TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 6.8

Mid-term evaluation

Grant Agreement nr	770376
Project runtime	May 2018 - April 2021
Document Reference	TR-D6.8-Mid-term Evaluation
Work Package	WP6 - End User Pilots
Deliverable Type	Report
Dissemination Level	PU- Public
Document due date	31 October 2020
Date of submission	4 November 2020
Leader	TUD
Contact Person	Cynthia Liem (c.c.s.liem@tudelft.nl)
Authors	Cynthia Liem (TUD), David Baker (GOLD), Ioannis Petros Samiotis (TUD), David Weigl (MDW), Nicolás Gutiérrez (UPF), Juan Sebastián Gómez (UPF), Maria Pilar Pascual (UPF), Emilia Gómez (UPF)
Reviewers	Aggelos Gkiokas (UPF)

Executive Summary

This deliverable contains the mid-term evaluation report for the prototypes developed in the context of TROMPA's five use cases. The deliverable is a continuation of the work reported in all earlier deliverables of WP6, which moved from mock-ups testing (D6.1), through a global planning (D6.2) for five use case specific prototypes, of which the technical status has been documented in D6.3-6.7. Currently, we present the user-facing evaluation results involving these prototypes. Due to the COVID-19 crisis, conducting user studies has turned out more challenging than initially foreseen: user studies and prototype showcases could not be performed in in-person setups, and overall, scheduling under remote conditions has become more uncertain. This required adaptations in the participant recruitment and study setup strategies, as already discussed in D6.3-6.7. As a consequence, with approval of our PO, this deliverable is published 2 months later than originally foreseen. For each of the use cases, we report on the current status, user-facing evaluation outcomes, and future prospects in a standardised reporting structure.

In the Music Scholars use case, as reported in D6.3, the originally planned physical Mahler annotation showcase in May 2020 could not take place due to the COVID-19 crisis; the transformation of this concept into a fully digital alternative is still ongoing at the time of reporting. However, an internal usability evaluation on the currently delivered annotation component (which will be important for transferring and presenting expert annotations to TROMPA-compatible digital formats) has been conducted with a technology-minded musicologist. This evaluation will lead to a more refined feature set in the upcoming technical iterations, that will further help in guaranteeing the success of a digital showcase.

In the Orchestras use case, following the reviewer-suggested focus shift towards young amateur orchestral players, and the project-wide interest in realising score digitisation in human-in-the-loop hybrid setups (also connecting to work under WP4), a Campaign Manager was delivered under D6.4. Through focus group studies with representatives from multiple student orchestras in The Netherlands, both at the amateur and young professional level, feedback has been obtained on the Campaign Manager, task design, and motivational considerations with regard to campaign participation. With this feedback, new technical iterations with improved and refined functionality are currently being planned, that will lead to a larger collaborative digitisation campaign in December.

In the Instrument Players use case, the rehearsal companion presented in D6.5 has been evaluated with music students at different professional levels. First of all, deeper insight was gained into the participants' rehearsal practice; next to this, feedback was solicited on the current rehearsal companion, and priorities could be set for the next prototype iteration, which will focus on robustness and reliability, and refine current visualisation and user feedback functionalities.

In the Choir Singers use case, the choral rehearsal functionality proposed in D6.6 has been taken to several choirs in the Catalonia area. Choir members were informed about the prototype, information about digitally support needs was obtained (which have become more relevant under the current COVID-19 crisis), and joint efforts have been started to digitise their repertoire, and make it accessible for their everyday practice through the Choir Singers pilot. Current feedback yielded a rich list of functionalities to be added and improved for the upcoming technical iteration.

Finally, for the Music Enthusiasts, the annotation campaigns proposed in D6.7 have now been run in the form of two online contests. We discuss how the current contests were set up, and how the

contest setup is being improved to get the annotations that will be relevant for further human-in-the-loop processing under WP3 and WP4.

Version Log		
#	Date	Description
v0.1	30 October , 2020	First review version sent out
v0.2	3 November, 2020	Major revisions done
v1.0	4 November, 2020	Revised finalised version

Table of Contents

1. Introduction 7

2. Music Scholars 8

2.1. Aim of the evaluation study 10

2.2. Participants 10

2.2.1 Recruitment strategies 10

2.2.2 Participant characteristics 10

2.3. Study protocol 10

2.4. Study evaluation outcomes 11

2.5. Impact on future work 12

3. Orchestras: Workshops with Student Orchestras 12

3.1. Aim of the evaluation study 13

3.1.1 Usability tests with Krashna Musika 14

3.1.2 Usability tests with student orchestras & young professionals 14

3.2. Participants 15

3.2.1 Recruitment strategies for Krashna Musika 15

3.2.1 Recruitment strategies for student orchestras & young professionals 16

3.2.2 Participant characteristics 16

3.3. Study protocol 18

3.3.1 Methodology 18

3.3.2 Predefined discussion points 18

3.3.3 Usability tests 20

3.3.4 Recruitment prospects 20

3.4. Study evaluation outcomes 20

3.5. Impact on future work 22

3.5.1 Motivational aspects 22

3.5.2 Campaign manager changes 22

3.5.3 Task design improvements 23

4. Instrument players 23

4.1. Aim of the evaluation study 23

4.2. Participants 24

4.2.1 Recruitment strategies 24

4.2.2 Participant characteristics 24

4.3. Study protocol 24

4.4. Study evaluation outcomes 25

4.5. Impact on future work 27

5. Choir singers 28

5.1. Aim of the evaluation study 28

5.2. Participants 29
5.2.1 Recruitment strategies 29
5.2.2 Participant characteristics 29
5.3. Study protocol 30
5.4. Study evaluation outcomes 32
5.5. Impact on future work 33
5.5.1 Improvements on user and content management 34
5.5.2 Improvements on score playback and voice recording 34
5.5.3 Improvements on choir synthesis quality 35

6. Music enthusiasts 36
6.1. Aim of the evaluation study 36
6.2. Participants 36
6.2.1 Recruitment strategies 36
6.2.2 Participant characteristics 37
6.3. Study protocol 37
6.4. Study evaluation outcomes 39
6.5. Impact on future work 40

7. Conclusion 43

1. Introduction

In this deliverable, we present the current user-facing evaluation results involving the prototypes developed in the context of TROMPA's five use cases. The deliverable is a continuation of the work reported in all earlier deliverables of WP6, which moved from mock-ups testing (D6.1), through a global planning (D6.2) for five use case specific prototypes, as documented in D6.3-6.7.

As already indicated in D6.3-6.7, the unexpected COVID-19 crisis has had an adversarial impact on timelines and the ability to conduct experiments in the ways that initially were foreseen. On the one hand, as user studies can currently not be run in physical circumstances, the crisis gave a stronger push towards developing prototypes that truly can be run online. At the same time, because of higher uncertainty about audience attention and commitment in crisis times, for conducting the user studies, a more local focus than initially foreseen was followed, in which recruitment strongly targeted entities and communities that were close to the TROMPA members. As a consequence, the TROMPA use cases are presently not yet showing the large-scale engagement that is in the TROMPA ambition, and different use cases are in different stages towards these ambitions. Still, current outcomes have been evaluated with relevant and representative audiences, and towards the closure of the project, the intention still is to show the benefit of TROMPA to audiences that are larger and more diverse than audiences that previously have been engaged with Music Information Retrieval (MIR) outcomes.

In this deliverable, we will report the current evaluation outcomes, general status, and future plans for all of the use cases: the Music Scholars in Chapter 2, the Orchestras in Chapter 3, the Instrumental Players in Chapter 4, the Choral Singers in Chapter 5, and the Music Enthusiasts in Chapter 6. In all cases, we follow the same structure, starting with a general introduction and update, then discussing the general aim of the evaluation studies reported, participant recruitment strategies and characteristics, the study protocol, evaluation outcomes, and the impact of these outcomes on future work. We conclude the deliverable in Chapter 7.

2. Music Scholars

Main effort under the music scholars use case has been geared towards preparing a digital Mahler showcase, in which a Mahler expert would annotate the first 10 pages of Mahler's 4 symphony, these annotations would be converted into a TROMPA-compatible digital format, and then be discussed between different additional Mahler experts.

As described in D6.3, the original intention was for the annotations to be presented and discussed in a physical setting at the 2020 Mahler Festival in Amsterdam. To this end, Mahler expert Paul Banks was recruited to be the initial annotator, and four further experts (Peter Franklin, Jeremy Barham, Alexander Wilfing and Marcel van Tilburg) were asked to respond with comments. Due to the COVID-19 crisis, this event had however been canceled; our initial annotator was struck with the COVID-19 virus himself, needing several months of recovery before being able to return to work. As we already described in D6.3, we chose to move to an online showcase instead, but this first needed for our initial annotator to be back in good health, and for more development to take place on performing and presenting both annotations and the additional discourse in TROMPA-compatible formats.

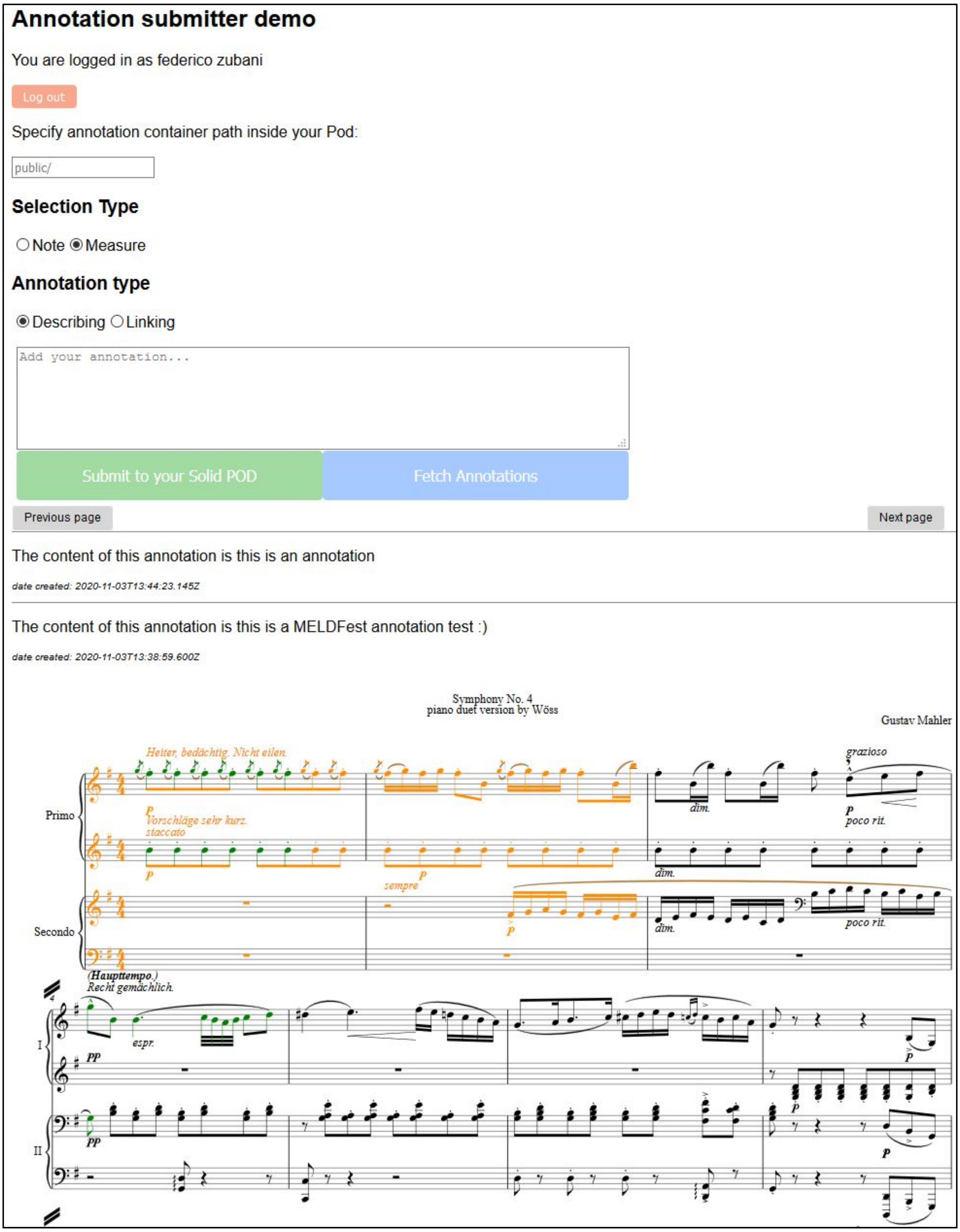
Presently, our initial annotator has provided his annotations in the form of a structured spreadsheet. In parallel, an annotation environment has been delivered, but discourse display is still under development. As a consequence, we have not yet been able to run a full Mahler showcase, and do a fully user-centered evaluation on this.

However, with our annotation environment now existing as a technical Score Edition component to the TROMPA infrastructure, we currently chose to evaluate the current usability of this component in the context of the Music Scholars use case. For this, we conducted an internal review using an expert musicological participant in order to provide feedback both on the usability and musicological relevance of the tools. The goal of this work was to investigate the usability of the current software infrastructure for expert annotation in a multi-modal environment.

Using an open ended response interview format, we presented the current version of an annotation environment that allows music scholars to link commentary on both audio and score-based primary sources within an annotation environment (see the screenshot in Figure 2.1.). The current version implements user-authentication using *Personal Online Datastores* (PODs) as proposed by the W3C Solid Project[1] (as described in deliverable **D6.5-Working Prototype for Instrument Players v1**), which gives users the means to retain complete control of their annotations, allowing them to keep their comments private or to release them publicly for general viewing as desired. Annotations can be made on pre-existing annotations, as a means of generating discussions or evaluations of subjective opinions which may in turn form the basis for scholarly or less formal discourse.

We present a summary of findings from the process here, as well as detailing how understandings from this interview will affect future iterations of the Music Scholars use case. Central to the findings at the mid-term evaluation are requests to be able to have cleaner integration with multi-layer text responses and smoother interfaces while working with both score and audio simultaneously.

[1] https://solidproject.org



Annotation submitter demo

You are logged in as federico zubani

Specify annotation container path inside your Pod:

public/

Selection Type

○ Note ● Measure

Annotation type

● Describing ○ Linking

The content of this annotation is this is an annotation

date created: 2020-11-03T13:44:23.145Z

The content of this annotation is this is a MELDFest annotation test :)

date created: 2020-11-03T13:38:59.600Z

Figure 2.1 Score-annotation interface (development/demo version), showing three annotations: two to sequences of notes (in green), and one to a pair of adjacent whole measures (in orange).

2.1. Aim of the evaluation study

The aim of the present study was to solicit feedback on the usability of the software tools developed in the Music Scholars use case for expert musicological annotations. We sought to better understand what types of annotations were afforded by the current system and to collect feedback as to what future improvements in the current infrastructure would enable researchers working in more humanistic disciplines to find value in the annotation environment.

2.2. Participants

2.2.1 Recruitment strategies

We recruited one expert musicologist new to the TROMPA project from a humanities background, but who has knowledge of software development to comment on the current state of the annotation environment. The expert musicologist had enough familiarity with software development to conduct the review without the use of a guided panel; our goal was to take an open-response approach to soliciting feedback in order to more immediately attend to issues in need of addressing.

2.2.2 Participant characteristics

The research investigation for the Music Scholars use case consisted of one internal expert musicologist who recently joined the TROMPA project. The participant holds a terminal degree from a School of Music in the United States, but is familiar with the cycle of how software is developed. Core parts of his Ph.D. training program focused on understanding formal analysis of 18th and 19th century Western symphonic music (e.g. sonata theory, formal analysis), so he was able to provide suggestions as to provide specific music theoretic suggestions based on contemporary research in American musicological discourse.

2.3. Study protocol

We chose to adopt an open ended approach to gain insights on the expert annotator environment. Though not complete in its final development, we introduced the musicologist to our annotator environment by sending him both the current build of the software, as well as several recordings and scores of the opening of Mahler’s Fourth Symphony and asked him to provide open ended feedback on what he could imagine using the environment for given his experience with other annotation environments.

We chose this approach, as opposed to taking a more guided one, in order to not lead any questions as to any of the original intentions of the platform. One of our main research questions was to investigate if themes that the expert musicologist discussed aligned with the initial intentions of the software development team.

Participant feedback was captured by having the participant use free-text response to note his initial experiences with the software and then subsequently use this response to guide a discussion with the software developers about future feature requests. The discussion between the participant and the team of software developers will then be used to guide future prioritisation of features of the annotation environment.

2.4. Study evaluation outcomes

The study was conducted over a three hour morning session in mid-October. The participant received both the software, the Mahler recordings and scores, and was instructed to provide open response feedback on what types of musicological questions might be aided with the current state of the annotation environment. The participant was encouraged to use the current state of the software as a stepping off point and to use their imagination in order to think of new features that might help future musicological investigation. We summarise the main issues raised by the session in Table 2.1

Category of Reflection	Specific Comments
Current Features	It is helpful to be able to click on specific MEI elements (notes, measures) but need some sort of annotation layer based on formal elements of the score that are not score dependent. Examples include cadential points in alignment with formal theory.
	Clicking score features leading to points of the audio could provide for helpful annotation when discussing how performers handle expressive timing choices
	It would be helpful to expand note/measure MEI elements to have listing of expressive timing terms used by Mahler in the score and link to their timing in the audio. Mahler was known for very explicit annotations in his original score and being able to navigate to listing of where conductors had some sort of artistic license in the performance would allow for easier management of features relating to score/performance analysis (also see Section 2.5, below).
Future Features	Alignment of multiple annotations of cadence points in music, helpful to be able to compare in score major points of disagreement are in formal processes (IAC, Transition themes, demarcations of the trimodular block)
	Page flip coordination with online audio playback
	Create a bank or dictionary of MEI elements that are of interest to performance practice and able to navigate the score audio environment using those elements, rather than notes themselves. This extends to rehearsal markings.

Table 2.1. Participant Responses

Reflecting on the response data, we end this chapter by collating many of the individual points made by the participant into broader categories that will serve as the basis to steer future development.

- ❖ Interface Interaction: Improvement of ability to move between score and audio using annotation level features
- ❖ Query by similarity matching: Once a subset of measures is selected in the score, the audio files are highlighted matching the range in the score. Allow users to then listen to audio within this framework and make, import, export annotations
- ❖ Page Flip: Synchronise page flip annotations across annotation environment

2.5. Impact on future work

Considering the results of the test case, we now consider how the results here will guide both future user studies and project iteration. In terms of future user studies, our next goal will be to carry out a similar user study with more individuals using a guided questionnaire. The questions used here will be derived from the topics generated from the case study presented here. The goal of this research will be to assess how well the annotator environment is able to accommodate some of the specific tasks that were brought to the team's attention in the first round of user testing.

In terms of future project iterations, we plan on continuing development on creating score audio alignment and planning deeper integration with the TROMPA contributor environment. For example, we plan to introduce features --in line with the interview feedback-- that would allow annotators to more easily link an audio file using a URL within the environment to score based MEI features. Creating this infrastructure will also enable the mass importation and sharing of annotations across and within musicological sources.

The development version of the interface as it stands at the time of writing only permits the selection and annotation of notes and whole measures containing notes; this restriction was as planned for the proof of concept and for demonstration purposes. For the Mahler showcase, we shall expand this feature to include some other elements of MEI score-encodings; in particular, tempo indications and other expression marks need to be made selectable. This will allow Mahler’s detailed and highly explicit performance instructions to be linked directly with time-points in a set of aligned audio recordings, so that the responses of different conductors can be directly compared.

In terms of our future timeline, we plan on setting the following goals:

- ❖ Page Flip: Ability to have score and audio alignment in the annotation environment by December of 2020
- ❖ Extra-score Annotation Commenting: Ability to select and navigate the score and audio environment using non-score based MEI elements by February of 2020
- ❖ Query by Similarity Matching: Ability to highlight score and navigate location in the audio tracks, ability to comment, export, import on all selections to TROMPA contributor environment by March 2020

3. Orchestras: Workshops with Student Orchestras

Initially, the Orchestras use case was focused on professional orchestras. We first studied the Content Owners perspective on musical score data, evaluating digital score interaction mock-ups with members of the RCO (**D6.1 - Final Mockups Testing**). However, these studies showed that the RCO’s orchestra members indicated interest in the concept, but no strong intention to practically move over to more digital workflows. Following this observation, feedback from our project reviewers at the first project review, and the general project need for hybrid human-in-the-loop workflows to get digitised scores in the first place (studied under WP4), we have refocused the attention in this use case to younger players (students), also including the amateur music scene. Apart from this shift in audience focus, to better align with the activities under WP4, the Orchestra use case also has been recentered around the process of collaborative, semi-automatic transcription of digital music score information, for which a crowd campaign manager has been developed, as described in deliverable **D6.4 - Working Prototype for Orchestras**.

We conducted the current evaluation studies for the Orchestra Use Case in the form of workshops, with Focus Group Discussion (FGD) and usability test activities. The subject of those workshops was the usage of semantically rich digital music scores, alongside incentivisation factors to participate in online music transcription campaigns. The evaluation study was conducted in two steps: at the end of August, a workshop was held with participants from the Delft student orchestra "Krashna Musika", to evaluate a transcription campaign running that period on the Campaign Manager. These gave insights in the current usability of the campaign manager, and the understandability of the current verification tasks. Beyond the verification tasks, we have been planning further tasks within the digitisation pipeline; to evaluate mockups for these new tasks, but also gauge interest and engagement of orchestras that may not be as technically biased as Krashna Musika, we held a series of workshops with representatives from other Dutch student orchestras.

Both evaluation workshops had similar structure, sharing a common goal to spark discussions with semi-expert orchestra members regarding their use of semantically rich, digital music scores. The participants' selection criteria and their characteristics slightly differed, as did some of the Task Designs and User Interfaces (UI). In this chapter, we will present both steps of our evaluation study showcasing the common aspects between the workshop with *Krashna Musika* and the rest of the workshops, while at the same time emphasising their unique characteristics and goals.

3.1. Aim of the evaluation study

The conducted workshops were organised in order to tackle our main research question: "*How can the crowd be incentivised to enable sustainable and scalable crowdsourced music data annotation creation?*". More specifically, we wanted to evaluate the following:

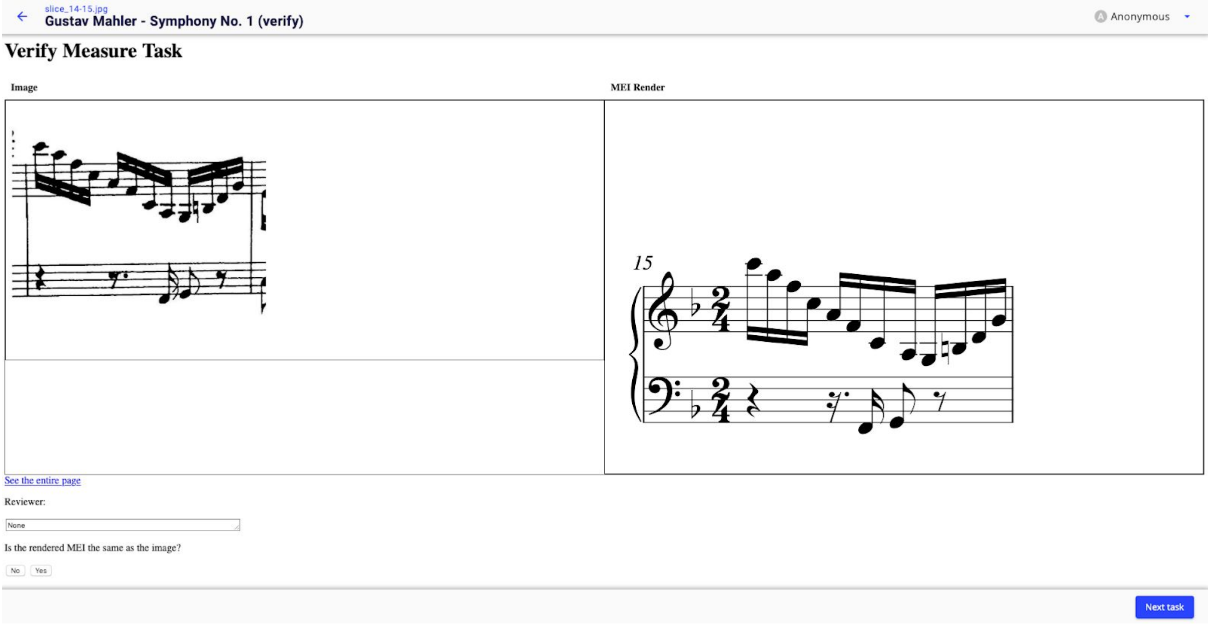
- ❖ How the use of semantically rich digital music scores could benefit orchestras during rehearsing and/or live performance?
- ❖ How crowd-assisted Optical Music Recognition (OMR) campaigns could enable the use of digital music scores?
- ❖ The usability and feasibility of crowd-assisted OMR campaigns
    - ➢ How to improve such campaigns? (motivation of orchestra members to participate and Task Design aspects)
- ❖ How could we motivate the general public (non-experts) successfully, to participate in such campaigns?

With respect to these main points of discussion, we were able to explore alongside the participants, the extent to which they have adopted newer technologies of music transcription and how they use them as members of their respective orchestras. Outcomes of these discussions helped to introduce the on-going efforts of the involved TROMPA partners, on the online crowd-assisted OMR campaigns, which aim to help on digitising music scores.

Finally, following the work of deliverables **D4.2 - Annotator Properties and Metrics** and **D4.4 - Hybrid Annotation Workflows**, we were interested in how the participants envisioned to encourage their colleagues and the general public to participate in transcription campaigns organised by TROMPA, appropriate to their expertise.

3.1.1 Usability tests with Krashna Musika

Through this workshop, we had the opportunity to test an on-going (at that time) TROMPA quality evaluation campaign, where the participants interacted with task interfaces in the Campaign Manager that was delivered in **D6.4 - Working Prototype for Orchestras**, to execute several verification tasks, as illustrated in Figure 3.1. The usability tests and discussions afterwards, focused on the overall experience using the Campaign Manager as a platform for music transcription campaigns, and features they would like to see in such a platform.



slice_14-15.jpg
Gustav Mahler - Symphony No. 1 (verify)

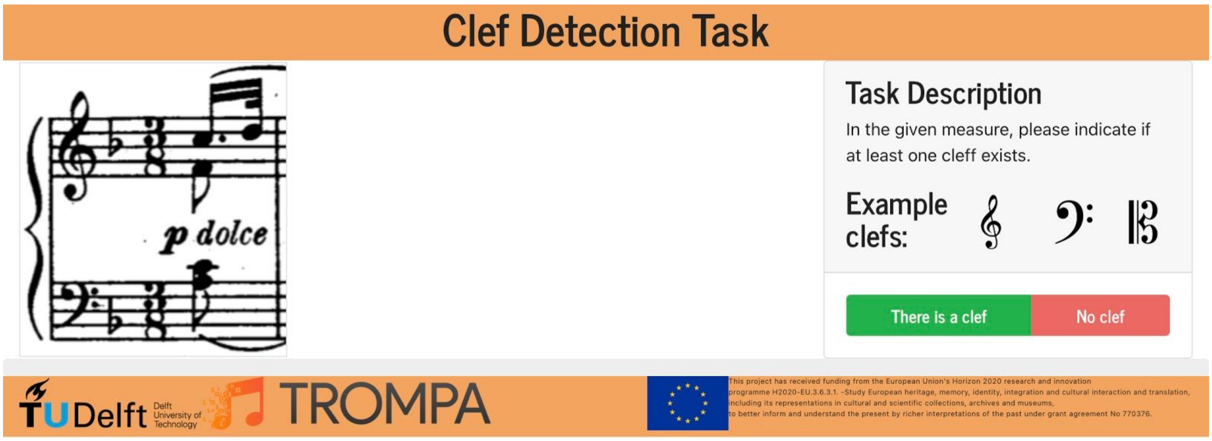
Verify Measure Task

Is the rendered MEI the same as the image?

Figure 3.1. Task Interface from Campaign Manager.

3.1.2 Usability tests with student orchestras & young professionals

In the workshops conducted with other student orchestra members and young professionals, we were able to test new mockups with new task designs. We used usability tests to evaluate these designs for their usability, accessibility and motivational factors of future users of our transcription platform (Figures 3.2, 3.3 and 3.4). Finally, we encouraged the participants to share their ideas and opinions on Task Design and UI elements for online music transcription, to find points where we could improve our design choices.



Clef Detection Task

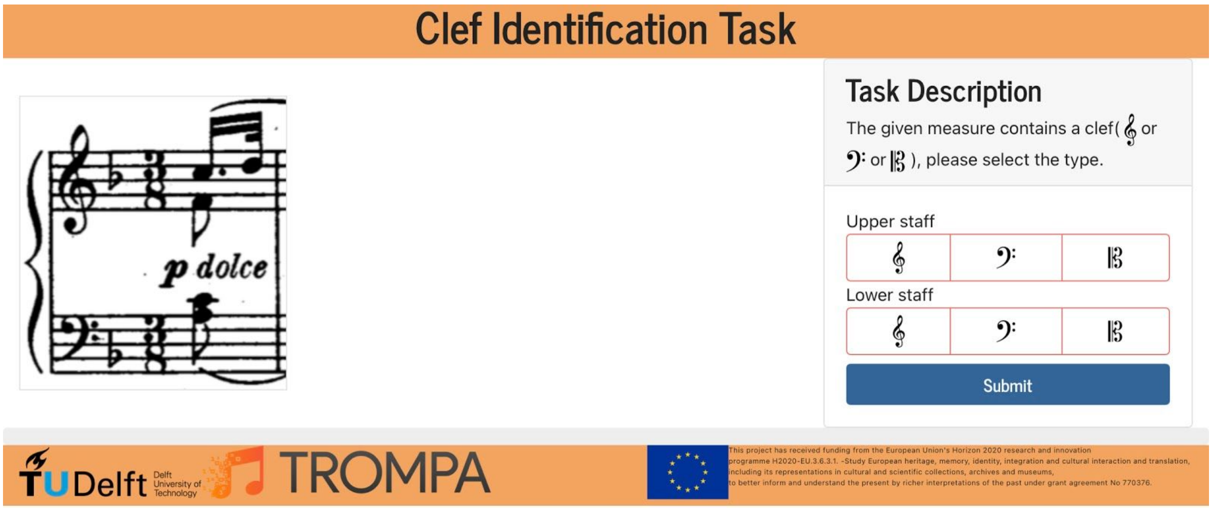
Task Description

In the given measure, please indicate if at least one cleff exists.

Example clefs: 𝄞 𝄢 𝄡

This project has received funding from the European Union's Horizon 2020 research and innovation programme H2020-EU.3.6.3.1. -Study European heritage, memory, identity, integration and cultural interaction and translation, including its representations in cultural and scientific collections, archives and museums, to better inform and understand the present by richer interpretations of the past under grant agreement No 770376.

Figure 3.2. Clef detection task mockup.



Clef Identification Task

Task Description

The given measure contains a clef(𝄞 or 𝄢 or 𝄡), please select the type.

Upper staff

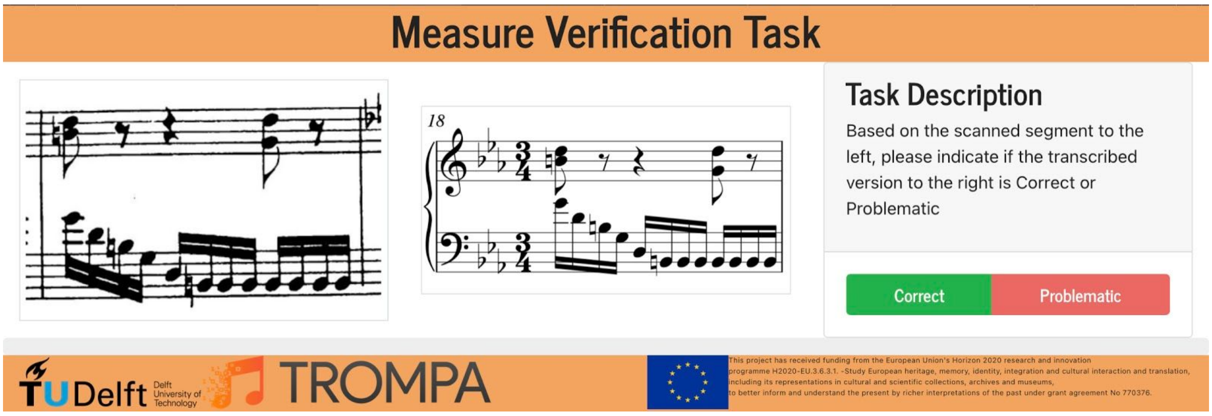
𝄞	𝄢	𝄡

Lower staff

𝄞	𝄢	𝄡

This project has received funding from the European Union's Horizon 2020 research and innovation programme H2020-EU.3.6.3.1. -Study European heritage, memory, identity, integration and cultural interaction and translation, including its representations in cultural and scientific collections, archives and museums, to better inform and understand the present by richer interpretations of the past under grant agreement No 770376.

Figure 3.3. Clef identification task mockup.



Measure Verification Task

Task Description

Based on the scanned segment to the left, please indicate if the transcribed version to the right is Correct or Problematic

This project has received funding from the European Union's Horizon 2020 research and innovation programme H2020-EU.3.6.3.1. -Study European heritage, memory, identity, integration and cultural interaction and translation, including its representations in cultural and scientific collections, archives and museums, to better inform and understand the present by richer interpretations of the past under grant agreement No 770376.

Figure 3.4. Measure verification task mockup.

3.2. Participants

3.2.1 Recruitment strategies for Krashna Musika

Krashna Musika is a music association associated with Delft University of Technology, having a symphony orchestra, a choir and a chamber music department. As such, members of Krashna Musika are closely involved with the Delft student population, meaning that they are both musically and technically inclined, and are in relative proximity to several of the TROMPA investigators. As such, we wished to start our first exploratory study with members of Krashna's orchestra. In close cooperation with a former board member of the Krashna Musika, a group of 8 participants was recruited, and the workshop was held on August 31, 2020.

3.2.1 Recruitment strategies for student orchestras & young professionals

With the help of the academy coordinator of the RCO, we reached out to the boards of all student orchestras in The Netherlands, as well as the board of Het Nationaal Jeugdorkest (NJO), two project orchestras for conservatoire students. While the direct aim was to recruit participants for the current workshop rounds, at the same time, we also wanted to already cast a broader net, to get a broader target audience interested, also for future workshops and evaluation rounds.

The choice was to both recruit with student orchestras (where players largely are musical amateurs), as well as the NJO (where players are young professionals in music instrument majors); the first audience does not engage with music professionally, but also has less professional infrastructure as a consequence, meaning that TROMPA technologies contributing to the public domain may be of direct benefit to their practice. Furthermore, student orchestras also have an important social component, which would help in stimulating recruitments in groups. At the same time, young professionals are future players of professional orchestras, which may be more open to innovation and digital technology than very established players. Therefore, we also found it important to engage them and get their feedback.

The orchestras were first mailed through their official contact addresses, with an accompanying letter explaining the general goals of TROMPA, and a request for their willingness to help us recruiting participants that could join pre-scheduled workshops on October 19 or 20, 2020. Following this, several orchestras (*Collegium Musicum, Quadrivium, NJO, Sweelinck, Nijmeegs Studentenorkest CMC, Amsterdams Studenten Orkest,* and *S. M. G. 'Sempre Crescendo'*) came back with multiple available members, often including the contacted board members themselves. As a consequence, on October 19 and 20, 2020, we ran 5 workshops with 30 participants in total. Orchestras that indicated interest, but could not make these workshops, have been added to a list of interested parties, and will be re-contacted for future studies.

3.2.2 Participant characteristics

We retrieved occupation and the level of music expertise of the participants through a selection of questions from the Goldsmiths Musical Sophistication Index (Gold-MSI). The compiled form of questions was:

- ❖ Please fill in your current occupation
- ❖ I have had formal training in music theory for __ years

- I have had __ years of formal training on a musical instrument (including voice) during my lifetime.
- I can play ___ musical instruments
- The instrument I play best (including voice) is ____
- I have experience designing User Interfaces

As expected, the participants' musical expertise was high, and we covered a wide spectrum of instrument specialty (14 different instruments in total). Most of the participants (83.3%) had no previous experience designing User Interfaces. More specifically, on Figures 3.5 and 3.6, we see that most participants had more than 3 years of music theory training and more than 10 years of formal instrument training. Finally, on Figure 3.7, we see a high versatility in instrument performance, with the majority of the participants being able to play more than 2 instruments.

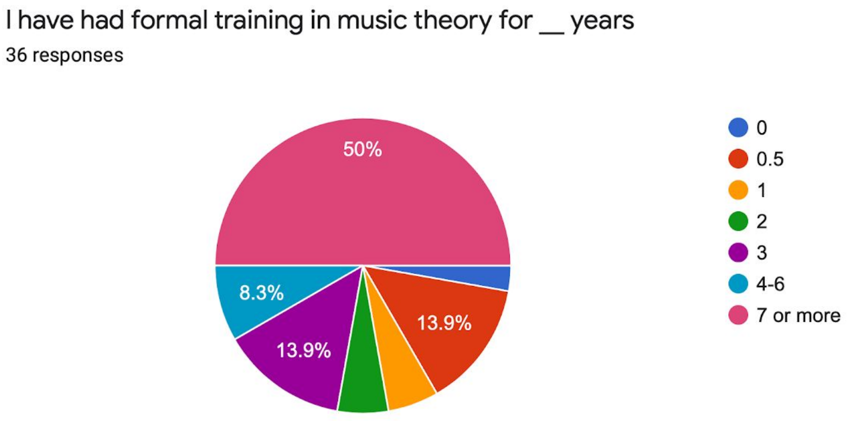


Figure 3.5. Participants' years of training in music theory.

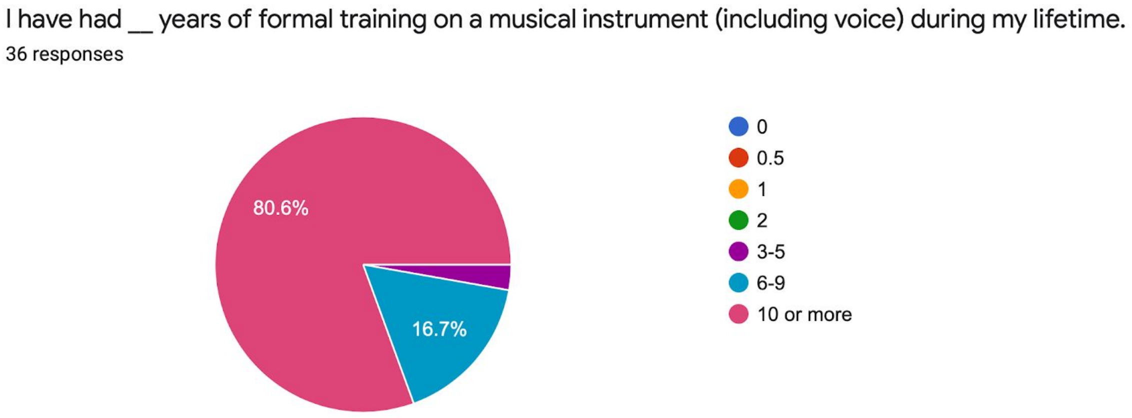


Figure 3.6. Participants' years of training on a musical instrument.

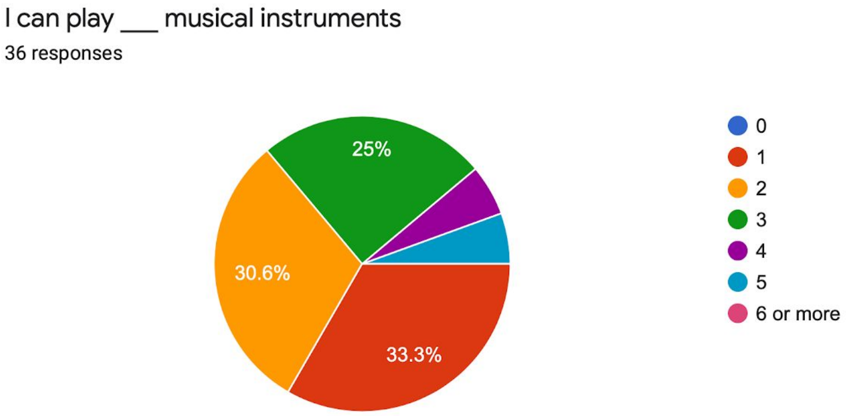


Figure 3.7. Number of instruments each participant could perform with.

While participants would introduce themselves in the workshops, no further detailed demographics were officially solicited; in the case of these current studies, we did not consider them to be of key value to our analyses, and thus followed our institute's general ethics guidelines to then not ask for these.

3.3. Study protocol

In this section, we describe the methods and protocols we used in our evaluation study. Due to the safety measures against COVID-19, that were present during our study, we conducted all interviews and discussions through the online video conference tool *Zoom*.

3.3.1 Methodology

Our study's goals were to gain insights on how and to what extent student orchestras use digital scores in their rehearsals and performances, but also their familiarity with digital tools for music transcription. To that end, we selected the FGD methodology to collect qualitative data on the topics of our discussions, which will later help us understand how to better conduct online transcription campaigns. During the workshops, there were always two researchers present. One acted as the facilitator of the discussions and was aware of challenges that student orchestras might face, empathising better with the participants, while the second acted as the notetaker. Both co-curated the discussions and the FGD were conducted in English.

Due to the diverse technical background of our participants in both parts of our study, we curated the discussion points in a way to find how familiar they were with digital means to access and edit their selected music scores. This helped us to adjust the extent to which we discussed the technical aspects of our work.

Before the start of each workshop, we handed each participant a consent form, where they were informed about the discussion notes we would log and the option to record our session. We made explicit that any data will be treated confidentially and that if one participant wouldn't feel comfortable to record the session, we would not do so. They were also free to withdraw their participation at any point, while they could choose to participate with video or not.

During the FGD, we demonstrated to the participants online interfaces, with which they were encouraged to interact and voice their opinions and discuss points to improve on their design. To measure the users' perceived satisfaction, we distributed the Post-Study System Usability Questionnaire (PSSUQ) to all the participants. Answers to all the forms alongside all of the discussion notes, were treated anonymously without any identifying features.

Even though the study wasn't conducted in person, we believe that it didn't hinder its effectiveness. Participants were already familiar with online video calls at the time of the study and we were able to conduct several workshops per day with participants who would otherwise be more difficult to gather together in a physical meeting point. Finally, the systems and mockups used were all available online, making them easily accessible by all the participants through their device of choice.

3.3.2 Predefined discussion points

For the purpose of our evaluation study, we conducted all workshops using the same outline of discussion points. As mentioned before, we adjusted the extent of technical details discussed, based on the background of the participants per workshop. The list below represents the main discussion points used as a guide during the workshops.

To assess the familiarity of the participants with digital music scores and discuss their use of them alongside digital transcription tools, we used the following questions:

- How familiar are you with semantically rich music scores?
- Do you use this kind of digital scores personally?
  - If yes: When (e.g. during practice) and/or how (e.g. tablet)?
  - If no: Why and what would make them more appealing?
- Do you use digital scores as an orchestra?
  - If yes: How do you incorporate them?
  - If no: Have you thought about using them? What are the reasons you haven't so far?
- How could/(or already do) you benefit from using digital scores?

Since the work for the Orchestra Use Case in TROMPA is focused on providing high-quality digital transcriptions of music scores, we wanted to find how familiar were the participants with Optical Music Recognition (OMR) workflows and tools. We discussed with them the following points, in order to showcase why it is a hard problem and what is the state-of-the-art.

- How we want to incorporate the crowd into OMR processes;
- Current state and limits of research.

We followed our discussions on OMR, with how the TROMPA project is working to improve OMR workflows by incorporating human-in-the-loop solutions. We discussed the transcription and improvement campaigns that we are planning and how they are organised. Finally, we explored the concept of crowdsourcing and how it can be applied on music score transcription workflows.

With the above discussion points, we set the scene to conduct our usability tests, which differed between the first workshop with Krashna Musika (when students were interacting with a running verification campaign in the campaign manager) and the rest of the workshops (when students were interacting with several mockups on a more diversified set of tasks: clef detection, clef identification, verification). Students were provided access to the relevant pages through links, and afterwards were asked to complete a usability survey, as presented in Section 3.3.3.

As through our workshop with Krashna Musika, we wanted to evaluate the quality of a transcribed music score, after the usability tests, we followed with:

- Discussions about the crowd manager platform (opinions, ideas on how we can improve it etc).

For the workshops with student orchestra members and young professionals, rather than discussing the crowd manager, we followed our usability tests on the provided mockups with:

- Discussions about the task mockups (opinions, ideas on how we can improve it etc).

After this, participants had a better sense of what could happen within a campaign, and we continued to discuss motivational considerations to take into account, when seeking to run such campaigns as part of the practice of the participants:

- How such a campaign could benefit their orchestras;
- How could they see themselves motivated to participate in such a campaign;
    - What about campaigns from other orchestras;
- How could the general public (non-experts) help such campaigns;
    - Types of tasks they believe they could successfully do.
- What could be a satisfying final product coming from such a campaign?
    - What music score elements could be tolerated to be missing?
    - In case of less than 100% coverage, what extent of transcription coverage would be satisfying enough?

3.3.3 Usability tests

As explained before in this chapter, we followed different usability tests on the workshop with *Krashna Musika*, which focused on testing the Campaign Manager and an existing transcription campaign; while during our workshops with other student orchestras and young professionals, we conducted usability tests on different Task designs. In all workshops, we measured the users' perceived satisfaction while using the provided interfaces, using the Post-Study System Usability Questionnaire (PSSUQ), distributed to all the participants. More specifically, the questionnaire contained the following questions:

- Overall, I am satisfied with how easy it is to use this system.
- I was able to complete the tasks and scenarios quickly using this system.
- I felt comfortable using this system.
- It was easy to learn to use this system.
- The system gave error messages that clearly told me how to fix problems.
- Whenever I made a mistake using the system, I could recover easily and quickly.
- The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
- It was easy to find the information I needed.
- The information was effective in helping me complete the tasks and scenarios.
- The organisation of information on the system screens was clear.
- The interface of this system was pleasant.
- I liked using the interface of this system.
- This system has all the functions and capabilities I expect it to have.

- Overall, I am satisfied with this system.

3.3.4 Recruitment prospects

Finally, we ended all the workshops with showcasing our plans for the TROMPA transcription campaigns, and sought to motivate and encourage participants to also consider participating in future campaigns and user studies, and help us with recruiting further participants from their own circles, as soon as new studies will be conducted.

Furthermore, as a token of gratitude for their time (with a workshop lasting 2h30 on average), participants were offered the choice between a membership to Entrée, the RCO's youth audience association, or an RCO CD.

3.4. Study evaluation outcomes

Considering familiarity with digital, semantically rich scores, participants had little familiarity with these, although several participants have been active with making transcriptions for themselves in programs such as Sibelius or Musescore. As the Krashna Musika participants were more biased towards engineering studies, and recruited by a member who is a computer science major, had programming experience and knew XML. This was not the case for the members for other orchestras, although one participant had once tried working with Lilypond encodings (but found it too difficult to work with those).

In their practice, participants commonly have been using scores from the IMSLP Petrucci Music Library, but mostly for printing them in physical form. Playing from and annotating on digital devices still is uncommon; as one participant indicated, annotation interaction through a digital device currently still seems to be slower than annotating with pencil on paper, which makes the integration of a digital device into daily practice less attractive.

Several participants had tried using OMR functionalities to convert PDFs into digital scores, but none of the participants had been successful with this. As such, current issues with OMR were recognised, and participants understood the concept behind hybrid digitisation workflows as proposed by us.

We found that overall, participants were satisfied with our Task Designs for both the on-going transcription campaign hosted on Campaign Manager, and the mockups for up-coming tasks (Figure 3.8). Beyond the questionnaire, organisers and participants had in-depth discussions on aspects of the different UIs, as well as what could be improved to make the user experience better and types of tasks that could help in the online music transcription. Out of these discussions, it did become clear that the verification task will need more explicit instructions on when a match between two fragments is good enough to be considered 'the same': especially the higher-expertise musicians (e.g. the NJO players) turned out sensitive to editorial differences (e.g. the fragments seemingly being at different positions in a score system).

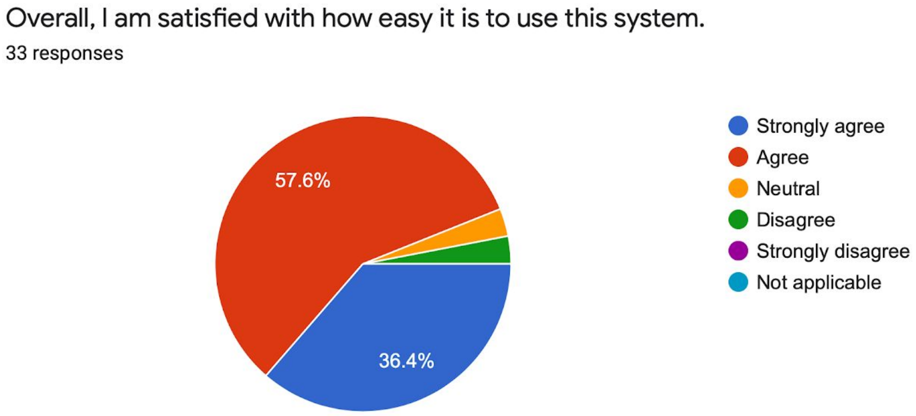


Figure 3.8. Participants' satisfaction with the provided interfaces.

As for motivational considerations, amateur and professional players come from different motivations and settings. Amateur student orchestras include a strong social component, and large volunteering willingness (also in their close social circles) to support the organisation of their orchestras. As such, they believed campaigns could work for their members, especially if they would involve tasks that can quickly be completed, and e.g. can be done on mobile devices. They also considered it feasible to not only engage their own players, but also friends and family wishing to support the orchestra, who e.g. would not be in the financial capacity to do this through financial donations. Furthermore, some willingness was indicated regarding supporting other orchestra's campaigns.

Young professionals were a bit more skeptical about participation willingness of their peers, which is in line with reluctance we noticed earlier with the RCO regarding the changing of existing workflows and practices. At the same time, if tasks would be well-integrated as small units that can quickly be done, without crossing music and rehearsal practice, they believed participation from more orchestra members may be possible, and indicated a high willingness to participate in future studies themselves. As for possible rewards for their participation, the young professionals indicated interest in 'premium' expert content surrounding the works they would help digitising; for example, a first edition or manuscript of the score, or annotations by famous players.

As for what would be a satisfactory end result, unless a score is fully digitised, it will not be usable as an edition that can be used in concert practice. Therefore, pushing for full digitisation is important. Furthermore, the participants indicated it would be useful to see progress.

3.5. Impact on future work

The impact of our evaluation study on the Orchestra use case is threefold. Through our workshops, participants seemed positive to participate in future TROMPA transcription and improvement campaigns. Their ideas and opinions had a catalytic effect on designing new system requirements and features for TROMPA's Campaign Manager. Finally, through their feedback, we were able to spot problems on our Task Design and UI and inspired us to bring new elements that can better integrate human feedback in our automated methods.

3.5.1 Motivational aspects

As mentioned in Section 3.3.4, at the end of our current evaluation sessions the participants were asked to consider participating in future TROMPA campaigns. Their reactions and overall sentiment throughout the workshops, indicate a strong possibility that participants of this study will also engage in future studies and online campaigns.

Throughout the workshops, there were several participants that were already familiar with digital music scores and digital transcription tools and shared the sentiment that methods so far have not yielded the results they would like. Participants who are frequently transcribing scores for their orchestras or during their studies, have found the process rather laborious and they seemed interested to better understand how Optical Music Recognition works and the challenges of the field.

After showcasing the work so far in TROMPA and the plans we have on music transcription campaigns, participants were discussing ways on how to make these campaigns more engaging and how they could meaningfully impact their way of rehearsing and sharing annotations. As discussed above, motivation and reward mechanisms may need to be differently designed for amateur and professional players. As for the young professionals who indicated interest in learning about expert annotations as a reward to task participation, we also presented the Music Scholars use case to them, and agreed we would also connect them as possible audiences to this use case.

3.5.2 Campaign manager changes

Through the workshop with Krashna Musika, we tested the design of TROMPA Campaign Manager. The feedback was very valuable and helped to understand how users, like the participants, would like to interact with the platform. For example, many users commented on the absence of mobile-friendly interfaces and how they could increase the likelihood of them contributing to the tasks hosted on the Campaign Manager. In general, their feedback helped to form the next iteration of the Campaign Manager, so we can integrate the proper features, making the platform more accessible and appealing to users.

In November, the Campaign Manager will be updated to be able to serve mobile-friendly tasks, and to support sequential digitisation through different categories of tasks. To be able to have a better-understandable means of showing progress towards completion, and increase the odds of completion of usable units, we will also schedule tasks such that they progress from page to page, rather than being served randomly throughout the score.

3.5.3 Task design improvements

Evaluating the designs of our tasks was one of our main goals of our study. Designing crowdsourcing tasks for music transcription, which will help Optical Music Recognition workflows, is a novel field of study, meaning that there are a lot of unknowns to research upon.

In our workshops with student orchestras and young professionals, we were able to test task mockups and measure their usability. The participants' input was of great value, since we were able to see how they interact with the UI, which parts they didn't like and ideas they had for improvements. This feedback became one of our main guides for future task designs, having impact on how we think users interact with the UI, what would assist them in their task and what could potentially increase their performance while executing them.

4. Instrument players

We have conducted a user study with students (piano majors) at the University of Music and Performing Arts Vienna (mdw), comprising of structured interviews into overall piano rehearsal habits and strategies, the potential role of digital tooling, and implications for pedagogical contexts (*with* or *as* a teacher). The interviews featured a demonstration of the Companion for Long-term Analyses of Rehearsal Attempts (CLARA) prototype at current state of development, featuring note-level multi-rehearsal alignment, multimodal navigation through rehearsal recordings, visualisation of tempo curves, dynamics, and performance errors, and Solid integration, as described in deliverable **D6.5-Working Prototype for Instrument Players v1**.

4.1. Aim of the evaluation study

The aims of the study are:

- ❖ To validate the user requirements gathered by means of a digital mock-up in our pilot study (**D6.1-Final Mockups Testing**)
- ❖ To evaluate the utility and suitability of the functionalities implemented in the current prototype (D6.5-v1), which were developed based on initial user testing with non-functional mockups (see Section 3 of D6.5-v1, and D6.1), and
- ❖ To guide further development effort in order to arrive at a maximally useful application by the end of the project.

4.2. Participants

4.2.1 Recruitment strategies

This prototype is focused on a target audience of musicians with advanced expertise in classical piano performance. As such, we have chosen to recruit student pianists (piano majors) at the University of Music and Performing Arts Vienna (mdw), through electronic postings on mailing lists, mdw social media accounts, and through physical posters placed around campus.

4.2.2 Participant characteristics

Five pianists have participated in the present phase of the study (evaluating the prototype as described in D6.5-v1) at time of writing. A further five participants have been recruited to complete the study in early November. As part of the post-interview questionnaire (Sec. 4.3) we are additionally inviting each participant whether they would consent to participation in the next phase of user studies (to be reported in the deliverable **D6.9-Final Evaluation**), which will focus on interactive use of the system in an experimental rehearsal session. Students tested thus far exhibit a range of expertise, including one Bachelors student in piano performance; two students with Masters degrees in piano performance and one with an MA pursuing further postgraduate studies; and one student with a Dr. artium (artistic doctorate) pursuing further specialised postgraduate study in chamber music. The participants spend between 20 and 40 hours on piano practice in a typical week (mean: 28.2, SD: 10.8), with a maximum of 28 to 55 hours (mean: 39.6, SD: 11.1). This diversity of experience corresponds to the scope of user audience envisioned for our application.

4.3. Study protocol

Due to the ongoing global pandemic, each session of this study was conducted remotely using the Zoom teleconferencing platform. Students participated in the sessions individually. Each session lasted approximately 1 hour, and involved experimenters David M. Weigl (CLARA developer, mdw) and Werner Goebl (pianist and performance scientist, mdw), alongside the participant. The interview was conducted primarily by D. M. Weigl, with additional contextualising and clarifying questions by W. Goebl applying piano performance domain knowledge.

At the beginning of each session, each participant was emailed three documents: an information sheet, a consent form, and a questionnaire (in English or German depending on participant preference). Participants were asked to read through the information sheet and consent form (see D1.3), and the terms of their participation were clarified. Upon consent, Zoom recordings were started to capture the session for transcription purposes. These recordings will not be shared beyond the two researchers involved in this study, and will be deleted after transcription is completed.

Participants were guided through a series of questions on their rehearsal practice, focusing on the following subjects:

1. A general description of their *rehearsal strategies*, both when initially learning a piece, and when rehearsing a piece for performance.
2. The *context* of their rehearsal sessions – reflecting on how often, how long, when (time, weekday, context in terms of daily routine), and where (e.g., university, practice room, at home) they rehearse, and whether / how it makes a difference.
3. The *purpose* of rehearsal – what is being practiced? One or many pieces; whole pieces, or sections? Which repertoire, and how is this decided? Who guides the rehearsal – the pianist, or a teacher? Does it make a difference? Are specific objectives followed during rehearsal? Which (concrete examples)?
4. *Rehearsal activity* – what happens during rehearsal? Are rehearsals recorded? Are annotations made? Are digital tools used? If so, which? What properties must such tools have or not have, in order to be used? What can digital tools offer? What's currently missing?

The discussion on the above four points typically lasts about 30 minutes. At this stage, the CLARA prototype is demonstrated via Zoom screen-sharing, using some example rehearsal takes of a Clara Schumann piece recorded by W. Goebl for demonstration purposes. Participants are walked through each feature of the prototype, starting with a view of the rendered score; the selection and playback of rehearsal recordings (demonstrating score-alignment through highlighting; dynamics and error visualisation, based on highlight colour; navigation of rehearsal recordings by clicking on score elements, or selecting larger structural segments from a drop-down menu; automatic and manual page turning; and finally, tempo curve visualisation, and navigation within and between rehearsal takes using tempo curves. Concepts around data ownership and sharing are briefly explained – that the data behind each rehearsal take is private by default, but that selected takes can be shared with specified others (e.g., teachers, colleagues) or made public, and that similar facilities are envisioned for score annotations.

Participants are given the opportunity to ask clarifying questions, and are then asked to reflect and provide honest feedback on the utility of the demonstrated prototype in light of the preceding discussion.

The session is then concluded with a final series of questions around pedagogical contexts:

5. *Pedagogical context* – could you envision using such a tool with your teacher? Could you envision yourself using such a tool when teaching your own students? What would be important in such uses? What properties would be required or need to be avoided?

Participants are then asked to return their filled in forms via email at their earliest convenience after the study is concluded. Upon receipt of their forms, they are sent a €20 voucher for Thalia, an Austrian highstreet and online bookshop chain, in exchange for their participation.

4.4. Study evaluation outcomes

A full evaluation incorporating responses from all 10 scheduled participants will be used to inform subsequent development as well as the next set of users studies (incorporating interactive rehearsal sessions), and will be reported in future deliverables (D6.5-v2; D6.9). Here, we summarise preliminary findings based on the initial batch of five participants tested.

- ❖ On *rehearsal strategy*, respondents differed in their reported initial approaches, with three indicating that they study the score of a piece (one incorporating listening to others' recordings) before beginning rehearsal, while the others began rehearsal renditions right away. Every participant mentioned the **annotation of fingerings** at a very early stage of rehearsal preparations. All but one participant mentioned practicing in a **slowed tempo** and building up speed to that anticipated for performance as rehearsals progress. In addition, one participant also mentioned playing slow pieces at an **increased tempo** and gradually slowing down to anticipated performance levels. One participant explicitly mentioned the **capture and study of performance recordings** at this stage, particularly to check for missed directives (articulations, dynamics). Two participants mentioned the benefits of performing rehearsal renditions in front of (one or several) others before a public performance, with one explicitly mentioning the effectiveness of **"simulated" audiences** (via video recording or Zoom call), which have become more relevant in the current pandemic situation.
- ❖ On *rehearsal context*, responses on typical daily **rehearsal hours** varied between 5 and 7, either in one session or split into two (morning and afternoon). Rehearsal **location** played a strong role for some participants and a weaker role with others – one participant asserting that practicing at home or at a dedicated rehearsal space at University makes little difference, others describing differences in rehearsal strategy (e.g., a greater need for focus) when practicing at home, and one participant not owning their own piano and thus relying entirely on rehearsals in dedicated practice rooms. Each participant reported significant disruptions to their pre-pandemic routines in the current situation. Three participants reported access to **electronic (MIDI-capable) instruments** at home. The affordances of silent practice were outlined as advantageous by several participants (lack of neighbourhood disturbance; ability to validate and rehearse "muscle memory" and knowledge of the piece), while others regarded pianos that do not feel or sound like the grand pianos typically used during performance and during practice at University as inadequate for rehearsal purposes.
- ❖ On *purpose* of rehearsal sessions, responses differed, with one student reporting typical deep, focussed concentration on a single piece over the course of a 7 hour rehearsal session, while others typically rehearsed several pieces, often guided by the programming of upcoming concert or competition events. One participant explicitly reported to never focus

on just one piece. Each participant reported focussing on **specific sections** as well as on full rehearsal run-throughs, with one reporting that the former frequently turn into the latter (i.e., what was intended to be a rehearsal of a certain section just carries on to the end of the piece). Each participant reported some mixture of elements in terms of who decides **repertoire** for rehearsal – the pianist through personal choice, or an external factor (e.g., the teacher, an upcoming competition) – interestingly with varying effects on motivation, with one participant reporting motivation almost exclusively with self-chosen pieces, and another reporting the opposite. Two participants reported rehearsing with no explicit **goals** in mind (other than general progress: "I rehearse what I'm not happy with"; "I want to play it better than I did yesterday"). Others report specific goals, e.g.: "be loyal to the text" [incorporate specific metronome markings, articulations or dynamics as written in the score]; "I want to play through and annotate fingerings on the first three score pages"; "play separate hands".

- ❖ On *rehearsal activity* – all participants **record** their playing at least occasionally; one using recordings routinely during rehearsal, and three others stating they "should" record themselves more often [because such recordings are deemed helpful]. One participant explicitly does not **revisit recordings** after immediate review, whereas another does so frequently, even consulting recordings from two years ago when a piece is rehearsed again after a long pause. The **utility of recordings** is appreciated by each participant. **Annotation of the score** during rehearsal includes **fingerings** (all participants), **metronome indications** (one participant), **circling of notes** (two participants), **key indications** (one participant), and **structural annotations** (e.g., divisions, patterns in virtuosic passages; one participant). One participant mentions not writing anything beside fingerings, but incorporating annotations (e.g., **dynamic markings**) provided by the **teacher**. Another reports writing **short descriptive notes** outside of the **rehearsal context** (e.g., studying the score on a train ride). Three participants incorporate **digital scores** (displayed on an iPad) into their rehearsal practice. In each case, a bluetooth pedal is used to turn pages. Of the remaining participants, one would be interested in switching to digital scores but cannot do so at present for budgetary reasons, whereas the other explicitly prefers reading from paper. For users of digital scores, additional tooling includes the **forScore app** (score viewing and annotating app mentioned by two participants), and the **Henle app** (score subscription service). Four participants also indicate the use of **IMSLP** for score acquisition. In terms of properties required for a digital tool to be useful, **focus** (lack of distractions, mentioned by three participants), **reliability** (mentioned by three) and **performance speed** (mentioned by two) were seen as most important, particularly in a concert performance context. **Affordability** was mentioned by one participant.
- ❖ On *pedagogical context* (responses given after demonstration of the CLARA prototype), responses were mixed. Three participants expressed concerns that teachers would not want to use such a tool due to the **additional time and effort involved** (also echoing comments during the user study pilot sessions in the first year of the project), whereas two saw potential for its use, with one reporting on experience with a digital platform that matches competition participants with experienced reviewers / judges as an example use case where such a tool would be particularly well placed.
- ❖ Feedback in response to the demonstration of the current CLARA prototype: the ability to **revisit rehearsal recordings** and to **navigate** these through interaction with the score was universally seen as useful, as was the ability to **visualise performance errors. Automated page turning** was explicitly deemed useful by three participants, though one requested

these to happen earlier than as in the demo shown to them (e.g., half a measure before the end of page, possibly configurable). Feedback on **tempo** and **dynamics visualisation** was mixed: three participants saw tempo visualisation as having limited (two participants) or no (one participant) utility; whereas two appreciated its usefulness, with one stating "the feature that appeals most is the tempo" (for checking variability and consistency with objectivity, and comparing to previous rehearsal renditions). Similarly, **dynamics visualisation** was seen as less helpful by two participants ("should be audible, not visualised"), but potentially very useful by two others ("as a sanity check to your own perception"; "checking that every voice in a fugue is audible"). One participant specifically proposed **aggregate dynamics measures** as a useful means of **performance error detection**, e.g., to verify that a crescendo specified in the score was reflected in performance – though outlined the need for the use of good editions to provide the necessary references. In broad terms, three participants saw utility for their personal rehearsal practice and could see themselves incorporating such a tool into their rehearsals; of the remaining two, one participant (the most advanced pianist) could see utility of such a tool at earlier points in her career; whereas one saw utility only in the automated page turning, but otherwise stating that "I don't see how this would help me develop my muscle memory … I don't see how it would help me in practice".

4.5. Impact on future work

The initial findings reported above already point to specific priorities in ongoing development of the prototype, and will also be revisited in the subsequent phase of user testing in interactive rehearsal sessions foreseen for late 2020 / early 2021. We expect further insights from the second batch of sessions in the current study, scheduled for early November 2020. Insights with implication on development obtained thus far include:

- ❖ The outlined expectations on **reliability** and **performance** of digital tooling easily translate into development priorities: while interactions with the current prototype are performant and interactive once loaded, the initial loading of the interface (~ 30 seconds) and the turning of pages (~1–2 seconds) are currently too slow and need to be improved.
- ❖ Comments on the usefulness of **error visualisation** (by most participants) and **dynamics visualisation** (by two participants) suggest that a more detailed focus on these areas may prove useful. Currently, these are visualised per-rendition, through colouring of the notes – as opposed to the tempo curves, which are visualised separately and can be used to compare between renditions. Here we envision the creation of further stand-off visualisation components. Possibilities include a visualisation of rehearsal renditions as horizontal lines or bars (distributed chronologically along the y-axis), which change colour to highlight inserted or omitted notes (performance errors) at corresponding note positions on the x-axis; and, aggregated dynamic curves (analogous to the current tempo curves), potentially with additional visualisation of dynamic cues from score directives. A means of visualising dynamics more granularly, e.g., to show the relative values of simultaneously sounding voices or notes, should be investigated.
- ❖ The diversity of expectations, and worry about digital distractions by some participants, suggest the usefulness of **personalisation** – to control what information is displayed and what is hidden, as well as other interaction aspects, e.g., the degree to which automated

page turning should come early (before the recorded music reaches the final note of the score).

- Score annotations are seen as useful to a degree by all participants, though a small palette of symbols for placement above notes (particularly fingerings) may be sufficient to cover the most important use cases.

5. Choir singers

We are currently working with six choirs that participate in the use of the pilot for choirs, which are having their repertoire available for the individual study of the singers. We have faced difficulties in the digitisation of scores due to the need to change their usual practice, more seasoned to work with physical scores. The changing scenario of concert commissions and the intermittence of face-to-face rehearsals due to COVID-19, has also caused delays in the works, both in the digitisation of the repertoire and in the subsequent revision of the voice synthesis. The 6 choirs coincide in the opportunity provided by the pilot to work on a repertoire at home, a common issue among all of them which has been emphasised in the pandemic scenario.

5.1. Aim of the evaluation study

After the initial user validation sessions of the pilot that we reported in the Deliverable D6.6, several modifications of the pilot were carried out to improve it. These changes were either already planned to be implemented and verified by the users as well as other aspects that were detected during testing. The aim of this evaluation study was:

- To validate the changes made to the pilot upon the studies conducted previously and described in Deliverable D6.6.
- To conduct studies in order to prioritise the functionalities to be implemented the development of the next version of the pilot.

5.2. Participants

5.2.1 Recruitment strategies

Some of the users of this study are members of the choirs that we have contacted in the previous evaluations, which they also helped us with useful information in order to contact other choirs that could be potentially interested in participating in the project. As reported in D6.6, we approached local amateur choirs around Barcelona where UPF and VL partners have strong links with the Catalan Federation Choirs (FCEC) and the associated partner ESMUC (Higher School of Music of Catalonia). This resulted to the following list of choirs that participated:

- Fuga de Lluïsos de Gràcia (Barcelona, Catalonia)
- Violeta de Centelles (Centelles, Catalonia)
- Cor de la UPF (Barcelona, Catalonia)
- Lloriana jove (Torelló, Catalonia)
- Gregorian de Iubis (Vic, Catalonia)
- Orfeó Vigatà (Vic, Catalonia)

A special mention is deserved to the collaboration agreement signed with Cantoría, a vocal ensemble created in 2016 that has its origin in the Escuela Superior de Música de Cataluña, that specialises in the interpretation of vocal polyphony of the Iberian Renaissance. Cantoría is a young and very dynamic musical team that, as of 2017, has received various recognitions and awards at the international level and that has led them to build a remarkable national and international career. Cantoria is developing the More Hispano project that aims to share the musical repertoire of the spanish renaissance of which they are experts with various choral groups. For the development of this project they were looking for a digital platform to promote the dissemination, diffusion and knowledge of this repertoire, as well as a research and academic environment that could share their concerns.

Thus, the TROMPA pilot for choirs will include the repertoire of the vocal ensemble so that it can be made available to the participating choirs. Currently, Cantoría has already made studio recordings. All this will be realised -if the health situation allows it- in a concert in Barcelona in which the participating choirs will interpret the repertoire of Cantoría.

5.2.2 Participant characteristics

One of the main criteria of the selection of the choirs has been the diversity of its members with respect to musical knowledge and studies. Consequently, we aim to obtain a usability result of the pilot for a group of people with very diverse musical and learning needs that must, at the same time, achieve a unified, common and compact result when presenting the repertoire of the choir publicly. On the other hand, we consider that the choirs that participate in the pilot can represent, in a reduced universe, the reality of the majority of the amateur choirs in Catalonia. The six participating choirs also guarantee a representation of different age groups that perhaps will generate different results in front of the proposed scenario.

We also want to highlight the participation of young choirs. Firstly, the UPF choir and the proactivity of its director, which allowed the 20 to 25 age group to be represented in the project, with its 22 participants. Together with the Lloriana Jove choir, with a median age of 25 years, they represent the youngest choirs of the list.

As mentioned, all choirs are based in the area of Barcelona, mostly in the territory of central Catalonia. Taking into advantage of contacts of co fact that can facilitate a subsequent meeting, if health circumstances allow it. At the moment, all the actions carried out with the choirs have been telematic or in person, with the presence limited to 6 people and with all the security measures. A summary of the demographics of the choirs is presented in Table 5.1.

Choir	Participants[2] (male/female)	Conductor	Ages	Musical skill distribution (high, medium, low)
Fuga	26 (6 / 20)	Female	26-52	30%, 40%, 30%
Violeta	28 (11 / 17)	Male	35-65	20%, 40%, 40%
UPF	22 (8 / 14)	Female	20-25	30%, 40%, 30%

[2] For the Fuga and Violeta choirs not all singers will participate in the pilot.

Lloriana jove	28 (9 / 19)	Female	20-25	15%, 20%, 65%
Gregorian de Iubis	11 (10 / 1)	Male	50-66	45%, 40%, 15%
Orfeó vigatà	40 (13 / 27)	Male	25-68	20%, 60%, 20%

Table 5.1. Summary of the sex, age and musical skill distribution of the participants.

5.3. Study protocol

For all choirs we followed the same procedure. Firstly, the project was presented to the choirs; in this first contact, it is possible to identify various elements of the choir that can give us clues in order to propose their participation in the project: diversity of musical levels, the ways of the individual practice of the choir members, the general organisation of the choirs and their internal procedures, as well as their interest and time to dedicate to the project. Next, they were asked about the repertoire they would like to include in the pilot, in digital format. Technical support is given for digitising scores if required. Afterwards, we proceed with the synthesis of the scores, their revision by the directors, to apply modifications to the scores if problems have arisen, as well as changes to solve the problems detected in the synthesis. Meanwhile, a training session was held with the singers of each choir where users individually access the pilot. Finally, users were registered so that they can use the pilot to study their repertoire and provide feedback by proposing improvements and detecting problems. Regarding ethics, we follow the procedures implied by the UPFs ethics approval.

Due to the situation caused by COVID-19, the activity of the choirs has been altered. After the summer, given that the first wave of the COVID-19 crisis had ended, all the choirs raised their objectives and recovered some actions that were temporarily canceled during the first period of confinement. Given that a few weeks after the rehearsals - with all the security measures and some singers who did not join the rehearsals - the health situation has become complicated again, the concerts and activities are canceled again making the planning of such activities uncertain. However, at the same time it can be an opportunity for the individual use of the pilot for choirs.

The main difficulty in putting into operation the choral practice of the members of the choirs has been the digitisation of the choral pieces; in many cases there were no digital versions of the repertoire, and many choirs had no previous experience on digital score programs and editing. The members of participating choirs might get prepared individually with MIDI files or recordings (whether their own, made by the director, shared by other choirs, or listening to recommendations of the pieces performed on a channel, such as YouTube), but they usually work with physical scores or in PDF.

The uncertainty and cancellation of concerts and activities have hampered the work of the managing boards of the choirs, as well as the directors, who have had to prepare extra material or modify their repertoires, without yet having any assurance about the present and future of their future activities, concerts, even their income and even some choirs do not know for the moment if they can count on all their members. As can be seen in Table 5.1, the male presence is low, and it is very likely that an event could be cancelled by the choir because there are insufficient men to ensure a good level of performance. Thus, on many occasions the directors have had to adapt the repertoire. Below we report a brief description of the current status of each choir and the difficulties encountered so far:

- **Gregorian de Iubis**: There have been delays in the digitisation of scores for this choir, as problems were generated with the synthesis engine of the pilot for the excerpt parts related to the encoding of the scores. The members of the choir have contributed to digitising the pieces, and currently the revision is pending. Although facing the imminent situation of confinement, this choir has already carried out the study for the use of the pilot.
- **Lloriana jove:** The directors have learned to digitise the scores with MuseScore and are in the process of synthesising. A session for explaining the pilot and its functionalities to the choir is pending.
- **Orfeó vigatà:** Currently a Bach song is being digitised. However there have been many delays in the return of the syntheses by the technical team, due to the volume of work caused by cancellation of concerts. It is proposed to speed up the training and start with the study of the aforementioned piece, although it is not certain that it can be interpreted and face-to-face rehearsals can be continued due to the COVID-19 situation.
- **UPF choir:** This choir began with the rehearsals later than the rest of the choirs, but the director has already sent us the scores and the comments about the syntheses. They have also done the pilot training, in recorded video format for later dissemination. Video recording of the pilot training is a good option for the choirs' members to revisit it whenever they need. We plan to follow this practice for the other choirs as well.
- **Violeta:** The choir is already using the pilot for individual study, although their Christmas concert has been canceled and they have opted for the recording of a CD packed by the city council and to be distributed with the municipal magazine. They have provided us with a first return with ideas and proposals on the use of the pilot.
- **Fuga:** The importance of speeding up the process for the individual study through the pilot has been expressed, since the face-to-face tests have been canceled.
- **Other choirs:** Through the director of Fuga, another choir named *Cor Mixt Friends of the Unió de Granollers* have shown interest in participating in the project, they have sent us their digitised pieces and the volume of work is being assessed to be able to incorporate them into the project.

To summarise, the current situation has caused difficulties in the management of the setup of the use of the pilot, making impossible to deliver a coherent and complete evaluation study for all choirs. However, the whole processes with choirs so far have given enough information and feedback in order to evaluate the pilot and provide a better and more appealing version in the next months, as will be described in the next subsections. Moreover, it is certain that the pilot offers to the choirs a valuable tool to support their activities when the health situation allows it.

5.4. Study evaluation outcomes

At the moment, we already have received feedback from some users of the Violeta choir, who have used the pilot for individual and group rehearsals. Although the group rehearsal is not a current functionality of the pilot, the choir used the pilot for group practice by amplifying the synthetic voices. Thus, the group practice is a desired functionality that could be explored as an additional option to be offered to the choirs. Specifically, Violeta has used it to rehearse and used the synthetic voices of the application as an accompaniment in order to strengthen their tuning and that the singers have the harmonic notion of the rest of the voices. A summary of the comments and proposals are presented in Table 5.2.

Issue category	Issue description / Proposal
Visualisation	Possibility of use on mobile phone (choir members are used to listen to MIDI / recordings through mobile for learning)
Voices	Functionality to put all voices in “mute” except the recorded one
Accompaniment/armony	Possibility of integrating a piano accompaniment of the pieces (or an orchestral simplification)
Sound quality	The soprano group indicates that the voice synthesis in some higher notes sounds metallic or strange (not out of tune)
Repertoire	Possibility of accessing more repertoire than the one owned by the choir
Piano roll visualisation	In some screens the text of the song lyrics did not appear on the screen
Interface	Individual case. 1 person has proposed to modify the colors and the appearance.
Tempo	The possibility of listening at a slower tempo is considered important for difficult passages (option already provided)
Accompaniment	The quality of the choral synthetic voices in not good when the sound volume is high (using mini-speakers)

Table 5.2. A summary of the feedback received from the choirs so far..

During the tests, the Gregorian de Lubis choir highlighted the importance of the organ accompaniment in its repertoire. The choir Orfeó Vigatà delivered a synthesis of Bach's Cantata No. 131 that also incorporated a piano reduction, with the intention that it might appear in the pilot. For now, we will work only with voices waiting to assess the implementation of this option.

The next step will be to schedule a group session for each choir, since at the moment the proposals have been collected directly through the director of each choir. We will study the most effective system for data collection, depending on the number of users, the interaction between them and the involvement of the director.

As a summary, there is a very positive response regarding what the pilot offers, its simplicity of use and its potential for integrating additional features. The pilot is very intuitive and is understood very easily and quickly by the users. It is worth mentioning the effort made by the older group to be present in online training and the use of virtual tools, which were already given in this group, but it has been accelerated because of COVID-19.

5.5. Impact on future work

Regarding further user studies for the pilot, by taking into account the evaluation studies so far, the next intermediate steps are to finish with the current studies, to review and deliver the digital musical material so that the choirs can individually work their pieces in a larger scale and to collect comments on the use of the pilot.

It is also proposed to incorporate the improvements already planned for the pilot in the next version, as well as new functionalities based on the contributions that were collected.

For the longer term, it is proposed to hold thematic sessions so that the choirs can debate on the possibilities that the pilot offers, the difficulties they have encountered in its use, the differences between people with different backgrounds, and the result of their work reflected as a concert / representation. We also propose to continue monitoring the feedback collected by the choirs during the use of the pilot, such as Violeta, as in the case of Violeta.

We also foresee the incorporation of a common repertoire, material from the vocal ensemble Cantoría, which would be made available to all the participating choirs for study and a subsequent joint action, in a face-to-face / online concert (to be confirmed according to the COVID-19 situation).

At the dissemination level, it is proposed to present the pilot in specialised magazines with musical content aimed at choirs, as well as to present it in a choral musical event at a local, national or European level.

Regarding the development of the Choir Singers Prototype, it has been driven by the user needs from its initial design stages. The user evaluation reported above has been the first time that end users could actually test the functionalities of the prototype. It includes: listening to synthesised renditions of the repertoire of their own choir; and recording and visualising the analysis of their rehearsals.

The feedback received will determine the feature requirements in the next prototype iteration towards the end of 2020, and subsequent iterations until the end of project (M36). We can group this feedback received in three categories: *a)* user and content management, *b)* playback and recording functionality; and *c)* choir synthesis quality. Next we detail our future plans for these three categories.

5.5.1 Improvements on user and content management

- ❖ Automated user registration: in the current version, the user management is done manually by database administrators (Voctro Labs staff). In order to scale it to larger user evaluations, we will add the functionality of self-registration for users. For the next iteration, the subscription of choirs will still be managed case by case, with the aim to have a complete automated user and choir management system before the end of the project.
- ❖ Users member of multiple choirs: this is a request for some participants that were members of two choirs. This feature was already considered in the back-end (database model), although has not been yet implemented on the front-end. A choir selection menu for those users will be implemented in the next iteration.
- ❖ Internationalisation support: the current version of the prototype is only in English. We plan to translate the strings to Catalan and Spanish to facilitate its usage by local choirs in Catalonia and Spain we are collaborating with.
- ❖ Automated repertoire creation: currently, adding new repertoire to the CSP is a manual process. Participant choirs have provided scores in MusicXML format, scores are manually revised, and the synthesis renditions are generated offline with Voctro Labs algorithms. Pieces are also manually added to the back-end database. We will automatise part of this

process allowing users (choir conductors or choir managers) to upload new pieces to the CSP, which can then be synthesised through Voctro Labs' Voiceful API.

- Public repertoire (without registration): the current version of the CSP requires a user registration. On the one hand it is a privacy requirement for managing user recordings, but also it allows a more controlled usage in the initial development stages. However for obtaining new users at a larger scale, this might be a limitation. We got a request by virtual choir initiatives (e.g. Stay at Home Choir) that can benefit from a more direct access to a public repertoire. We will study the necessary changes to the user and content management (front-end and back-end) to offer the possibility of public repertoire. Functionalities for non-registered users will be limited, e.g. not having access to recording storage, notes sharing with choir members, etc.

5.5.2 Improvements on score playback and voice recording

- Turn Page button: this is a request consisting in implementing a way to turn the score page before the last note is played. This is a common practice for choir singers, who read and memorise the last bar, and turn the page to read the first bar. We will implement a button at the bottom right corner so that singers can easily press it while holding the tablet.
- Instrumental track: it is common that choir repertoire contains an instrumental accompaniment by orchestra or piano. This has been extensively requested in the user studies. We will implement a new audio track to the CSP, allowing choirs to upload an instrumental track, in the form of an audio MP3/WAV file. In the next iteration, we expect the instrumental track to be perfectly synchronised in time with the MusicXML score.
- Practice tracks (expressive performances with varying tempi): in addition to the functionality of synthesising scores with artificial voices, some choirs requested the possibility to upload already available practice tracks[3]. We will implement a new functionality to upload one audio file per part. Also we shall support score alignment with a time-varying tempo, as we find in real expressive performances. This new feature will be implemented in different stages, and was one of the requests for the collaboration with the professional choir Cantoría. Timing information shall be manually annotated (e.g. tempo tapping).
- Repetitions: currently, if the score has repetitions, we manually edit the MusicXML/MEI score using external software (eg. MuseScore) to unroll the repetitions. We will implement a way to programmatically do this score unrolling when loading the pieces on the prototype. Having unrolled scores allows to keep the same functionality for both score and piano-roll visualisation and internal time management.
- Divisi: current version of the engine synthesises the divisi, but the visualisation and selection is not supported on the CSP. In the next iteration, we will first ensure that the score-following visualisation works properly in case of divisi. Also we will decide how a final solution shall be implemented, for example, offering a sub-track for each divisi with a radio button to select it.

[3] Practice tracks are audio recordings for each part (e.g. four tracks in the case of SATB scores) that choir conductors share to choir members. These tracks can typically be a singing recording or also a piano reference recording with the melody.

5.5.3 Improvements on choir synthesis quality

- Portamento in pitch contour: in the current version of the synthesis engine, the generation of a pitch contour from score uses a rule-based algorithm designed for generic singing style (e.g. pop vocals). This algorithm generates smooth transitions for note to note, which introduce an unnatural portamento in case of choir repertoire with long notes (e.g. at slow tempi). The next iteration of the prototype the pitch contour will use a new generative model based on Deep Learning that shall overcome this limitation.
- Vibrato in pitch contour: same as above for the portamento, the current version of the algorithm to generate the pitch contour introduces vibrato, especially for long notes. This is not common in choir repertoire. The next iteration of the prototype the pitch contour will use a new generative model based on Deep Learning that shall overcome this limitation.
- Correct diphthongs synthesis in German: revision of the vowel to stretch in case of long diphthongs.
- Correct mispronunciations in Latin: some Latin words are not correctly transcribed phonetically, leading to mispronunciations. Current transcription in Latin is done by doing first a transcription on the Spanish rules, and applying some additional Latin-specific rules. We will revise the cases, and add new rules to the transcription algorithm.
- Synthesis naturalness: the current version of the synthesis engine uses a parametric vocoder (WORLD) which produces a sound with an artificial character. This is especially noticeable on sustained vowels at the constant pitch. We will update the synthesis engine with a newer algorithm that, together with other improvements on the timbre modelling, uses a Wavenet vocoder to produce a more natural sound.

6. Music enthusiasts

In the context of the Music Enthusiasts use case, several user-centered evaluations have taken place along the duration of the project so far to refine requirements, validate usability of a minimum viable product and to determine the most suitable incentives to implement in these early stages. The results of these evaluations can be found in deliverables **D6.1 - Final Mock-ups Testing** (Section 5) and **D6.7 - Working Prototype for Music Enthusiasts v1** (Section 3.4). During the last 6 months an evaluation study was conducted, divided in two annotation contests (first contest from June 29 to July 5, 2020; second contest from October 14-20, 2020). The current evaluation study description focuses on the first contest, since by the time of this report, collected data for the second contest was still under analysis.

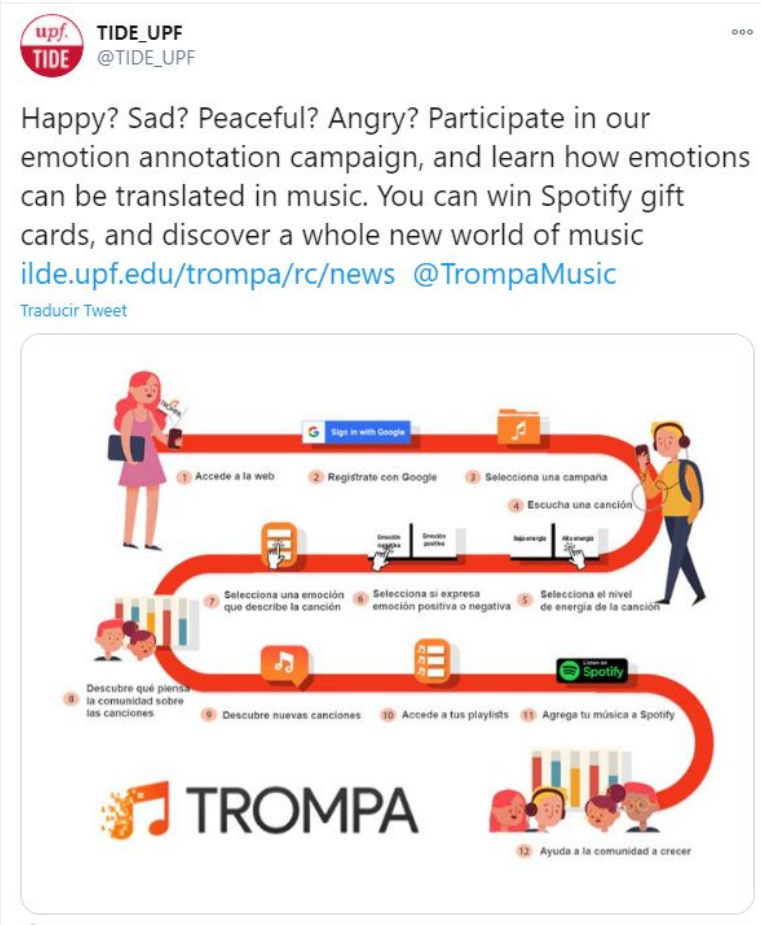
6.1. Aim of the evaluation study

The aim of this evaluation study was to evaluate the usability and workflow of the pilot in a real setting (participants using the ME platform by their own with their own devices), as well as to determine the impact of some of the implemented incentives (e.g. scoring system, contributors' ranking, music recommender system based on emotional content) and the quality of the annotations. Likewise, the evaluation study allowed to assess the scope of some of the dissemination mechanisms available, e.g. mailing lists and social networks. This evaluation study focused on user behavior data collected through the platform[4] .

6.2. Participants

6.2.1 Recruitment strategies

UPF-MTG and UPF-TIDE networks (Twitter, mailing lists, etc.) were the main recruitment strategy (Figure 6.1). The dissemination strategy focused on, but was not limited to, UPF students (i.e. Higher Education students). TROMPA social networks were also used to disseminate the contest. The pilot currently runs in English and Spanish, so the call for participation messages were disseminated in both languages.



TIDE_UPF
@TIDE_UPF

Happy? Sad? Peaceful? Angry? Participate in our emotion annotation campaign, and learn how emotions can be translated in music. You can win Spotify gift cards, and discover a whole new world of music ilde.upf.edu/trompa/rc/news @TrompaMusic

Traducir Tweet

upf. TIDE
Sign in with Google
1 Accede a la web
2 Regístrate con Google
3 Selecciona una campaña
4 Escucha una canción
Emoción negativa
Emoción positiva
Baja energía
Alta energía
7 Selecciona una emoción que describe la canción
6 Selecciona si expresa emoción positiva o negativa
5 Selecciona el nivel de energía de la canción
8 Descubre qué piensa la comunidad sobre las canciones
9 Descubre nuevas canciones
10 Accede a tus playlists
11 Agrega tu música a Spotify
Listen on Spotify
TROMPA
12 Ayuda a la comunidad a crecer

Figure 6.1. Example of a dissemination tweet promoting the contest.

6.2.2 Participant characteristics

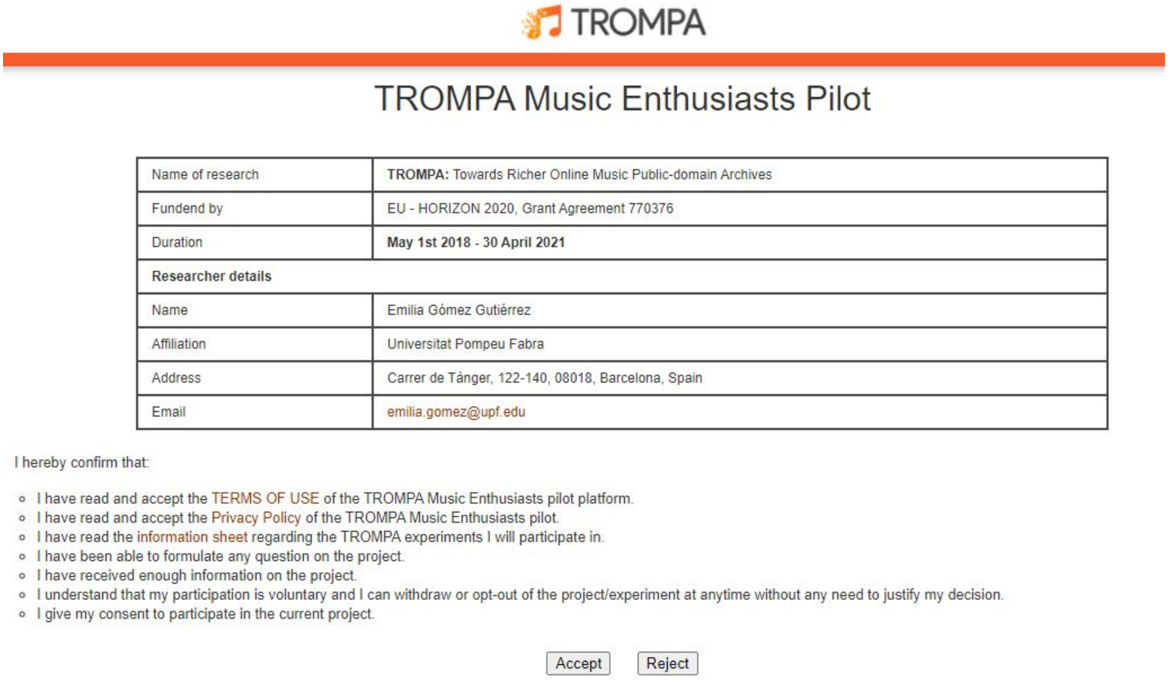
- Participants were English and/or Spanish speakers.
- For the first contest, 32 active participants generated 694 annotations.
- For the second contest, 23 active participants generated 655 annotations (these data have not been analysed yet).

[4] https://ilde.upf.edu/trompa/

6.3. Study protocol

During the contest period, participants had to complete as many annotations as they could in order to obtain an external reward (i.e., Spotify gift cards for the first contest and BandCamp gift cards for the second contest) . The winners of the contest were defined using the scoring system designed as an incentive mechanism. General rules for the contest were defined as follows:

- Participants must login in the pilot. In order to register, users must accept the TERMS OF USE of the platform (Figure 6.2). The information sheet is presented to the user, where detailed information about the collected data, use of this data for research purposes, as well as the privacy policy of the pilot is described.
- Once they were registered, they were able to annotate.
- Participants must complete at least one of the available campaigns (different from the Tutorial campaign).
- In case of a tie, the winner of the prize will be determined as follows:
    - Highest amount of valid annotations done during the contest period.
    - Highest amount of completed campaigns during the contest period.
    - Highest amount of annotations during the same access to the platform.
    - If the tie persists, the prize will be awarded by lottery.



TROMPA

TROMPA Music Enthusiasts Pilot

Name of research	**TROMPA:** Towards Richer Online Music Public-domain Archives
Fundend by	EU - HORIZON 2020, Grant Agreement 770376
Duration	**May 1st 2018 - 30 April 2021**
Researcher details	
Name	Emilia Gòmez Gutiérrez
Affiliation	Universitat Pompeu Fabra
Address	Carrer de Tànger, 122-140, 08018, Barcelona, Spain
Email	emilia.gomez@upf.edu

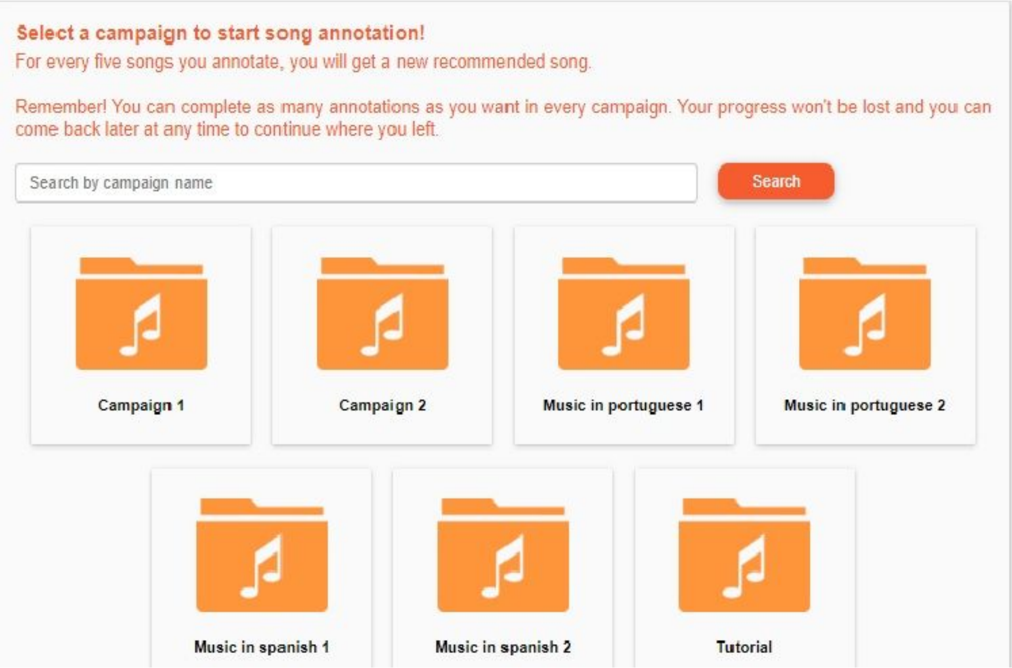
I hereby confirm that:

- I have read and accept the TERMS OF USE of the TROMPA Music Enthusiasts pilot platform.
- I have read and accept the Privacy Policy of the TROMPA Music Enthusiasts pilot.
- I have read the information sheet regarding the TROMPA experiments I will participate in.
- I have been able to formulate any question on the project.
- I have received enough information on the project.
- I understand that my participation is voluntary and I can withdraw or opt-out of the project/experiment at anytime without any need to justify my decision.
- I give my consent to participate in the current project.

Figure 6.2. Screenshot of the *terms of use* section of the pilot. Links to the information sheet and privacy policy are available for users.

A contest structure was selected as the protocol for this study in order to evaluate the pilot in a real controlled setting. Likewise, data was collected through the user behavior data collected within the platform. Furthermore, the main aim of the study was to evaluate the usability of the pilot based on the quality of the obtained annotations.

Previous to the contest, there were 84 songs that were annotated by participants of previous experiments described in deliverables D6.1 and D6.7, contained in two campaigns (Campaign 1 and Campaign 2). For the contest, 4 additional campaigns were incorporated within the platform so the participants could explore different types of music, namely Music in Portuguese 1, Music in Portuguese 2, Music in Spanish 1, Music in Spanish 2 (Figure 6.3). Additionally, we added a question for the participant to select the reasons for which they decided on the arousal, valence and emotion annotations. Our aim is to be able to understand if the annotations were made by using musical properties, rather than personal felt experience (i.e. to verify if the participants understood the information provided though the different sections of the platform explaining the aim of the pilot and the differences between *perceived* and *induced* emotions).



Select a campaign to start song annotation!

For every five songs you annotate, you will get a new recommended song.

Remember! You can complete as many annotations as you want in every campaign. Your progress won't be lost and you can come back later at any time to continue where you left.

Figure 6.3. List of available campaigns for the first contest.

6.4. Study evaluation outcomes

Based on the collected data during the first contest, the following conclusions were drawn:

- ❖ Figure 6.4 shows the distribution of annotations across the different songs to be annotated. Previous to the contest (songs 0 - 84), songs were annotated a similar amount of times since all participants were asked to annotate all the songs (as described in section 3.4.1 of D6.7-v1). New songs added for the contest (songs 85 - 189) were only annotated between 1 and 5 times, since participants were free to annotate as many songs as they wanted from all the available songs (songs 0 - 189) during the contest period.

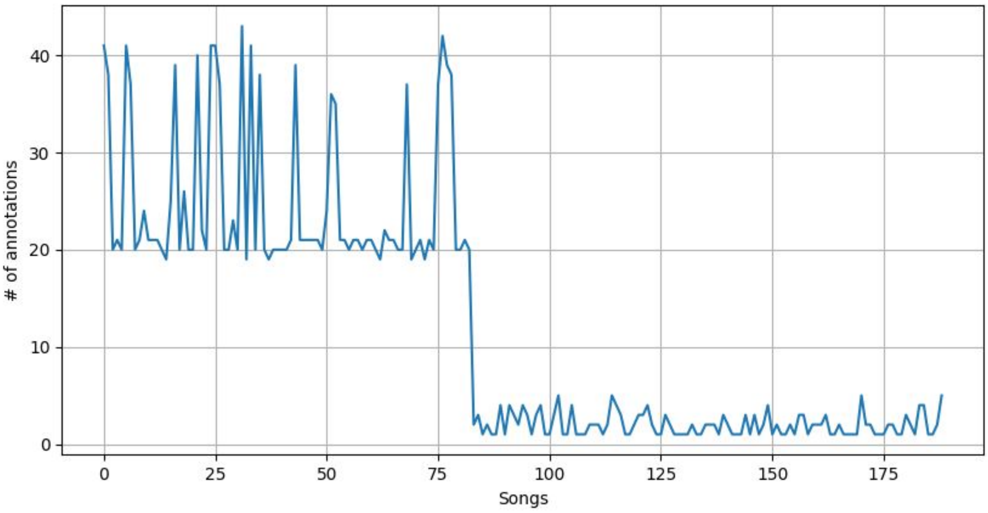


Figure 6.4. Annotation distribution after the first contest.

❖ Figure 6.5 shows words count for the reasons for which participants decided on the arousal, valence and emotion annotations. In general, we find that: (1) associations of arousal relate to tempo, articulation, and sound level, (2) association of valence relate to harmony, timbre, modes (major or minor) and melody contour, and (3) association with specific emotions is usually related to felt emotion. We evaluated the reliability of the collected data using Krippendorff's coefficient α to understand the importance of inter-rater agreement on the collected annotations. In summary, we obtained: (1) $\alpha_{Arousal} = 0.486$, (2) $\alpha_{Valence} = 0.364$, and (3) $\alpha_{Emotion} = 0.170$. This result shows that users still associate their selection about emotions with the induced emotions instead of perceived emotions.

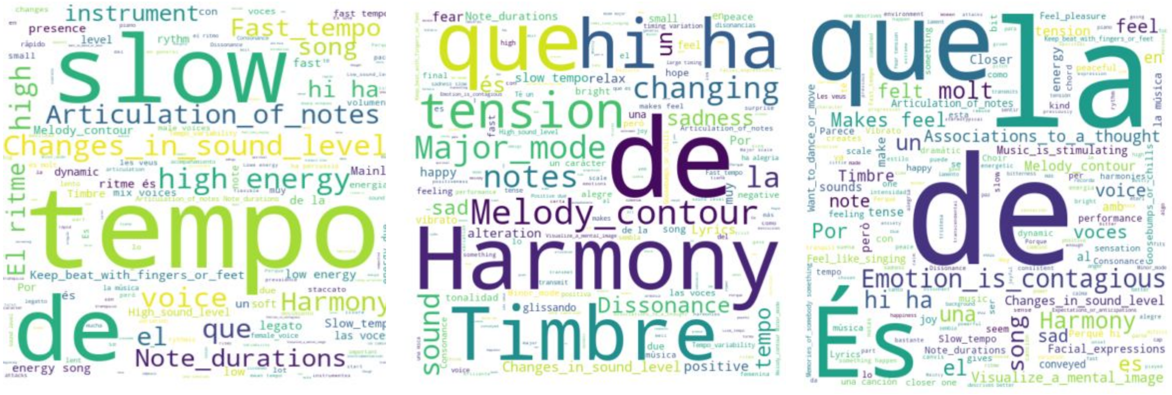


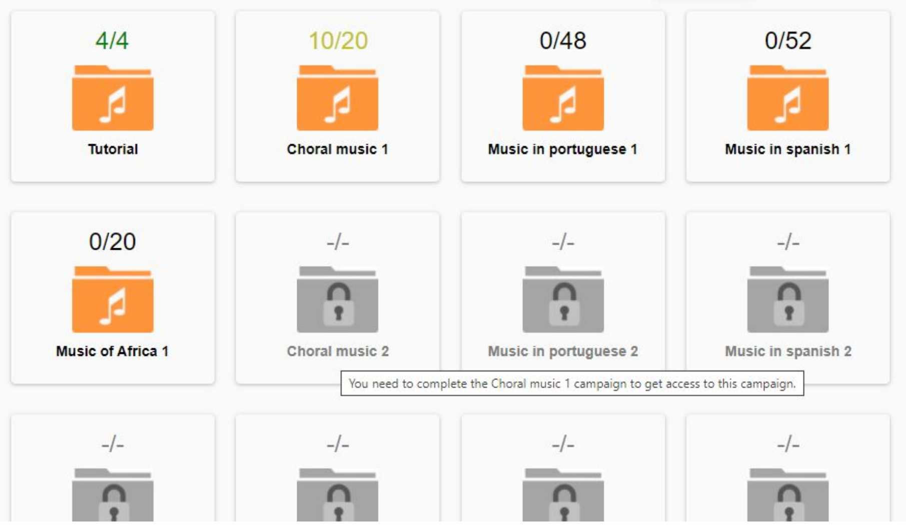
Figure 6.5. Word clouds of the reasons for annotations in TROMPA ME dataset.

❖ Scoring system and the ranking visualisation were effective incentives to generate a contest with tangible rewards. The system is reliable for evaluating user performance in a specific time period. Nevertheless, no user behavior data was collected to determine how participants interacted with the recommended music.

6.5. Impact on future work

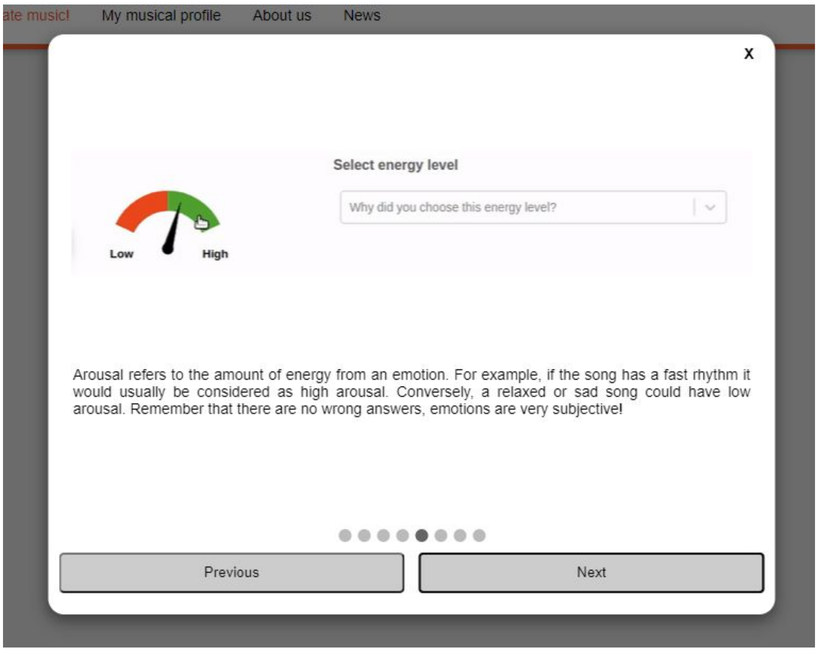
The outcomes of this study had a direct impact on the pilot design. Based on the conclusions, several modifications and new features were implemented in the platform (and are tested and validated during the second contest). The following list summarises the modifications made based on the outcomes of the first contest:

- Given the low amount of annotations per song during the first contest period, an “unlock levels” gamification approach was implemented, in which new songs appear to be unlocked after completing a specific campaign (Figure 6.6). In this way, we can increase the amount of annotations of songs that need more annotations.
- A more detailed tutorial on how to annotate is needed in order to improve the quality of the annotations. For that reason, a new tutorial section of the platform has been designed (Figure 6.7). This tutorial is shown automatically to new users (and users that have not annotated before the implementation of the tutorial), while it will remain available anytime in the annotation screen.
- It is necessary to focus on the development and improvement of the recommendation system and the visualisation of the recommendations in order to generate new incentives within the platform beyond the external tangible rewards used by now.
- Even when the scoring system provides valuable information about users’ behavior, it is necessary to implement new dashboards with valuable information for the users in order to engage them to keep annotating. At the moment, a new ranking scoring dashboard has been designed (Figure 6.8) to explore the top annotators within specific time ranges.
- It is necessary to collect user behavior data to determine how participants interact with all the sections of the platform. Modifications in the platform workflow have been implemented to collect user analytics within the platform, focusing on user clicks.



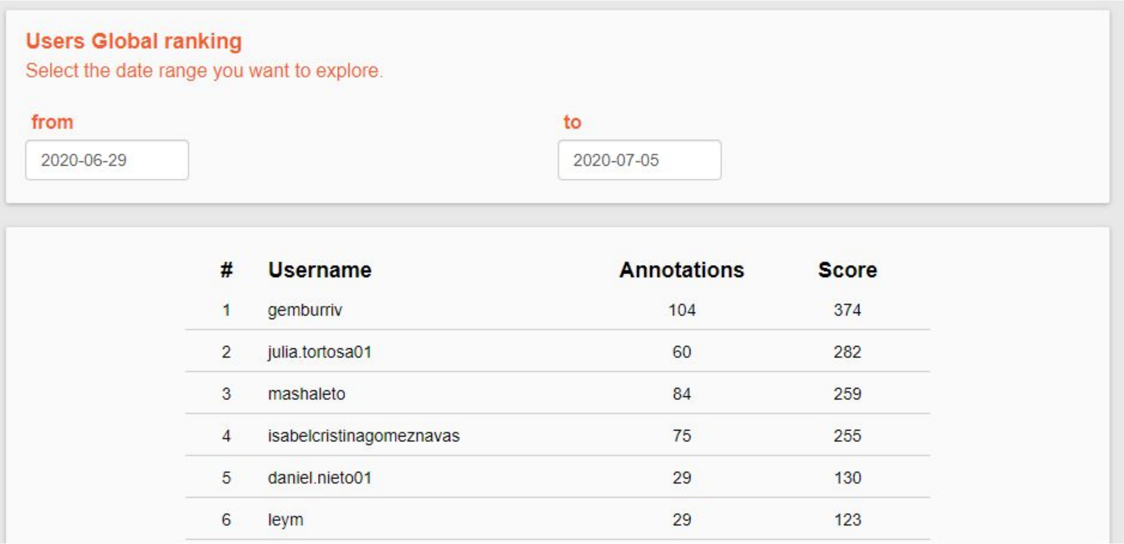
You need to complete the Choral music 1 campaign to get access to this campaign.

Figure 6.6. VIsualisation of the “unlock levels” gamification approach.



Arousal refers to the amount of energy from an emotion. For example, if the song has a fast rhythm it would usually be considered as high arousal. Conversely, a relaxed or sad song could have low arousal. Remember that there are no wrong answers, emotions are very subjective!

Figure 6.7. Visualisation of the tutorial section popup message. Different stages of the platform are explained in the slides.



Users Global ranking

Select the date range you want to explore.

from: 2020-06-29

to: 2020-07-05

#	Username	Annotations	Score
1	gemburriv	104	374
2	julia.tortosa01	60	282
3	mashaleto	84	259
4	isabelcristinagomeznavas	75	255
5	daniel.nieto01	29	130
6	leym	29	123

Figure 6.8. Score ranking dashboard. Users can filter by date range.

7. Conclusion

In this deliverable, we have reported on the current evaluation status for the prototypes delivered under TROMPA's five use cases. Despite COVID-19 circumstances, in all cases, relevant user audiences have been reached, and relevant feedback has been obtained for the next prototype iterations. With the findings reported in this document, we will perform, integrate and evaluate these next iterations over the remaining months in the TROMPA project, which will lead to updated Prototype Deliverables D6.3-D6.7 in M34 of the project, and a final evaluation report D6.9 to be delivered by M36 of the project.