

D1.2 Data Management Plan

Project acronym	TELMi
Grant Agreement nr	688269
Project title	Technology Enhanced Learning of Musical Instrument Performance
Call identifier	H2020-ICT-2015
Topic	ICT-20-2015
Type of action	Research and Innovation Action
Start date of project (dur.)	February 1st, 2016 (36 months)
Report availability	PU-Public
Document due Date	July 30th, 2016
Actual date of delivery	July 30th, 2016
Leader	UPF (internal review RCM)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

Reply to	Oscar Mayor (oscar.mayor@upf.edu)
Additional main contributors (author's name / partner acr.)	Rafael Ramírez (UPF), Aaron Williamon (RCM), George Waddell (RCM), Gualtiero Volpe (UNIGE)
Document reference	TELMI-D-WP4–UPF-20160730-D1.2DataManagement Plan
Document status	Final



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

Table of contents

[Executive Summary](#)

[2 Introduction](#)

[3 List of Prospective Datasets to be Shared](#)

[3.1 Raw Data Resulting from Multimodal Performances Acquisition](#)

[3.1.1 Description](#)

[3.1.2 Types of Data \(Generated and Collected\) and use of Standards](#)

[3.1.2.1 MoCap Data](#)

[3.1.2.2 Video and ambient audio](#)

[3.1.3.3 Instrument audio](#)

[3.1.3.4 Video, Audio, IR, Depth and Mocap data from Kinect for windows v2 sensor](#)

[3.1.3.5 IMU data \(Accelerometer, Gyroscope, Magnetometer\) from XOSC Sensors](#)

[3.1.3 Data Sharing and reuse](#)

[3.1.4 Archiving and Preservation](#)

[3.2 Music Education Datasets and Users Feedback](#)

[3.2.1 Description](#)

[3.2.2 Types of Data \(Generated and Collected\) and use of Standards](#)

[3.2.3 Data Sharing and reuse](#)

[3.2.4 Archiving and Preservation](#)

[3.3 Public Database Repository Data](#)

[3.3.1 Description](#)

[3.3.2 Types of Data \(Generated and Collected\) and use of Standards](#)

[3.3.3 Data Sharing and reuse](#)

[3.3.4 Archiving and Preservation](#)

[3.4 Additional Guidelines for the Data Management Plan](#)

[3.4.1 Discoverability and Accessibility](#)

[3.4.2 Additional Archiving and Preservation Requirements](#)

[3.4.3 Compliance with Ethics Requirements and Protection of Personal Data](#)

[4 Conclusion](#)

[5 References](#)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

Executive Summary

Within the TELMI project, qualitative and quantitative multimodal data will be acquired, managed and integrated as part of numerous core activities including teacher and student performances, user feedback, expert annotations and evaluation.

The sharing of this data in the form of structured datasets can prove to be of high value both to the TELMI consortium and its stakeholders, particularly in the the Music Education sector.

This deliverable outlines the data management practices that will be followed by the TELMI consortium in order to achieve these objectives.

The main goals of this deliverable are the following:

- To outline the potential types of datasets that will be publically shared for the duration of the TELMI project.
- To lay a common foundation for data management across the consortium and ensure interoperability of data & metadata among the partners.
- To gauge each partner's willingness to openly share datasets, and catalogue different sharing strategies.

We foresee the release of various types of datasets:

- User surveys, questionnaires, evaluation data and letters of consent from experts
- Raw data acquired from multi-modal performances
- Public multimodal database of expert and student recordings

Each of these cases is accompanied by its individual Data Management Plan.

We conclude this deliverable with a list of additional considerations to be taken in order to ensure robust access mechanisms to the data, as well as the necessary documentation that we outline the key Work Packages with which this deliverable is interrelated.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

Background

As an Innovation Action project under the Horizon 2020 EU Framework programme, TELMI participates in the pilot action on open access to research data. All projects participating in the pilot are required to develop a first version of the Data Management Plan (DMP), a document that outlines the project's strategy for handling and sharing research datasets. This deliverable presents the first version of the DMP for the TELMI project; given that the DMP is not a static document, but rather a document that is expected to evolve and gain depth throughout the duration of the project.

Therefore, this document implements a first draft of the project's DMP. Besides outlining a series of potential datasets to be released, this deliverable is also instrumental in fostering data sharing as a philosophy within the TELMI consortium, as well as a mechanism for gathering important information on the pre-existing data management practices of the consortium members.

This DMP deliverable has a significant connection with a number of Work Packages within TELMI. Raw Data Resulting from Multimodal Performances Acquisition is connected with WP3, Music Education Datasets and Users Feedback is mainly related with WP2 and Public Database Repository Data is connected with WP6.

2 Introduction

As a project whose core concept is centred on the combination of multiple modalities to assist in music education, TELMI will involve the acquisition and management of multimodal data in several different scenarios (teacher performances, student performances, interviews, questionnaires, etc).

It is in the interest of the consortium to have an open data policy to the maximum degree affordable, both to standardise best practices in recording, storage and data sharing, and to motivate and facilitate the advancement of research in multiple fields.

Where traditional research data (such as training or benchmarking datasets for machine learning algorithms) are concerned, sharing helps in improving the performance and quality of research results, avoiding the duplication of efforts associated with dataset creation and fostering collaboration across institutions both within the EU and abroad.

The main objectives and goals of this deliverable are:



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

- To outline the potential types of datasets that will be publically shared for the duration of the TELMI project. We anticipate the release of several datasets within the timespan of the TELMI project, and their contents and purpose to vary greatly depending on the activity that produces them.
- To lay a common foundation for data management across the consortium and ensure interoperability of data & metadata among the partners. In order to minimize the effort needed to share the collected data, we must ensure that the data management practices of each member of the TELMI consortium are aligned, both within the consortium as well as with the current practices in the Music Education domain. To this end, this deliverable documents these practices and proposes a series of data formats and metadata standards.
- To gauge each partner's willingness to openly share datasets, and catalogue different sharing strategies. While the value of open data sharing is undeniable, it is necessary to ensure that sharing practices are in line with the main objectives and strategic planning of the consortium partners regarding confidentiality when datasets may contain sensitive or personal information. For that purpose, each type of dataset is accompanied by the outline of the sharing strategy.

3 List of Prospective Datasets to be Shared

3.1 Raw Data Resulting from Multimodal Performances Acquisition

This consists of the raw data (motion capture, audio, video, and possibly sensors) directly captured by the TELMI recording platform during music performances by students, teachers and masters.

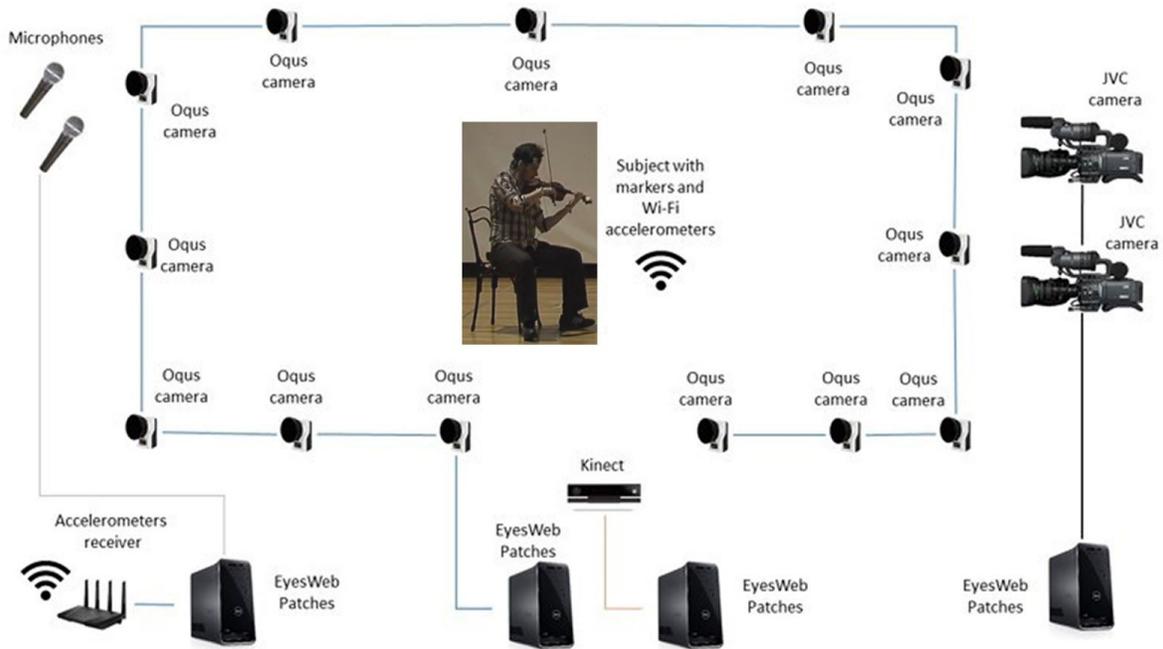
3.1.1 Description

The provisional overall architecture of the platform for multimodal recordings is shown in the following figure (first release of the recording platform expected at Month 8).

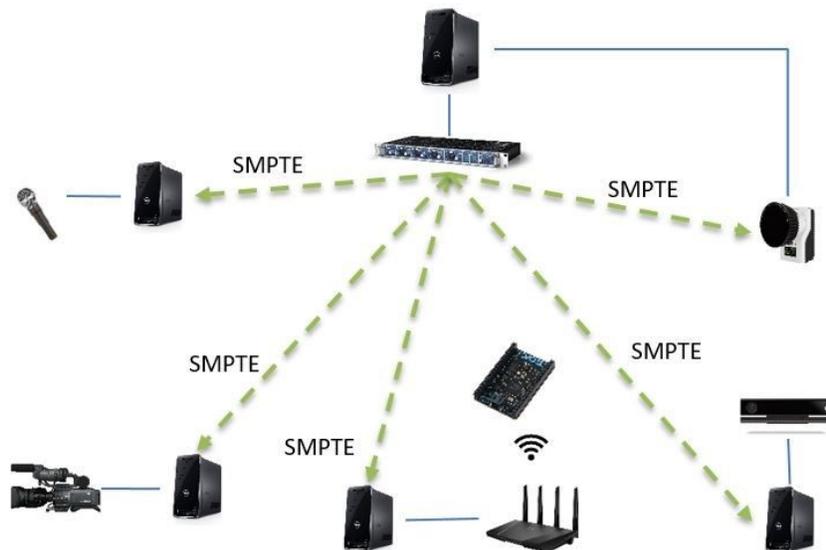
- The performer's movements are captured by a Qualisys Motion Capture system endowed with thirteen cameras.
- Two further broadcast quality video cameras observe the scene, one from the front and one from the side.
- A Kinect for windows v2 sensor further observes the scene from the front, providing video and depth map data.
- The performer wears a set of markers and of rigidbodies composed of a fixed number markers, tracked by the Qualisys Motion Capture System (see Deliverable D3.1 for more details).
- Tracking of the violin and of the bow is performed with real and virtual markers (see Deliverable D3.1 for more details).
- Microphones are placed both in the environment and on the music instrument (see Deliverable D3.1 for more details).
- A set of Inertial Measurement Units (IMUs) to measure hands and trunk movements may be included in case they are deemed relevant for experiments.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269



Synchronization is guaranteed by the EyesWeb platform¹ (see figure below). EyesWeb generates the reference clock used by all the recorders. The generated reference clock is sent to each device in a compatible format. In particular, the Qualisys Motion Capture system receives the reference clock encoded in an audio stream using the SMPTE format. Also the two broadcast video-cameras and the *Audio recorder* use SMPTE encoded as an audio signal. The *IMU recorder* receives the reference clock via network, through the OSC protocol.



¹ http://www.infomus.org/eyesweb_ita.php



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

To guarantee synchronization, EyesWeb keeps track of every recorded frame or sample, and of the timestamp when the data was received. Not all streams can be hardware-synchronized (e.g., with a genlock signal). To afford this problem, a software synchronization is performed by EyesWeb that stores the absolute time at which the data was received. This information is then used when playing back the data. IMU sensors or Kinect are examples of devices which are synchronized in this way.

Further recordings will also be carried out with cheaper motion capture technologies (e.g., Polhemus, see Deliverable D3.1) and with low-cost devices (e.g., Kinect and common video cameras) in order to enable downscaling of prototypes to low cost devices.

Recordings will follow the ethical procedures established in the TELMI Consortium. Where data is to be used in the public database then performers will provide both research consent and release copyright ownership of the recordings to the TELMI consortium for use in the public database, project dissemination, and marketing. Musicians will have the option to release copyright under the condition of anonymity (with identifying features removed from video recordings), otherwise they must explicitly grant the project the use of their likeness and identifying information (see also Sections 3.3.3 and 3.4.3).

3.1.2 Types of Data (Generated and Collected) and use of Standards

The following types of data will be produced during the recording sessions:

- MoCap data from the Qualisys and the Polhemus motion capture systems
- Videos and ambient audio from two professional video cameras
- Instrument audio from the player's instrument
- Video, Audio, IR, Depth Information and Mocap data from a Kinect for windows v2 sensor
- Optional IMU data (Accelerometer, Gyroscope, Magnetometer) from XOSC IMU Sensors

3.1.2.1 MoCap Data

Mocap Data will be saved and stored as QTM and TSV files: the QTM format is a binary and proprietary format by Qualisys, whereas TSV is a plain text format that can be read by any text editor and is used in EyesWeb XMI.

3.1.2.2 Video and ambient audio

The Video and audio streams will be stored using the following encoding:

- AVI file format
- 1280x720 50FPS video with MPEG4 codec
- 320 Kbps stereo audio with MP3 codec (ambient audio in the first channel and encoded SMPTE signal in the second channel)

3.1.3.3 Instrument audio

Audio streams will be stored as stereo AIFF or WAV files containing the instrument signal in the first channel and the encoded SMPTE signal in the second channel

3.1.3.4 Video, Audio, IR, Depth and Mocap data from Kinect for windows v2 sensor



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

The Kinect Video will be stored as a 1920x1080 30FPS (variable fps) AVI video file with mpeg4 codec. IR and Depth streams are stored as a 512x424 30FPS AVI video file with mpeg4 codec; audio will be stored as a single channel AIFF file; MoCap Data will be stored as TSV files.

3.1.3.5 IMU data (Accelerometer, Gyroscope, Magnetometer) from XOSC Sensors

In case recordings include IMU Data, this will be stored as plain text files containing timestamps and data streams of each sensor.

3.1.3 Data Sharing and reuse

Raw Data will be stored internally by UNIGE. Cleaned and ready-to-use data will be made available for public access. EyesWeb patches will be made available to playback the publicly-available data, and to convert it to other commonly used formats. As an example, the data files (IMU or MOCAP sensors) can be exported to the CSV format, to be imported in the RepoVizz database. The audio-video files can be converted to different formats (e.g., MOV, MP4, MPEG).

3.1.4 Archiving and Preservation

Raw data will be stored internally on a dedicated NAS server. Such a NAS is configured for Raid 5 redundancy, allowing a disk failure with no data loss. Moreover, a copy of the data is preserved and archived on an offline portable hard-disk.

3.2 Music Education Datasets and Users Feedback

Over the course of the TELMI project, data will be collected for the purpose of guiding and implementing the pedagogical framework of the project and evaluating the efficacy of the TELMI systems. These efforts will be led by the RCM via TELMI Work Package 2: Music Performance Pedagogy.

3.2.1 Description

In establishing the pedagogical framework for the project, data will be collected from violin students and teachers regarding their current teaching and learning practices and use of technology and where technology may be developed to address the challenges they face. These data will be collected, analysed, stored, and disseminated following standard research practices outlined by the British Psychological Society (BPS) and their Code of Human Research Ethics, including guidelines outlining the obtaining of informed consent and of maintaining participant anonymity. Where data is to be used in the public database then musicians will be asked to sign copyright ownership of the files and, if desired, permission to use identifying information, to the TELMI partners as described below.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

3.2.2 Types of Data (Generated and Collected) and use of Standards

The following data types will be collected:

- **Audio/video recordings (interviews and workshops):** recorded via hand-held recorders into .mp3/.mp4 format. Transcribed to text file (.doc) by project partners or by external services (e.g. www.rev.com).
- **Consent/copyright forms:** delivered, signed, collected, and securely stored in hardcopy.
- **Recordings (performance):** recordings of performance via audio, video, or motion capture will be processed as described in Sections 3.2 and 3.4.
- **Questionnaires:** collected in hardcopy or electronically via the online platform SurveyMonkey (www.surveymonkey.com). The first of these questionnaire can be found in Appendix B of D.2.1.Review of Violin Methods with Complementing Review of Technologies. Data will be stored as .xls, with quantitative data processed via IBM SPSS and qualitative data via NVivo.
- **Violin exercises:** collected as electronic PDFs, converted to .xml format for use in the public database (see 3.4). Exercises will be drawn primarily from the public domain (where composers have been deceased for a period exceeding 70 years, following EU copyright regulations) and, where required, licensing purchased from the publishers.

3.2.3 Data Sharing and reuse

A clear division will be maintained between data collected for research purposes and data intended for public users of the TELMI system.

1. Research Data only: following the guidelines of the BPS, consent forms approved by the Conservatoires UK Research Ethics council will be delivered to participants that guarantee that their anonymity will be maintained within data collected. They will be informed that these data can be used within and in the public dissemination of the project, but all identifying information will be removed. This will include questionnaires and recordings of workshops and interviews.
2. Where audio and video recordings are collected to be used for public dissemination in the database, the musicians will provide both research consent and release copyright ownership of the recordings to the TELMI consortium for use in the public database, project dissemination, and marketing. Copyright forms will be adapted from those used by the RCM Studios. Musicians will have the option to release copyright under the condition of anonymity (with identifying features removed from video recordings), otherwise they must explicitly grant the project the use of their likeness and identifying information.

Where possible, research data will be collected and disseminated following the open data policy of the Royal Society.² Empirical data will be made publically available in an anonymized format through the TELMI Public Database (see 3.4 below) or, if that is not suited for purpose, a publicly available repository such as Dryad or

² See <https://royalsociety.org/journals/ethics-policies/data-sharing-mining/> for the full policy.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

Figshare. The data will not be made publically available in cases where the nature of the information covered might compromise the participants' anonymity. In such cases, we would consider releasing extracts of the data to third parties upon request (e.g. for verification).

3.2.4 Archiving and Preservation

Data will be stored on the project databases as outlined in 3.4 below.

3.3 Public Database Repository Data

During the TELMI project a set of multi-modal recordings of performances will be captured including teachers and students. This data will be used in the TELMI prototypes as well as to refine analysis algorithms developed during the project. The raw data acquired from performances will be analyzed and enriched with feature extraction techniques to build the public datasets to be hosted online.

For this public database, the repovizz platform [1] will be mainly used. Repovizz (<http://repovizz.upf.edu>) is an integrated online system capable of structural formatting and remote storage, browsing, exchange, annotation and visualization of synchronous multi-modal, time-aligned data. Motivated by a growing need for data-driven collaborative research, repoVizz aims to resolve commonly encountered difficulties in sharing or browsing large collections of multi-modal data. At its current state, repovizz is designed to hold time-aligned streams of heterogeneous data: audio, video, motion capture, physiological signals, extracted descriptors, annotations et cetera. Most popular formats for audio and video are supported, while Broadcast WAVE or CSV formats are adopted for streams other than audio or video (e.g., motion capture or physiological signals). The data itself are structured via customized XML files, allowing the user to (re-) organize multi-modal data in any hierarchical manner, as the XML structure only holds metadata and pointers to data files. Datasets are stored in an online database, allowing the user to interact with the data remotely through a powerful HTML5 visual interface accessible from any standard web browser; this feature can be considered a key aspect of repovizz since data can be explored, annotated or visualized from any location or device. Data exchange and upload/download is made easy and secure via a number of data conversion tools and a user/permission management system. The repovizz platform is physically hosted at an internal server in the DTIC-UPF infrastructure.

3.3.1 Description

All datasets in repovizz (public database) include a description field that can contain the information above mentioned. For additional information a web page will be generated containing more structured information and additional fields of all datasets generated during the project containing professional musicians pieces. This web page will contain cross links to the datasets stored in repovizz.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

3.3.2 Types of Data (Generated and Collected) and use of Standards

The data gathered will be mainly consisting in music exercises and pieces that are commonly used as learning material for violin training. RCM will be responsible for selecting the pieces and exercises to be recorded and the professional musicians that will record them.

For audio data professional microphones and bridge pickups will be used, for video low cost and professional cameras will be used and additionally mocap data will be acquired using an electromagnetic fields sensor and an optical motion capture system.

Once the different data streams from different modalities are recorded they need to be time synchronized between them and formatted accordingly to be compatible with formats accepted in repovizz. The following formats are used for each type of data:

- Audio: any common audio format that can be decoded by ffmpeg (wav, mp3, ogg, flac, aac, etc). Once uploaded to repovizz original audio streams are kept in the server but additionally are converted to wav files at a sampling rate of 44.1Khz and 16 bits for audio feature extraction and web friendly mp3 and ogg files are generated.
- Video: any common video format that can be decoded by ffmpeg (mp4, avi, mkv, mov, webm, etc). Once uploaded to repovizz original video streams are kept in the server but additionally are converted to webm and mp4 at a resolution of 720p to make it compatible with standard html5 browsers.
- Time varying Signals / Descriptors: csv containing a header line as defined in repovizz tutorial [2]
- Musical Scores: music xml (compatible with musescore open source software)
- Mocap Data: multiple csv files for each marker coordinate as defined in repovizz tutorial [2]
- Annotations: txt files containing lines with time and label information as defined in repovizz tutorial [2]

3.3.3 Data Sharing and reuse

In the case of TELMI, the public database stored in the repovizz infrastructure will serve as a sharing and visualization platform, allowing third parties to download data as well as visualize it in a user friendly way just opening a url in a browser.

Data sharing and reuse of data will be guaranteed once the data is uploaded to the public database (repovizz). Being repovizz an online web based solution it makes easily to share datasets and individual streams within each dataset.

A RESTful api [3] allows users to access and use all data stored in repovizz programatically.

Using the API users can browse, search, list, and download datasets and individual streams contained inside.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

3.3.4 Archiving and Preservation

All data acquired during the TELMI project and uploaded to the public database (repovizz) will be guaranteed to be available within a minimum of 6 months after project completion. This embargo period is requested to allow time for additional analysis and further publication of research findings to be performed. Nevertheless the data won't be deleted as repovizz platform might be further maintained with other funds after this period.

3.4 Additional Guidelines for the Data Management Plan

Besides simply providing a sharing mechanism, data management in H2020 poses a series of requirements for the access mechanisms to the data, as well as the documentation and characteristics of the data itself. Below we outline our plan to satisfy these requirements.

3.4.1 Discoverability and Accessibility

All data sets stored on the public database are searchable online, and uniquely identified using a randomly generated ID. Also datasets in the public database can be cross-linked through a unique url. This unique url can be included in related publications, deliverable documents, or detailed description documents or web pages that explain their contents and meaningful context information.

Barring specific restrictions imposed by the TELMI partners to ensure that there is no conflict with their strategic planning, the aforementioned datasets will be released under a Creative Commons (CC) license (specific CC license details will be analysed individually as they depend on the contents of each dataset).

3.4.2 Additional Archiving and Preservation Requirements

The Public Database will be hosted in DTIC - UPF server's infrastructure, and it takes advantage of the UPF's storage and backup facilities:

- Data is backed up on a type-class basis: mission-critical (user's data, virtual machines, scientific output, etc) and static (scientific datasets, intermediate files, HPC filesystems, etc).
 - Mission-critical data is backed up:
 - Three times per day, locally (00:00, 08:00, 16:00) and retained for three days. Granularity: 9 (3x3)
 - Once per day, remotely (00:00), to Jaume Primer remote datacenter, and retained for two weeks. Granularity: 15 (15x1)
 - Static data is backed up:



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

- Two times per day (00:00, 18:00), locally, and retained for one week. Granularity: 7 (7x1)
- Backups are processed automatically based on snapshot technology on a time-scheduled basis.
- Standard recovery processes are available: Samba sharing (previous versions), NFS sharing, Qtree and volume restore.

The raw data collected and archived at UNIGE are stored on a dedicated NAS server, with RAID 5 configuration. Data is backup on an offline hard-disk for additional redundancy.

3.4.3 Compliance with Ethics Requirements and Protection of Personal Data

All the TELMI consortium members are well aware of the ethical aspects of the project, and will take into account rules and legislation at national and institutional level in their respective countries when collecting potentially sensitive personal data. In order to enable participants to make informed decisions, detailed documentation (e.g. informed consent and/or terms of participation) will be prepared on a case-by-case basis, outlining the information that will be gathered, the intended use within the project, and any applicable risks associated with a potential public dissemination (e.g. Video data reveals the identity of the user). All data will only be stored, used and/or shared when participants and/or legal entities to which ownership of the data can be credited have given their express informed consent for publication; participants and/or legal entities will be given the option to allow the public dissemination of the data and/or to disseminate the data in an anonymous or traceable way. See section 3.2 Music Education Datasets and Users Feedback above for further details regarding the collection of data from users.

The treatment of personal information and sensitive data will be done in accordance to the Ethics and Security specifications outlined in section 5 of the TELMI proposal. All personal information and data intended to be private will be stored in an administrative database housed in a secure location with appropriate protections by the partner(s) responsible for its collection and/or generation. All participants will be assigned a unique ID, by means of which the anonymized data (when applicable) will be shared within the consortium. All the master records and traceable tokens prone to enable the mapping between the anonymized data and the real identity of a given user will be protected by appropriate safety measures and only accessible to the principal investigator of TELMI within the partner(s) institution(s) responsible for the data collection and/or generation.

4 Conclusion

This deliverable presents a series of guidelines and best practises regarding the Data Management Plan within the consortium, one for each type of dataset that we are planning to release within the timeframe of TELMI. While future updates of the Data Management Plan are expected to add specificity and depth, this



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269

deliverable lays the foundation for data collection, generation and management practices, as well as the sharing conditions of the datasets.

5 References

[1] Mayor, O., Llop J., & Maestre E. RepoVizz: A multimodal on-line database and browsing tool for music performance research. 12th International Society for Music Information Retrieval Conference (ISMIR 2011)

[2] Repovizz Tutorial (<http://repovizz.upf.edu/misc/repoVizzTutorial.pdf>)

[3] Repovizz RESTful API documentation (<http://repovizz.upf.edu/repo/api-doc/index.html>)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688269