPA use cases, including the technical and repertoire requirements for each of them, and detail the contributions expected to be made to the repositories through our use cases. Finally, we summarize all target repertoires and we discuss technical details required for further developments in the project.

In this deliverable, we consider the following existing music repositories: **IMSLP** covers around 124,000 unique works, providing scanned scored images (~405,000) that can be exploited in our use cases (e.g. for conversion of scanned images to symbolic format); symbolic scores (~37,000) that can be directly used for the use cases or for training methods on WP3; and audio recordings (~ 47,000) of performances. It can also serve as a place to deposit TROMPA contributions to the Public Domain.

Similar to the IMSLP, the **Choral Public Domain Library** (CPDL) is a sheet music archive, in this case focusing on vocal and choral music in the Public Domain. It contains around 10,000 unique works, most of which are available as scanned score images in PDF format.

MuseScore[1] is an online platform that allows the users of the MuseScore software to publish and share their music scores online. MuseScore mostly contains user-provided transcriptions of music pieces that are not limited to classical.

ECOLM is an electronic corpus of lute music, predominantly notated in tablature. The total size of the ECOLM repertory available to TROMPA is about 2,000 encoded pieces, which is supplemented by a further 6,000 whose encodings have been translated from different formats.

Early Music Online (EMO) is a collection of 300 books of printed music from before 1600, which have been digitised and made available as images by the British Library. They cover almost all genres and styles of music from the 16th century, including works by all the leading composers of the age. The music is almost exclusively vocal, for one to twelve voices, mostly printed in part-books. Also in the early music field are two relevant public collections of Spanish music suitable for choral singing:

[1] www.musescore.com

BDH, containing facsimiles of c100 books of early printed vocal music, and **TLdV**, which comprises about 2,000 encoded scores of similar music.

MusicBrainz is a community-based repository that stores information about artists, their recorded works and the relationships between them. It describes a large variety of commercial music (including western classical music), consisting of around 1 million artists and 18 million tracks.

AcousticBrainz provides audio descriptors (rhythm, key, genre tags, mood tags) for almost 4 million tracks identified using MusicBrainz identifiers.

Kunst der Fuge consists of around 19,000 MIDI files of western classical music, some of them published under a Creative Commons License.

The **CDR Muziekweb catalogue** contains music in all music styles, including popular, jazz, world music and classical. CDR strives to collect all music released in the Netherlands; this is the only criterion to be included in the collection. As of 2018, the collection holds over 600,000 CD's, 300,000 LP's and 25,000 music-DVD's. In total, there are 579 music styles in use. For classical music, the collection has albums in all styles and genres, from all style periods, labels and countries.

The **Vienna 4x22 Piano Corpus** consists of score-aligned performance and audio recordings of 4 short piece excerpts by Mozart, Schubert, and Chopin, each of which performed by 22 professional pianists in 1999.

The **humdrum data repository** is a collection of musical scores in the Humdrum (kern) file format, containing large parts of the standard repertoire for piano solo, string quartet, and choirs, including composers such as J. S. Bach, Domenico Scarlatti, Haydn, Mozart, Beethoven, Hummel, Chopin.

Over the past years, the National Library of the Netherlands has worked on digitizing the newspapers that were published in the Netherlands over the past centuries. The **Delpher** platform provides access to dutch newspapers from the years 1618–1995, amounting to over 12 million scanned newspaper pages. The newspapers from 1618–1876 are considered to not have any copyright-protected material, and a full download of all full texts in OCR, ALTO and XML format is available. This information will be useful to increase and improve contextual understanding of how musical works and persons historically were perceived.

Biblioteca Digital Hispanica contains around 5,000 high resolution pages of Spanish music of all periods. In particular, it has a major component of 16th-century printed vocal music from a period when Spanish composers were highly esteemed. It contains much parallel repertory with **2.17 Tomás Luis de Victoria**, which gives an opportunity for exploring possibilities for practical TROMPA linkage between various manifestations of a given work. This collection, privately compiled by Nancho Álvarez, contains about 2,000 choral works by leading Spanish composers of the 16th century, Victoria, Morales, Guerrero, Vásquez and others. The music is highly suitable for singing by amateur choirs, as it is not too difficult; furthermore, its essential simplicity (relative to later music) makes it ideal for testing TROMPA's methods for OMR, score annotation and audio alignment components as well as our interfaces for music scholars and choral singers.

This deliverable also connects existing public domain archives with the different pilots, defining core repertoire for the pilots to be enriched during the project. Each pilot corresponds to one of five TROMPA use cases, respectively targeting music scholars, orchestras, instrument players, choir singers, and music enthusiasts.

In the **Music scholars use case,** scholars will be able to find connections between music works on multiple levels: from co-occurrences of melodies, harmonic and rhythmic progressions, to the large-scale structural similarities of musical works. The score is the main starting point and anchor for such research. Our aim for this pilot is to provide an interface for the selection and display of musical

scores (sheet music) from the TROMPA collections, for annotating them, and for searching them (by text or by example). The repertoire for the initial version of the music scholars' pilot will mainly comprise early music from the 16th century from different resources. As the project progresses, we shall provide facilities for digital enquiries by making the search API publicly available and publishing the bulk of data collected in TROMPA as a public linked open data dataset.

The **Orchestras Pilot** will aim to digitize all symphonies by Gustav Mahler, a core repertoire of most orchestras, making them available free of charge. TROMPA will develop crowdsourcing technology to engage music lovers in encoding (out of copyright) scores available as digitized score images from IMSLP. Using RCO orchestra members' and librarians' expertise, we will develop a tool to extract good quality instrumental parts from these scores. The technological and technical requirements of the orchestras' use case is twofold. Firstly, scanned score image analysis software in conjunction with crowd-sourced annotation mechanisms will be deployed in order to encode Mahler's music scores in digital format (MEI, or MusicXML). The second technical requirement is the capability of annotating music scores and sharing annotations. All the digitized score encodings of the Mahler Symphonies that will be derived from the use case will be deposited in public domain musical archives. Moreover RCO will offer its most recent annotations of one of the Mahler symphonies for digitization, as well as other archives such as annotated orchestral scores of Willem Mengelberg and a big part of the available Mengelberg Concertgebouw recordings.

The **Instrument Players Pilot** provides musicians engaging in rehearsal or performance with a "Performance Companion" system capable of characterising performative aspects of their playing. By alignment of performance recordings and metadata with musical score encodings, the characteristics can be assessed and compared against those derived from other performances or reference recordings. Initially, the pilot will focus on pianists performing Beethoven's piano works (primarily his Sonatas, Variation works, and Concertos). Performance recordings and characterizations produced by the Performance Companion will be captured and published, contributing to the available repertoire of Beethoven recordings. We will work towards producing a complete set of Beethoven piano work encodings over the course of the project. The technological components required for this system can be split into two groups: score alignment and performance characterisation. Score encodings created for the pilot will be made publicly available. Score segmentations, created manually and automatically as part of the score alignment task, will be associated with these encodings and published under open licenses. Finally, performance metadata and recordings (with performer permission) will be made publicly available.

The goal of the **Choir Singers Pilot** is to assist amateur choir singers during individual performance and to provide functionality for the choir conductor to create repertoires and to listen to performances by choir members. Users of the pilot will be able to synthesize existing scores, sing-along with the synthesized voices, and receive feedback on their performance. The accompanying voices will be available for music in Spanish, Catalan, Latin, English and German, and the pilot will then focus on pieces in these languages from the repertoires considered below. A set of selected pieces is provided for the first iteration of the pilot, and several composers are selected as representative of the target languages: Tomás Luis de Victoria, Anton Bruckner and Josquin des Prez. This pilot will be mostly based on the audio processing techniques to be developed in Task 3.3, where we will research and develop techniques for audio synthesis of choir singing. The pilot will collect data from users to improve the voice synthesis algorithms. Users should be able to provide general ratings for the synthesis (e.g. by rating the overall quality of the synthesis) and to make timestamped annotations for the generated material, i.e., allowing the user to input free text comments to inform about specific problems (e.g. "this phoneme sounds weird at this point in time

in the soprano voice"). Through the use of the pilot, synthesized versions of the scores will be generated and stored, accessible through public URLs. These synthesized versions will be associated to the scores. The same applies for recordings of performances by users of the pilot (needed for providing automatic feedback), although in this case, their addition to the repertoire will depend on getting appropriate permission from the user.

In the **Music Enthusiasts Pilot** we will provide interaction mechanisms with musical cultural heritage content targeted at people that, although lacking formal musical knowledge, are interest in learning more about music. The main study of this use case is to build a music recommendation system, focused on classical music, with the possibility of integrating user feedback and annotations of the content. The music enthusiasts use case will be focused on the existing music collection of CDR Muziekweb. We will focus on the classical music repertoire of this library, but we will not limit ourselves to that. Techniques on higher level semantics extraction, such as emotion classification,will be adopted and evaluated during the use case (Deliverable 3.2). This use case will contribute new evaluation and ground truth data on the CDR repository, e.g. opinions/ratings about the outcome of a recommendation system, emotion tags and other annotations of music pieces.

The target repertoires for these user pilots come from many **separate repositories** and may have **diverse representation** schemas. We store metadata from each target repository in the WP5 TROMPA Contributor Environment (CE) and we will link **representations** the same real world item in **different repositories** to each other so that information about these items can be shared regardless of where that information comes from. The CE uses an internal data model that is primarily based on the schema.org[2] ontology schema, and all metadata items that will be stored in the CE will be mapped on this schema. We make the effort to maintain the MusicBrainz repository in close sync with the data available in the CE, linking this data where possible to the metadata from each target repository. Guidelines that describe how metadata can be imported into the CE, and how this data from external repositories can be linked to existing metadata in the CE is being developed. These guidelines are in development and are described in detail in Deliverable D2.3 - Complete Requirements and in the 2nd version of the Deliverable D5.1 - Data Infrastructure. Two tools have been developed until this stage of the project for loading data. One tool for **storing MusicBrainz metadata** in the CE, and one tool that **stores metadata in the CE** and links this metadata to **actual data** (scores, audios, videos, performances) that is stored in external repositories.

[2] https://schema.org

Version Log		
#	Date	Description
v1.1	23 October 2019	Initial version submitted for internal review
v1.2	30 October 2019	Revised version after internal review
v2.0	31 October 2019	Final version submitted to EC

Table of Contents

1. Introduction

This document is the 2nd version and the final version of the deliverable D3.1 Data Resource Preparation. It is in the scope of Work Package 3 and it aims to provide information about the target repertoires that will be used in TROMPA use cases, as well as how these repertoires will be imported in the TROMPA ecosystem. To an extent, the contents of this version are identical to the previous version[3], but for reasons of clarity and completeness, since this is the final version of the deliverable, we preferred this approach instead of providing only the differences with the previous version.

The elements of the selected repertoires are connected to existing online repositories, and will be exploited and enriched in the TROMPA use cases. The structure of the deliverable is as follows. First, we present an updated list of most of the common Public Domain music repositories that are considered during the definition of the target repertoires for the TROMPA use cases (Section 2.1). In the rest of Section 2 (Sections 2.2 - 2.15) we provide an informative overview for each of these repositories; what volumes of data are contained, the corresponding data licence, an overview of the musical styles as well as the uses of these repositories. In Section 3 we present the five TROMPA use cases, with an overview of each one, the technical and repertoire requirements for each, as well as the contributions we expect to make to public repositories through each use case. In Section 4 we summarize all target repertoire, and in Section 5 we discuss the technical details about integrating repertoire data / metadata to the Contributor Environment..

2. Review of Public Domain Repositories

2.1 Overview of the Public Domain Repositories

Table 2.1 summarizes all the public repositories that we consider in the context of TROMPA. This list includes all the major public domain repositories that target western classical music. Some of the repositories presented in the previous version are omitted, since they will not finally exploited in TROMPA. In the following subsections we will provide details for each of these repositories.

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

[3] https://trompamusic.eu/deliverables/TR-D3.1-Data_Resource_Preparation_v1.pdf
[4] https://imslp.org/
[5] http://www.cpdl.org/wiki/
[6] https://musescore.com/

ECOLM[7] - 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[8] - 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[9]	Automatically extracted features for 10 million music recordings (of all types of recorded music), JSON, 400GB
MusicBrainz[10]	Metadata for recorded music, 1.4m artists, 2m releases, 20m recordings[11], Accessible via webservice (XML, JSON) or as a Database Archive
Kunst der Fuge[12]	19,300 MIDIs
Humdrum-data Repository[13]	Over 4,100 scores encoded in Humdrum (kern) format
CDR Muziekweb catalogue[14]	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[15]	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[16]	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[17] - 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.

[7] http://www.ecolm.org/

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

[9] https://acousticbrainz.org/

[10] https://musicbrainz.org/

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

[12] http://kunstderfuge.com/

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

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

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

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

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

TLdV[18] - 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 2.1. TROMPA repositories to be considered in the user pilots.

The following sections provide a more detailed explanation of the different repositories to be considered within TROMPA.

2.2. IMSLP Petrucci Music Library

International Music Score Library Project (IMSLP) was started in 2006 and consists mainly of scans of old musical editions of western classical music that are in the Public Domain. Moreover, it hosts scores by contemporary composers who wish to share their works under a Creative Commons license. IMSLP is an associated partner in TROMPA consortium.

IMSLP is a great source of information and can be potentially used for many tasks. IMSLP covers around 124,000 unique works. It is a great source providing scanned scored images (~405,000) that can be exploited in the use cases (e.g. for conversion of scanned images to symbolic format); symbolic scores (~37,000) that can be directly used for the use cases or for training methods on Work Package 3 (Task 3.2 - Music Description, Task 3.4 Visual Analysis of Scores, Task 3.5 Alignment of Musical Resources); and audio recordings (~ 47,000) of performances. It can also serve as a place to deposit TROMPA contributions to the Public Domain (e.g. new digitized scores, performances). Apart from the raw score images, IMSLP also hosts scores in digitized symbolic format (MIDI, MusicXML, MSCZ) and audio recordings. Detailed information about the supported formats can be found on their website[19]. IMSLP is structured so that each represented data entity (e.g. composer, work, media resource) can be directly accessed via a URI. Moreover, IMSLP has a great community, with millions of users and around 6,000 active contributing members. TROMPA can take advantage of this community to incentivize users to participate in TROMPA use cases (see **Deliverable 2.1 - Early Requirements**, section 2.1).

2.3. Choral Public Domain Library

Similar to the IMSLP, the Choral Public Domain Library (CPDL) is a sheet music archive, in this case focusing on vocal and choral music in the Public Domain. It contains around 10,000 unique works, most of which are available as scanned score images in PDF format. There are also some works in symbolic music format such as MusicXML, MuseScore and audio performances[20]. These data are

[18] https://www.uma.es/victoria/
[19] https://imslp.org/wiki/IMSLP:File_formats
[20] http://www1.cpdl.org/wiki/index.php/Template:Legend

directly accessible with a public URI. This repository can be used as a resource for defining the repertoire of the Choir Singers use case, as well as a resource for training/developing/evaluating methods related to the choir singing synthesis (Task 3.2 - Music Description, Task 3.3 - Audio Processing). Since it is a community based archive, it can also be used to deposit TROMPA contributions such as new scores, user performances or synthesized choir works. It can also be used as a hub to find people (e.g. choir singers) that might be interested in participating in the choir use case.

2.4. MuseScore

MuseScore[21] is an online platform that allows the users of the MuseScore software to publish and share their music scores online. MuseScore mostly contains user-provided transcriptions of classical, western popular, and jazz music. Musescore.com contains about 3,000,000 scores, around 10% of which (~300,000) are shared. Although this repository predominantly contains amateur encodings which may not be of suitable quality for use within TROMPA, it provides a candidate repository for the contribution of digitized scores generated by the project. Recently, Musescore.com features special sub-projects dedicated to individual works that are encoded in a concerted effort by several editors, resulting in high-level encodings of important works of the standard repertoire[22]. Most of the data is available in various formats (museScore format, pdf, MusicXML, MIDI) which can be exploited in some use cases.

2.5 ECOLM: An electronic corpus of lute music

ECOLM is a collection of encodings of music for the lute - an instrument of central importance in European music history between c1500 and c1800 - and related instruments (bandora, cittern, theorbo, etc.). The music was almost entirely notated in tablature, a form of notation giving specific physical instructions for its performance (rather than the abstract description of conventional notation) dependent on various aspects of the instrument, such as its tuning. The musical styles are of three basic types: free compositions in more or less improvisatory style (including preludes, ricercars, fantasies and toccatas); arrangements of vocal (and occasionally instrumental) ensemble music (madrigals, chansons and church music); dance music.

Apart from its historical significance, this music is important for testing certain assumptions concerning the semantics of music. For example, music in tablature does not specify precise chromatic pitch, but rather the intervals between notes; similarly, the duration of individual notes which commence together at a certain time is not specified, so determining voice-leading has to be an interpretive act. Although this is explicitly a historical repertory, it has special value in testing methods that could be applied directly to the vast repertory of online ‘tabs’ which are shared today by enthusiasts of folk, jazz and various popular musics.

The total size of the ECOLM repertory available to TROMPA is about 2,000 encoded pieces, which is supplemented by a further 6,000 whose encodings have been translated from different formats. They approximately cover the period 1500 to 1800, but the subset of this extra material up to c1650, and thus corresponding with ECOLM, amounts to about 5,000 pieces.

[21] www.musescore.com
[22] https://musescore.com/opengoldberg

All the music is available in TabCode, a data-entry format which covers most of the notational features of the tablature of the renaissance period. Ongoing work on an MEI format for tablature will allow translation to TabMEI during the TROMPA project period, though this is currently in development. Also, tools for importing and processing tablature into Music21 are under development and will soon become available for processing this data alongside that in conventional notation. The ECOLM project has developed a display, playback and editing web-based interface which could potentially be used within TROMPA, though the completion of the TabMEI format specification will allow early web-browser rendering (eventually interactive) through Verovio.

The combination of TabMEI and Verovio allows for full Linked Data exploitation, owing to the use of common xml:ids for all tablature objects as they are processed or translated into other compatible formats (e.g. within the TROMPA Data Infrastructure).

Currently the ECOLM data is stored in a MySQL database at Goldsmiths, which is backed-up and maintained professionally, and can be accessed by SQL queries (subject to permissions). The extra non-ECOLM tablature data currently resides on a filestore at Goldsmiths currently without public access, but freely available to TROMPA.

2.6 Early Music Online

Early Music Online (EMO) is a collection of 300 books of printed music from before 1600, which have been digitised and made available as images by the British Library. They cover almost all genres and styles of music from the 16th century, including that by all the leading composers of the age. (Approximately 10% of the collection is of music in tablature, mostly for lute, and has been mostly incorporated into ECOLM, see above.) The music is almost exclusively vocal, for one to twelve voices, mostly printed in part-books, each for a different singer; it therefore consists entirely of monophonic parts, not scores. Approximately half the resource comprises sacred music, the rest is secular. EMO does contain some instrumental ensemble music, some of which is itself derived from vocal originals.

While a large amount of 16c vocal music is available in modern editions (see IMSLP Petrucci Library, above), coverage is not total or consistent, being mainly focussed on the leading composers of the time. A large subset, approximately 32,000 page-images, of the EMO corpus has been subjected to Optical Music Recognition (OMR) with the specialist Aruspix program, and the MEI encodings thus produced, despite the inevitable OMR errors, have been shown to be useful for indexing purposes (a la Google Books); generating scores from these pages is, however, a formidable challenge beyond the scope of TROMPA. We hope to be able to align EMO data with encoded scores from modern editions, so that users could follow automatic links from modern editions to facsimile pages and vice versa.

Recent experiments at Goldsmiths with state of the art indexing techniques have led to new projects which will provide a very efficient musical content-based search interface to a much-enlarged page-image resource based on EMO but larger by a factor of 10 (with many extra images contributed by music libraries in Europe and the US). When this is in place (during the TROMPA project period), it can be incorporated into the scholarly toolkit for TROMPA.

The BL's public interface to EMO currently allows viewing of jpeg files, soon to be provided with the option to download higher-resolution tiff images. A set of the high-resolution image files, together with Aruspix and MEI files of the recognised music as well as specialist indexes, is maintained on a private server at Goldsmiths, currently without public access, but freely available to TROMPA.

2.7 MusicBrainz

MusicBrainz is an open data music database that stores information about artists, their recorded works (album titles, track titles, length of track, release date and country, cover artwork and other metadata), and the relationships between them. It is a community based repository, and these entries are maintained and updated by volunteer editors. It contains a large variety of commercial music (including western classical music), consisting of around 1 million artists and 18 million tracks. Apart from bibliographic metadata, recordings described by MusicBrains are also associated with derived audio feature metadata published by the AcousticBrainz project (see sec 2.8). MusicBrainz data is directly accessed by public URIs.

The fact that MusicBrainz provides metadata as either public domain facts, or under a Creative Commons License means that it is a great resource of information that can be used in TROMPA for several purposes as:

- **Use of unique identifiers**: MusicBrainz unique identifiers and metadata information can be the main reference for the music entities (composers, works) that will be used in TROMPA.
- **Repository**: MusicBrainz can be used as a repository to add new entities (composers, performers, works), and other metadata as tags that will be created in the TROMPA use cases.
- **User communities:** MusicBrainz active users (~250,000) can potentially be incentivised to participate in TROMPA use cases (e.g., Music Enthusiasts)

2.9 AcousticBrainz

AcousticBrainz is an open data music database that provides various audio descriptors. It is growing fast and currently almost 4 million tracks are indexed. AcousticBrainz uses MusicBrainz identifiers and the Essentia[23] library from UPF for the extraction of features. These features descriptors are related to rhythm (bpm, danceability), tonality (key, chords), timbre tags (vocal, genre) and mood tags (happy, sad, relaxed). Potential use of this repository is:

- **Use of acoustic features**: Existing acoustic features in AcousticBrainz can be used to facilitate research under Work Package 3 and for some of the use cases (e.g., Music Enthusiasts)
- **Storage of acoustic features**: Can be used as a repository to store low-level descriptors that will be extracted during the use cases.

2.9 Kunst der Fuge

Kunst der Fuge consists of around 19,000 MIDI files of western classical music, some of them published under a Creative Commons License. This repository can be potentially used in some of the TROMPA use cases, such as Music Scholars or Instrument Players.

2.10 Humdrum-data Repository

The humdrum-data repository contains a collection of musical scores in the Humdrum (kern) file format for use with Humdrum-processing software, containing large parts of the standard repertoire

[23] https://essentia.upf.edu/documentation/

for piano solo, string quartet, and choirs, including composers such as J. S. Bach, Domenico Scarlatti, Haydn, Mozart, Beethoven, Hummel, Chopin. This repository is provided online[24] and is constantly updated. The Chopin complete solo repertoire should be online within the coming years as well as a wide range of Polish music (personal communication with Craig Sapp, 2019).

The Humdrum file format (.kern) is easily converted to MEI through Verovio and may be accessed by an online interactive viewer[25]. For a subset of this corpus, a special online interface is provided for choir singers that allows to modify and personalize the score engravings for practical use in choir singing (Bach-370-chorales[26]). Craig Sapp is currently also working on encodings of Renaissance polyphonic music corpora[27,28].

2.11 CDR Muziekweb catalogue

The CDR Muziekweb catalogue contains music in all music styles, including popular, jazz, world music and classical. CDR strives to collect all music released in the Netherlands; this is the only criterion to be included in the collection. As of 2018, the collection holds over 600,000 CD's, 300,000 LP's and 25,000 music-DVD's. In total, there are 579 music styles in use. For classical music, the collection has albums in all styles and genres, from all style periods, labels and countries. There are six major classical categories, each with multiple styles: orchestral works; concerts; chamber music; solo concerts; vocal music and miscellaneous (including e.g. early music and electronic music and musique concrête).

The dataset and structured metadata could be used for various purposes. Because of the large and diverse amount of classic works and titles it could function as a reference set to match other collections to. The music data can be used for feature analysis or other forms of MIR. 30 seconds samples of the music (because of copyrights) can be embedded on every site, as previews or for educational purposes. Available file types for music are FLAC, MP3 and AAC, for video MP4. As of september 2018, there are 7,748,978 digital audio files available, 2,090,791 of them classical music. In total there is 590TB in use for audio and video storage.

CDR Muziekweb uses its own unique identifiers for artists/composers/performers ('contributors'), work titles and releases/albums. Personal names are linked to the ISNI-database (and Wikipedia). Links with Musicbrainz, Discogs and Wikidata are in an experimental phase. The data is stored in two separate locations, the first in the library and a second in an external storage centre. For the main entities (contributors, work titles and releases/albums) permalinks are available on the website muziekweb.nl. The metadata can be accessed through a API implemented as a REST webservice. Audio without a license is only accessible in 30 seconds clips.

2.12 The Vienna 4x22 Piano Corpus

The Vienna 4x22 Piano Corpus [1] consists of 4 short piece excerpts by Mozart, Schubert, and Chopin, each of which performed by 22 professional pianists in 1999. The sound of the performances was recorded by stereo microphones and additionally the performance parameters by an embedded computer system built into a Bösendorfer Imperial concert grand piano. In addition to the 22

[24] https://github.com/humdrum-tools/humdrum-data
[25] https://verovio.humdrum.org/
[26] https://chorales.sapp.org/
[27] http://josquin.stanford.edu/
[28] http://www.tassomusic.org/

individual performances, an artificial performance (No. 23) is provided for the Chopin pieces that was artificially rendered as the average of the expressive parameters of all 22 performances (timing, velocity, see [1, 2].

The performance data is available in recorded uncompressed audio (wav), MIDI format, the original Bösendorfer format, as well as in a textual match file format (".match") that contains the musical score information aligned to the performance data on a note-by-note basis.

The Vienna 4xx22 Piano Corpus may serve as training data set for machine learning projects on individual performance style as well as a basis for testing use case scenarios with professional instrumentalists.

The data corpus is stored at a permanent storage at mdw[29], persistently linked by a DOI[30]. Each file has a persistent URI and may be addressed with linked data. Still missing in this data set are encodings of the scores in MEI (or MusicXML), modifications of the existing alignments (".match") to correspond with these encodings and making this corpus compatible with a Linked Data structure.

2.13 Delpher newspaper collection of the National Library of The Netherlands

Over the past years, the National Library of The Netherlands has worked on digitizing the newspapers that were published in the Netherlands over the past centuries. The newspapers have been scanned, automatic Optical Character Recognition and segmentation has been performed, and subsequently the results are indexed and made accessible through the Delpher[31] platform. Presently, newspapers from the years 1618-1995 can be accessed through the Delpher interface, amounting to over 12 million scanned newspaper pages. The newspapers from 1618-1876 are considered to not have any copyright-protected material, and a full download of all full texts in OCR, ALTO and XML format is available through the Delpher open newspaper archive[32] (111 GB in total). For newer newspapers, beyond manual search in the public Delpher portal, the National Library can distribute API keys on request for research purposes.

The newspaper corpus is not formally a music corpus, but recently, efforts have started to consider the music-related information in it in more structured ways [3]. This information will be useful to increase and improve contextual understanding of how musical works and persons historically were perceived. This will be particularly valuable to TROMPA's music scholars and music enthusiasts use cases.

To truly move towards a systematic and usable music corpus, as described in [3], both syntactic challenges (in particular, dealing with OCR errors) and semantic challenges (e.g. non-trivial entity resolutions) will need to be navigated. Focusing on named entities indicating music-related people or organizations (composers, artists, bands), based on a list of musical contributors from the CDR Muziekweb catalogue, a list of possible Wikidata entities have been identified with the Mix'n'match tool. Out of 47,400 potential entities, it presently is investigated whether and where these entities can be found in the Delpher corpus. As one way to perform filtering, a hard rule is that an entity cannot occur before its date of conception or date of birth. As soon as this cleaning step is finalized, a

[29] (https://repo.mdw.ac.at/projects/IWK/the_vienna_4x22_piano_corpus/
[30] https://doi.org/10.21939/4X22
[31] https://www.delpher.nl/
[32] https://www.delpher.nl/nl/platform/pages/helpitems?nid=513&scrollitem=true

linked musical entity research corpus with Delpher article references will be released to the community, from which the TROMPA project will be able to directly profit in its use cases.

2.14 Biblioteca Digital Hispánica

The Spanish National Library has assembled this virtual collection from various music libraries to represent the historical legacy of Spanish music of all periods. In particular, it has a major component of 16th-century printed vocal music from a period when Spanish composers were highly esteemed. This collection is not just of interest to music scholars, but also to choral singers and music enthusiasts. It contains much parallel repertory with **2.15 Tomás Luis de Victoria**, below, which gives an opportunity for exploring possibilities for practical TROMPA linkage between various manifestations of a given work. The music is printed in separate part-books, each giving a single voice-part, rendering the pages suitable for optical music recognition (OMR) using the open-source program Aruspix. This outputs MEI files, which can be used to generate indexes or allow possibilities for automatic alignment with encoded scores or recordings (despite the presence of inevitable noise from the OMR process). The music is available as high-resolution PDFs from which TIFF files for OMR can easily be converted. There are about 100 such digital facsimiles on the web-site, which we estimate to be about 5,000 pages of music. There does not appear to be an API for automated download, but the modest scale of the collection means that the data can be harvested manually.

2.15 Tomás Luis de Victoria

This collection, privately compiled by Nancho Álvarez, contains about 2,000 choral works by leading Spanish composers of the 16th century, Victoria, Morales, Guerrero, Vásquez and others. (Many of the original sources can be found in **2.6 Early Music Online** and **2.14 Biblioteca Digital Hispánica**, above.) The music is highly suitable for singing by amateur choirs, as it is not too difficult; furthermore, its essential simplicity (relative to later music) makes it ideal for testing TROMPA's methods for OMR, score annotation and audio alignment, as well as our interfaces for music scholars and choral singers. The music is all encoded as Lilypond, a high-quality engraving format which can easily be converted to MusicXML or MEI, and thus made available to the whole of TROMPA; for some works individual voice parts have also been provided in MIDI format, which can give valuable data for TROMPA's choir singing components. In many cases, the web-site gives links to video or audio performances of the music, often on YouTube, which could easily be incorporated into TROMPA as annotations. There is no public API mentioned on the site, but manual/semi-automated download of the files is feasible.

2.16 Video Data

Video data can be potentially exploited in some of the use cases. YouTube is a vast source of videos that can be potentially used in TROMPA use cases, as for example videos of orchestras performing classical pieces. Moreover video data can be provided during the use cases, such as video recordings of choirs during their performance, or of instrument players. The potential use of video data and respective repositories are to be defined in the next months during the use case definition (see Section 6 - Conclusion).

2.17 Scientific Repositories

Apart from the repositories that will be used in the use cases, there are several repositories that will be used or created during TROMPA for research purposes. An example is the Choral Singing Dataset, that contains the individual audio recordings of 16 singers of the Anton Bruckner Choir from Barcelona (Spain) performing 3 different pieces a cappella, together with their associated MIDI files, is used to train synthetic voices (Task 3.3 - Audio Processing) in conjunction with the ESMUC Choir Dataset[33]. This corpus along with all datasets that will be created will be deposited in TROMPA Zenodo community[34]. More details on existing datasets or datasets that will be created during TROMPA for research purposes are provided in current and future versions of **Deliverable D3.2 - Music Description**.

3. User Pilot Repertoire Requirements

Each of the User Pilots will target specific repertoires. An initial set of requirements for the use cases are described in Deliverable 2.1 - Early requirement.

3.1 Music Scholars

The status of the pilot with respect to repertoire, technical requirements has not been changed since the previous version.

3.1.1 Music Scholars Pilot Overview

Music scholars – i.e., those involved in professional or amateur pursuit of deep knowledge about music for its own sake rather than for commercial reasons – are interested in finding connections between music works on multiple levels: from co-occurrences of melodies, harmonic and rhythmic progressions, to the large-scale structural similarities of musical works. The score is the main starting point and anchor for such research. Our aim for this pilot is to provide an interface for the selection and display of musical scores (sheet music) from the TROMPA collections, for annotating them, and for searching them (by text or by example). Thus it relates to many other aspects of the project, in particular the Score Edition component (Task 5.2) of WP5, the TROMPA Contributor Environment. At this stage, we expect the annotations to be nothing more than simple textual comments (which might include hyperlinks using URIs) and at this point shall not have developed any special annotation interface beyond textual linking.

3.1.2 Music Scholars Pilot Repertoire

The repertoire for this version of the music scholars' pilot will mainly comprise early music from the 16th century from resources such as IMSLP (2.2), CPDL (2.3) and Tomás Luis de Victoria (2.16). These have the advantage of relative musical simplicity, and suitability for choral singers (see 3.4.2, footnotes 19 and 23). MEI encodings can easily be obtained by file conversion, and we do not envisage any serious issues with the use of Verovio for rendering the scores in the interface.

[33] Escola Superior de Música de Catalunya (ESMUC) is an associated partner of the TROMPA project
[34] https://zenodo.org/communities/trompa/

3.1.3 Technological and Technical Requirements

The music can be presented in two ways: as (a) PDF graphical images of scores or (b) as graphical renderings using the open-source Javascript library, Verovio, of music encoded in the MEI format (in use throughout the TROMPA project). In many cases, type (a) scores may further be amenable to optical barline recognition (using the open-source program Gamera) which will allow alignment at measure level with an encoded score if one exists; this will be of much value to music scholars. Another kind of type a score would be page images of original manuscripts or printed books (from, say, the 16th century); in the latter case, fairly good encodings can be generated automatically (from good quality images) by optical music recognition using the open-source program Aruspix.

For the final version users should be able to work with the system in private or public, and be able to control the access to their documents and annotations (sets of connections made within a set of scores and/or recordings). However, for this pilot we shall work with whatever level of access control is currently provided by the Contributor Environment.

Users should be able to search the musical contents of scores selected by the user, and in suitably indexed TROMPA collections. Upon selecting a section of a music score, the users should be able to choose a search mode and submit the selected notes as a search query. For this first pilot, the modes will be restricted to either one-dimensional sequences of notes from within a single score voice, or to 'chordal sequences' based on the simultaneously-sounding collections of notes from all the voices or instruments; the length of these sequences will also be constrained for technical reasons, to an extent to be determined during testing.

The search results will be available on a separate page and the request identified by the page's URL, so that users can share the search results page with others. The detailed search request specification and the results will be saved in a suitable format which can be exported by the user if desired. For the final version of the interface, the search functionality will be exposed as a public API, though we do not expect to implement this feature in this version.

3.1.4 Contribution to Public Domain Archives

In the final version we shall provide facilities for digital enquiries by making the search API publicly available and publishing the bulk of data collected in TROMPA as a public linked open data dataset.

3.2 Orchestras

The status of the orchestras pilot with respect to repertoire, technical requirements has not been changed since the previous version of this deliverable.

3.2.1 Orchestras Pilot Overview

In this pilot we will aim to digitize all symphonies by Gustav Mahler, a core repertoire of most orchestras, making them available free of charge. Giving the pilot a kick start, the RCO will have three orchestral scores digitized in Finale through Donemus, a subcontracted publisher. The Fourth Symphony has already been digitized by Donemus during the Phenicx[35] project, so it can be used as well. In this use case, TROMPA will develop crowdsourcing technology to engage music lovers in encoding (out of copyright) scores available as digitized score images from IMSLP. Using RCO orchestra members' and librarians' expertise, TROMPA will develop a tool to extract good quality instrumental parts from these scores. In order to test the technology mentioned above, a chamber

[35] http://phenicx.upf.edu/

ensemble from the RCO will be playing a public domain work (to be chosen from the RCO's chamber music schedule) from a digitized chamber music score generated within the TROMPA project.

3.2.2 Orchestras Pilot Repertoire

As stated in the previous sub-section, the orchestras pilot repertoire will be focus on the Mahler symphonies:

- Symphony No. 1 in D major 'Titan' (4 movements, orchestra only) approx 55 minutes
- Symphony No. 2 in C minor 'Auferstehung' (5 movements, orchestra, choir and vocal soloists (german lyrics)) approx 90 minutes
- Symphony No. 3 in D minor (6 movements, orchestra, children's choir and 1 vocal soloist (german lyrics)) approx 95 minutes
- Symphony No. 4 in G major (4 movements, orchestra and 1 vocal soloist (german lyrics)) approx 56 minutes - already digitized by Donemus
- Symphony No. 5 in C sharp minor (5 movements, orchestra only) approx 70 minutes
- Symphony No. 6 in A minor (4 movements, orchestra only) approx 82 minutes
- Symphony No. 7 in E minor (5 movements, orchestra only) approx 80 minutes
- Symphony No. 8 in E flat major 'Symphonie der Tausend' (2 movements, orchestra, choirs and 8 vocal soloists (german and latin lyrics) approx 80 minutes
- Symphony No. 9 in D major (4 movements, orchestra only) approx. 90 minutes
- Symphony No. 10 in F sharp major: I. Adagio (1 movement, orchestra only) approx. 31 minutes
- Das Lied von der Erde (6 movements, orchestra and 2 vocal soloists (german lyrics)) approx. 63 minutes - optional

The data for these repertoire be will be retrieved by either RCO's content, or by digitizing scanned scores from IMSLP.

3.2.3 Technological and Technical Requirements

The technological and technical requirements of the orchestras use case is twofold. Firstly, scanned score image analysis software will be deployed in order to encode Mahler's music scores in digital format (MEI, or MusicXML). This will be done under the **Task 3.4 Visual Analysis of Scores** in conjunction with the crowd-source annotation tasks in **Work Package 4 - Crowd Annotation and Incentivisation** The second technical requirement is the capability of annotating music scores and share annotations. Feedback from the musicians and orchestra librarian will help getting a good understanding of the specific demands of playable parts. By considering RCO orchestra members' and librarians' expertise feedback, an annotation tool will be developed in **Task 5.2 Digital Score Edition**, with which the musicians will be annotating their parts individually during rehearsal at home, and sharing them among their fellow-musicians before meeting for rehearsal together. During rehearsal together they will annotate individually, and one musician will make annotations that concern all parts. Rehearsing would involve tablets/e-readers in order to be able to make annotations. Performance could be done from either tablets/e-readers or paper, including the printed annotations. Feedback from the musicians will ensure most and possibly all possible standardized types of annotations will become available.

3.2.4 Contribution to Public Domain Archives

All the digitized score encodings of the Mahler Symphonies that will be derived from the use case will be deposited in public domain musical archives. Moreover RCO will offer its most recent annotations of one of the Mahler symphonies for digitization. It can also offer the archive of annotated orchestral scores of Willem Mengelberg, the Concertgebouw Orchestra's chief conductor from 1895 until 1945 for digitization, as well as roughly 95% of the available Mengelberg Concertgebouw recordings, that could be matched to these scores. The RCO offers also all annotated instrumental parts from many Harnoncourt RCO performances, which were so different from normal practice that these parts were kept out of the normal library and therefore remained untouched since the last time they were used with Harnoncourt. Whereas the RCO's normal orchestral scores are used over and over again and current annotations are often the product of many years of performance under different conductors, these Harnoncourt parts are pure representations of one conductor.

3.3 Instrument Players

3.3.1 Instrument Players Overview

This user pilot provides musicians engaging in rehearsal or performance with a "Performance Companion" system capable of characterising performative aspects of their playing – as well as the produced sound – in real-time. By alignment of performance recordings and metadata with musical score encodings, the characteristics (derived features) can be assessed and compared against those derived from other performances, including reference recordings (e.g., a gold-standard studio performance of a particular piece; a user-contributed YouTube recording), or the performer's own previous renditions, allowing the evolution of performance characteristics to be tracked over time. The UI for this pilot implements a specialised view of the TROMPA Digital Score Edition (DSE) component capable of score-following (note highlighting; automated page turning) synchronised both with real-time performance and recorded playback.

3.3.2 Instrument Players Repertoire

Initially, the pilot will focus on pianists performing Beethoven's piano works (primarily his Sonatas, Variation works, pieces, and Concertos). Performance recordings and characterizations produced by the Performance Companion will be captured and published (pending performer permission), contributing to the available repertoire of Beethoven recordings. Toward this goal, we have converted a public collection of kern (humdrum) encodings of every Beethoven Sonata (generated by Craig Sapp) into MEI format, as well as manually generating MEI encodings of several Beethoven piano solo works: 32 Variations in C minor, WoO 80; 15 Variations Op. 35 ("Eroica Variations"); 6 Bagatellen Op. 126; and on the occasion of Clara Schumann's bicentenary on Sept. 13, 2019 her Romanze ohne Opuszahl in a minor. We will work towards producing a complete set of Beethoven piano solo work encodings over the course of the project.

3.3.3 Technological and Technical Requirements

The Music Information Retrieval (WP3) components required for this system can be split into two groups: score alignment and performance characterisation. The score alignment side is responsible for generating metadata descriptions aligning information streams produced during the performance (including audio, MIDI, and/or OSC signals, as well as other performance-related metadata such as

piano key trajectory measurements) with score positions (e.g. by reference to MEI element identifiers). The performance characterisation side is responsible for applying feature extraction algorithms in order to derive quantitative descriptions of specified aspects of the incoming information streams, making them available for (human or automated) assessment and comparison. At minimum, derived features will need to describe audio and MIDI performance streams to produce lower-level (e.g., note onset times and dynamics / velocities) and higher-level (e.g. tempo tracking, timing profiles) characterizations. By reusing standard data models (e.g. Audio Feature Ontology) and associating feature data with provenance information describing the responsible feature extraction algorithm and its version number, development on these MIR components can proceed in parallel with development of the performance companion system's workflow.

Of course, performance characterizations are only valuable to a performer if they can be readily understood and mapped to the performer's in-situ experience in front of the keyboard. The data visualisations and user-interaction mechanisms required to optimally present this sort of information so that it may be positively incorporated to inform future performance remain an open research question. Data from user studies conducted with music students studying at MDW are expected to address such questions and inform the pilot's development.

The use of standard data models, as well as web-addressable representations (using media fragments) opens up both performance information streams and musical score resources, as well as their aligned intersection, to human or algorithmic annotation. Performances employing the Performance Companion will thus become available for annotation in real-time. Foreseeable use-cases include: piano teachers providing pedagogical feedback on students' performance recordings; music enthusiasts referencing or comparing particular performance fragments in community discussion; and music scholars incorporating such fragments to evidence points of scholarly discourse. User annotations will be authored using annotation tooling built into the DSE component. Annotations will be represented within the TROMPA CE using the W3C Web Annotation standard.

3.3.4 Contribution to Public Domain Archives

Score encodings created for the pilot will be made publicly available. Score segmentations, created manually and automatically as part of the score alignment task, will be associated with these encodings and published under open licenses. Finally, performance metadata, public annotations, and recordings (with performer permission) will be made available through the TROMPA CE, with snapshots of public data periodically published on the TROMPA Zenodo community platform.

3.4 Choir Singers

For the choir singers pilot some additional datasets are described that were used for the singing voice synthesis engine. Moreover the file format of the files considered in the pilot are defined.

3.4.1 Choir Singers Overview

The goal of the Choir Singers Pilot is to assist amateur choir singers during individual (solo) performance/practice. It also provides functionality for the choir conductor to create repertoires and to listen to performances by choir members. Users of the pilot will be able to retrieve synthetic versions of existing scores, sing-along with the synthesized voices, and receive feedback (based on

automatic assessment) on their performance. The accompanying voices will be available for music in Spanish, Catalan, Latin, English and German.

3.4.2 Choir Singers Repertoire

The choir singers use case repertoire involves choral pieces in five target languages, namely **Latin**, **Spanish**, **Catalan**, **English** and **German**. Potentially the pilot can work with **any piece** from existing repositories such as CPDL and IMSLP, that is in symbolic format (e.g. musicXML, MIDI, MEI) and in one of the target languages. In the early stage of this pilot, we will initially focus on the following composers and languages. These will be:

- Tomás Luis de Victoria for Spanish[36].
- Anton Bruckner for Italian and Latin[37].
- Josquin des Prez for Latin[38].
- Locus Iste by Anton Bruckner[39] (Latin).
- Niño Dios by Francisco Guerrero[40] (Spanish).
- El Rossinyol, a traditional song[41]. (Catalan).
- Hallelujah by George Frideric Handel[42] (English).

During September 2019, we performed new choir singing recordings with the Francesc Valls choir in Barcelona to create new choir singing synthesis models as defined in **D3.3 - Audio Processing**. The repertoire for these recordings was defined together with the choir (to include songs in their existing repertoire) and includes fragments from the following pieces:

- Jesu meine freude by J.S. Bach (German)
- In the beginning by Copland (English)
- Niño Dios, d'amor herido by Francisco Guerrero (Spanish)
- Glòria a Vós (Spanish, Catalan, English)
- Wer nur den lieben Gott läßt walten by Georg Neumark (German, English, Catalan)
- Formosa sou, Maria by Segarra (Catalan)

These songs will also be included in the repertoire for initial testing of the pilot.

As the pilot progresses the repertoire will be updated and we will synthesize a large amount of pieces that exist in CPDL and IMSLP. For instance, in CPDL, in the category of "4-part choral music", there are ~4626 scores in MusicXML, of which 3267 are either in Latin, Catalan, English or Spanish. We will make the pilot able to deal with any of these pieces or additional ones in digital format provided by the TROMPA choir singers community.

3.4.3 Technological and Technical Requirements

This pilot will be mostly based on the audio processing techniques to be developed in Task 3.3, where we will research and develop techniques for audio synthesis of choir singing. The pilot will collect data from users to improve the voice synthesis algorithms. Users should be able to provide

[36] http://www3.cpdl.org/wiki/index.php/Tom%C3%A1s_Luis_de_Victoria
[37] http://www3.cpdl.org/wiki/index.php/Anton_Bruckner
[38] http://www3.cpdl.org/wiki/index.php/Josquin_des_Prez
[39] http://www3.cpdl.org/wiki/index.php/Locus_iste_(Anton_Bruckner)
[40] http://www2.cpdl.org/wiki/index.php/Niño_Dios_d%27amor_herido_(Francisco_Guerrero)
[41] http://www3.cpdl.org/wiki/index.php/El_Rossinyol_(Traditional)
[42] http://www3.cpdl.org/wiki/index.php/Hallelujah_(from_%27Messiah%27)_(George_Frideric_Handel)

general scores for the synthesis (e.g. by rating the overall quality of the synthesis) and to make timestamped annotations for the generated material, i.e., allowing the user to input free text comments to inform about specific problems (e.g. “this phoneme sounds weird at this point in time in the soprano voice”).

The pilot will be compatible with any piece included in the CE in MusicXML and MEI formats that contain lyrics in English, Spanish, Catalan, German or Latin, provided that the files are correctly formatted[43].

3.4.4 Contribution to Public Domain Archives

Through the use of the pilot, synthesized versions of the scores will be generated and stored, accessible through public URLs. These synthesized versions will be associated to the scores from which they are generated. The same applies for recordings of performances and annotations made by users of the pilot (needed for providing automatic feedback), although in this case, their addition to the repertoire will depend on getting appropriate permission for the user. As discussed in Section 3.4.2 we plan to **synthesize a large amount of choral pieces**; the audio will be deposited to public domain archives and the respective acoustic features to AcousticBrainz. All this information will be linked and accessible through the CE.

3.5 Music Enthusiasts

3.5.1 Music Enthusiasts Overview

In the music enthusiasts use case we will provide interaction mechanisms with musical cultural heritage content targeted at people that although they don’t have a formal musical knowledge, they are interest in learning more about music. The main study of this use case is to build a training tool for emotion recognition in music, focused on classical music, in order to investigate the factors that influence users agreement, generate and evaluate suitability of different recommendation algorithms trained with the collected data, and analyze how the taste for classical music evolves with the interaction with a music recommendation system, as well as aspects of human perception such as emotion.

3.5.2 Music Enthusiasts Repertoire

Music enthusiasts use case will be focused on the existing music collection of CDR Muziekweb, owned by Stichting Centrale Discotheek, which is a member of the TROMPA consortium. We will focus on classical music repertoire of this library, but we will not limit ourselves to that; since we will investigate the interaction of people that are not familiar with classical music, we will also use more commercial repertoire as well. Moreover we will consider a very small and focused repertoire of classical music in order to conduct experiments related to the music emotion/mood and classical music.

[43] In a reduced number of cases, we have observed for example that lyrics are coded using inappropriate fields in the digital score (e.g. because the user who transcribed the song has used annotations, not linked to the note, instead of *lyrics* fields). In this case, it is not possible to automatically match the lyrics to their corresponding note.

3.5.3 Technological and Technical Requirements

The music recommendation system that will be built for the music enthusiasts use will depend on the music similarity and recommendation engine that will be developed in Task 3.2 Music Description (see deliverable D3.2 - Music Description) as well as other descriptors that will be used in this similarity engine. Moreover techniques on higher level semantics extraction, such as emotion classification, will be adopted and evaluated during the use case (Deliverable 3.2).

3.5.4 Contribution to Public Domain Archives

The contribution of this use case will not be new data, but rather new metadata on the CDR repository, such as:

- Tags and annotations of music pieces (musical tags, tags related to style, tags related to mood/emotion).
- Opinion about music excerpts (e.g. ratings)
- Opinion about the outcome of a recommendation systems (e.g. like/dislike)
- Audio descriptors of music pieces.
- Evaluation of music performances

4. Overview of Target Repertoires

In this section we summarize the repertoires contained in the repositories (section 2) in terms of their applicability to each of the use cases (section 3). The following table (Table 4.1) summarizes the target repertoire and the corresponding use cases.

Repertoire	Repository	Data Format
Music Scholars		
Early music from the 16th century	❖ IMSLP ❖ CPDL ❖ Tomás Luis de Victoria ❖ EMO ❖ ECOLM ❖ Biblioteca Digital Hispanica	❖ PDF scores ❖ MEI scores
Orchestras		
Gustav Mahler: Symphonies Nos. 1-9 + Adagio from Symphony No. 10 + (optional) Das Lied von der Erde	❖ RCO private repository ❖ IMSLP	❖ Paper scores ❖ scanned scores ❖ symbolic Scores
Instrument Players		

Beethoven piano works	❖ Humdrum-data repository ❖ IMSLP ❖ YouTube ❖ MusicBrainz metadata ❖ AcousticBrainz metadata ❖ Other potential repositories	❖ symbolic score encodings ❖ scanned score images ❖ audiovisual recordings ❖ bibliographical metadata ❖ audio feature metadata
Choir Singers		
Targeted initial repertoire (see 3.4.2)	❖ CPDL ❖ Tomás Luis de Victoria	❖ MusicXML ❖ MEI ❖ MIDI
Any choral piece in English, Latin, Spanish, Catalan and German.	❖ CPDL ❖ IMSLP	❖ MusicXML ❖ MEI ❖ MIDI
Music Enthusiasts		
Commercial music for music recommendation system	❖ CDR website data ❖ MusicBrainz metadata ❖ AcousticBrainz metadata	❖ Audio files ❖ Metadata information ❖ User profile data
Mahler's / Beethoven's' specific selected repertoire for music emotion on classical music	❖ CDR website data ❖ MusicBrainz metadata ❖ Other potential repositories	❖ Audio files ❖ Metadata information ❖ User profile data

Table 4.1. List of TROMPA use cases repertoire and the corresponding repositories

5. Data Resource Preparation

In this Section we present how TROMPA repertoire data and metadata will be stored in the TROMPA Contributor Environment (CE) as defined in WP5. The contents of this section reflect the advancements made on the CE since M18. In subsection 5.1 we will describe how metadata (artist/work information) is represented and how links to actual data objects (scores, audios, performance, user generated content) are represented in the CE. In subsection 5.2 we will present the procedure that should be followed in order to store data and metadata in the CE.

5.1 Metadata Representation to the Contributor Environment

The CE uses an internal data model that is primarily based on the schema.org[44] ontology schema. It uses well-defined entities to describe the TROMPA resources defined by the deliverables **D2.3 - Technical Requirements and Integration**[45] and **D5.1 - Data Infrastructure v1**.[46] The internal data model of the CE is presented in Figure 5.1. All metadata items that will be stored in the CE will be mapped on this schema.

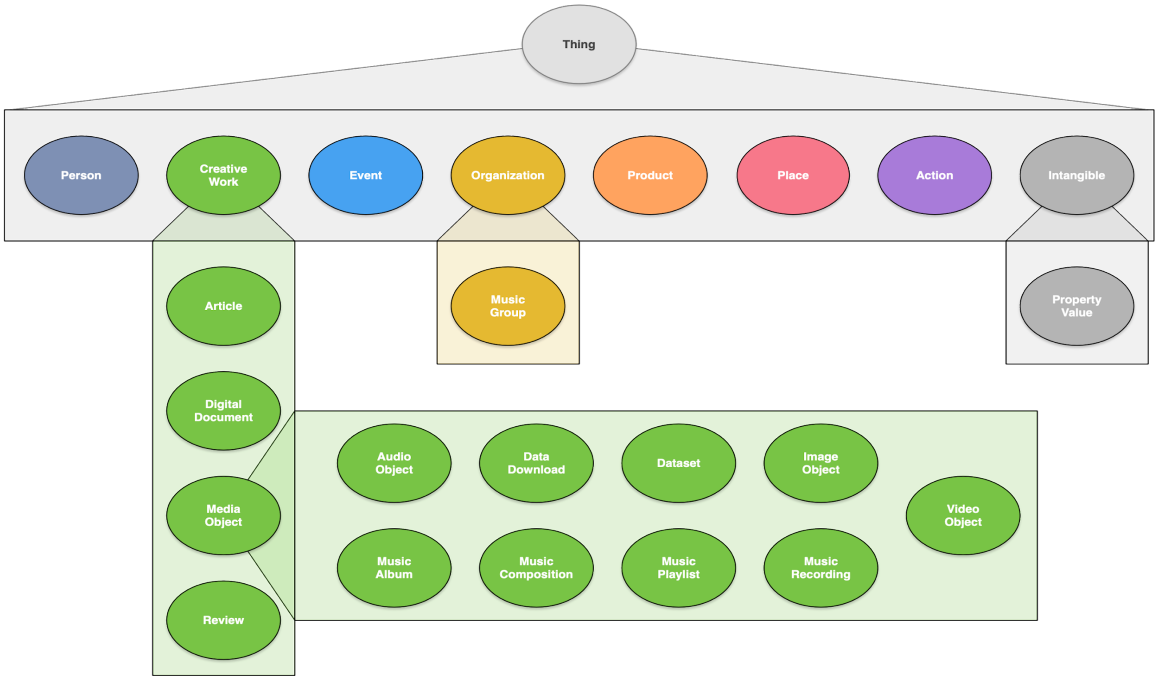

Figure 5.1. Schema.org base types, plus TROMPA relevant extensions

5.2 Storing Target Repertoires in the Contributor Environment

The target repertoires come from many separate repositories as presented in Table 4.1, and may have diverse representation schema. We will store metadata from each target repository in the CE,

[44] https://schema.org
[45] This deliverable is confidential and available only to the consortium.
[46] https://trompamusic.eu/deliverables/TR-D5.1-Data_Infrastructure_v1.pdf

along with any repository-specific identification system that exists for that repository. Each item in the CE will have a uniquely identifying URI. Within the CE we link representations of the same real world entity in different repositories to each other so that information about these entities can be shared regardless of where that information comes from. We make the effort to maintain the MusicBrainz repository in close sync with the data available in the CE, linking this data where possible to the metadata from each target repository. Recently added/updated MusicBrainz data will be added to the CE on a regular basis, and we will make an effort to contribute CE data originating from other repositories and contributors to MusicBrainz. This way we will contribute freely available metadata to existing open databases that can be used by anyone.

Guidelines that describe how metadata can be imported into the CE, and how this data from external repositories can be linked to existing metadata in the CE is being developed. These guidelines are in development and are described in detail in Deliverable D2.3 - Complete Requirements and in the 2nd version of the Deliverable D5.1 - Data Infrastructure[47]

Two tools have been developed until this stage of the project for loading data. One tool for storing MusicBrainz metadata in the CE, and one tool that stores metadata in the CE and links this metadata to actual data (scores, audios, videos, performances) that is stored external repositories. Although these tools are more related to WP5 - Data Infrastructure, for clarity of the representation we describe them in this deliverable. In the next sections we will provide details on these tools

5.2.1 Storing MusicBrainz metadata in the CE

UPF provides a tool[48] to automatically import metadata from MusicBrainz to the CE. This tool takes as an input the MusicBrainz Identifier of a recording or a release that is present in MusicBrainz and imports as much metadata as possible to the CE. The tool includes release, recording, work (movement and overall work), composer, and performer information if it is present in MusicBrainz. This tool does not ensure that the metadata is present or complete in MusicBrainz. It is up to the user importing the data to ensure that the metadata to be imported is present in MusicBrainz. Some aspects of this tool, such as authentication and how frequently it will be run, are still in development and will be clarified in upcoming deliverables (D5.1 - Data Infrastructure) . More information about how to run the importer tool can be found in the source repository for the tool.

5.2.2 Loading metadata and data from other repositories

The CE is designed to store only the metadata for the repertoires that are to be used in TROMPA. Any content that the metadata describes (scores, music recordings, the results of computational algorithms) will be stored in external locations. The CE will contain references to publically available URLs where this content can be obtained from. Where possible, we link to external primary resources as the source of this content. For example, we link to IMSLP to refer to the location of scores of musical works, and could link to YouTube for existing video recordings. It is expected that partners who create additional data, or who require data which is not publically available, to host this data on premises or in an additional storage location provided as a complementary service in the CE (Deliverable 5.1, section 4.5). We provide guidelines[49] that describe how content can be linked to

[47] https://trompamusic.eu/deliverables/TR-5.3-Data_Infastructure_v2.pdf
[48] https://github.com/trompamusic/ce-musicbrainz-import
[49] In Deliverable D2.3 - Complete Requirements

the CE. These guidelines are continually being updated as the development of the CE continues and the requirements of partners become more clear. We also provide sample code to explain to partners how to upload content to the service.

6. Conclusion

This is the final version of Deliverable D3.1 - Data Resource Preparation that describes the repertoire that is exploited in the TROMPA use cases and the corresponding repositories. As shown, there is a large variety of repertoires that exist in several repositories and data formats. These repositories will be the basis upon the use cases are built and will be enriched during the TROMPA use cases. All the repositories will be represented within the Contributor Environment, which will store all the repertoire metadata and the intelinks between the data repositories.

More technical details related to the loading of data to the CE will be given in the upcoming deliverables **D2.3 - Technical Requirements and Integration** (M18, M36) and the final version of **D5.1 - Data Infrastructure** (M30).

7. References

7.1 Written references

[1] Goebl, W. (1999a). Analysis of piano performance: towards a common performance standard? Society of Music Perception and Cognition Conference (SMPC99), Evanston, USA, August 14–17, 1999.

[2] Goebl, W. (1999b). The Vienna 4x22 Piano Corpus, 4 pieces performed by 22 professional pianists, doi:10.21939/4X22 (Open Data, 31 Jan 2017).

[3] Liem, C.C.S. (2018) Music in newspapers: interdisciplinary opportunities and data-related challenges. In Proceedings of the 5th International Conference on Digital Libraries for Musicology (pp. 47-51). ACM.

7.2 List of abbreviations

Abbreviation	Description
IMSLP	International Music Score Library Project
CPDL	Choral Public Domain Library
ECOLM	Electronic Corpus of Lute Music
EMO	Early Music Online
BDH	Biblioteca Digital Hispánica
TLdV	Tomás Luis de Victoria

ISNI	International Standard Name Identifier
API	Application Programming Interface
REST	REpresentational State Transfer
DOI	Document Object Identifier
OCR	Optical Character Recognition
OMR	Optical Music Recognition
CE	Contributor Environment
Partner	**Description**
UPF	University Pompeu Fabra
CDR	Centrale Discotheek Rotterdam
RCO	Royal Concertgebouw Orchestra