FuturePulse

D7.2 – Data management plan v1

February 28tht, 2018

Author/s: Daniel Molina (BMAT)

Contributor/s: FuturePulse Consortium

Deliverable Lead Beneficiary: BMAT

This project has been co-funded by the HORIZON 2020 Programme of the European Union. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use, which may be made of the information contained therein.

Deliverable number or supporting document title	D7.2 Data management plan v1
Type	ORDP: Open Research Data Pilot
Dissemination level	Public
Publication date	28-02-2018
Author(s)	Daniel Molina (BMAT)
Contributor(s)	FuturePulse Consortium
Reviewer(s)	Vasilis Papanikolaou (ATC), Geoffroy Peeters (IRCAM)
Keywords	DMP, Data management plan, ORDP
Website	www.futurepulse.eu

CHANGE LOG

Version	Date	Description of change	Responsible
V0.1	26/12/2017	Skeleton	Daniel Molina
V0.8	14/02/2018	Draft for peer-review	Daniel Molina
V1.0	28/02/2018	First version	Daniel Molina

Neither the FuturePulse consortium as a whole, nor a certain party of the FuturePulse consortium warrants that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

The commercial use of any information contained in this document may require a license from the proprietor of that information

Funded by the European Commission

Table of Contents

Funded by the
European Commission

1 Executive Summary

This document defines the data policy and data management procedures in *FuturePulse: Multimodal Predictive Analytics and Recommendation Services for the Music Industry*, Grant Agreement number 761634 ICT H2020. The purpose of this DMP is to provide an overview of the main elements of the data management policy that will be used by the Consortium with regards to the project research data.

This document corresponds to the first version of the DMP, delivered in Month 6 of the project. The next version of the DMP will be updated in Month 24 (D7.4). Thus, the DMP is not a fixed document but will evolve during the lifespan of the project (from September 1st 2017 to August 31st 2019).

This document follows the Guidelines on FAIR Data Management in Horizon 2020[1] and the Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020[2]. The document follows the template provided by the European Commission in the Participant Portal and will be submitted to European Commission as D7.2 Data Management Plan.

FuturePulse participates in the Open Research Data Pilot[3]. The Open Research Data Pilot aims to make the research data generated by Horizon 2020 projects accessible with as few restrictions as possible, while at the same time protecting sensitive data from inappropriate access.

The consortium has chosen Zenodo[4], the open research repository from OpenAIRE and CERN, as the central scientific publication and data repository for the project outcomes.

[1] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[3] https://www.openaire.eu/what-is-the-open-research-data-pilot

[4] https://zenodo.org

Funded by the European Commission

2 Key terminology: Definitions and acronyms

AWS	Amazon Web Services
EC	European Commission
DMP	Data Management Plan
Data Repository	a) General term used to refer to a destination designated for data storage. b) Setup within an overall IT structure, such as a group of databases, where an enterprise or organization has chosen to keep various kinds of data (https://www.techopedia.com/definition/23341/data-repository).
DOI	Digital Object Identifier: persistent identifier or handle used to uniquely identify objects, standardised by the International Organisation for Standardisation (ISO).
FTP	File Transfer Protocol
GDPR	EU General Data Protection Regulation (https://www.eugdpr.org)
H2020	Horizon 2020 Programme
Metadata	Administrative information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.
Zenodo	Catch-all OpenAIRE compliant repository for EC funded research, hosted by CERN since May 2013 (https://zenodo.org).

Funded by the
European Commission

3 FuturePulse project

3.1 Abstract

Music is one of the fastest evolving media industries, currently undergoing a transformation at the nexus of music streaming, social media and convergence technologies. As a result, the music industry has become a mixed economy of diverse consumer channels and revenue streams, as well as disruptive innovations based on new services and content distribution models. In this setting, music companies encounter daunting challenges in dealing successfully with the transition to the new field that is shaped by streaming music, social media and media convergence. The availability of huge music catalogues and choices has rendered the problems of recommendation and discovery as key in the competition for audience, while the continuous access to multiple sources of music consumption have resulted in a dynamic audience, characterised by a highly diverse set of tastes and volatility in preferences which also depend on the context of music consumption.

To serve the increasingly complex needs of the music ecosystem, FuturePulse will develop and pilot test a novel, close to market music platform in three high-impact use cases:

- Record Labels,
- Live Music,
- Online Music Platforms.

The project will help music companies leverage a variety of music data and content, ranging from broadcasters (TV, radio) and music streaming data, to sales statistics and streams of music-focused social media discussions, interactions and content, through sophisticated analytics and predictive modelling services to make highly informed business decisions, to better understand their audience and the music trends of the future, and ultimately to make music distribution more effective and profitable. FuturePulse will offer these capabilities over a user-friendly, highly intuitive and visual web solution that will enable the immersion of music professionals in the realm of music data, and will support them to make highly informed and effective business decisions.

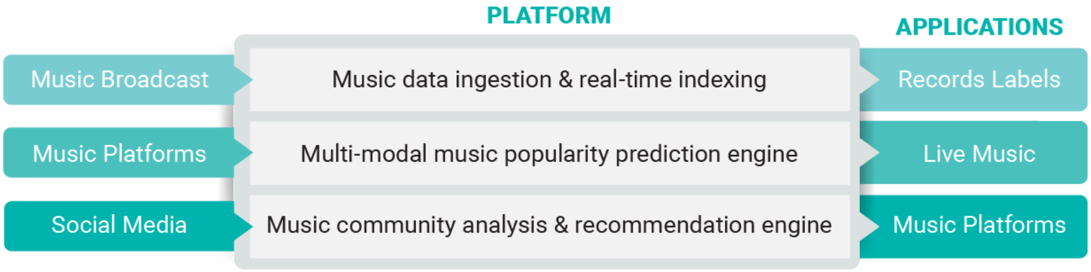
3.2 Project Scope and Objectives

In response to the industrial needs of the music industry the FuturePulse project has identified the following six specific technological and innovation objectives:

- Objective 1: Deliver a single tool for collecting and accessing music data from a diverse set of sources.
- Objective 2: Deliver a set of data-driven services for estimating the current and future popularity of songs, artists and genres.
- Objective 3: Deliver a set of services for enhanced audience analysis and management.
- Objective 4: Integrate music data collection, mining, and visualisation in a scalable Software-as-a-Service (SaaS) platform.
- Objective 5: Perform large-scale pilots on three clearly defined music segments.

Funded by the
European Commission

- Objective 6: Develop and execute a comprehensive dissemination and exploitation plan and pave a clear path to market.


PLATFORM
APPLICATIONS
Music Broadcast
Music Platforms
Social Media
Music data ingestion & real-time indexing
Multi-modal music popularity prediction engine
Music community analysis & recommendation engine
Records Labels
Live Music
Music Platforms

The project will result in a number of high-quality outcomes that will form the basis for the exploitation plan of the project. These are illustrated in figure above and include the following:

- A robust and extensible multi-source music data ingestion and real-time indexing framework.
- A multi-modal music popularity prediction engine: This will produce short- and long-term predictions for popularity indices about specific artists, albums, songs, styles and genres, given a variety of incoming signals.
- An online music community analysis framework and a music recommendation engine.
- An integrated scalable cloud-based platform offering the full spectrum of FuturePulse services.
- Three market-driven applications serving the needs of record labels, event organisers and music platform.

Funded by the European Commission

3.3 Project participants

BMAT licensing S.L.	Daniel Molina dmolina@bmat.com	bmat
Athens Technology Center S.A	Vasilis Papanikolaou v.papanikolaou@atc.gr	ATC ATHENS TECHNOLOGY CENTER
Ethniko Kentro erevnas Kai Technologikis Anaptyxis	Akis Papadopoulos papadop@iti.gr	iti Information Technologies Institute
Musimap S.A.	Frédéric Notet frederic@musimap.com	musimap cognitive technologies
Institut de Recherche et de Coordination Acoustique Musique	Geoffroy Peeters Geoffroy.Peeters@ircam.fr	ircam Centre Pompidou
Playground Music Scandinavia AB	Anders Engström anders@playgroundmusic.com	playground music scandinavia
Bass Nation	Tommy Vaudecrane tommy.vaudecrane@gmail.com	Agency
Sountrack Your Brand Sweden AB	Daniel Johansson danielvinkar@gmail.com	Soundtrack Your Brand®

3.4 Coordinator contact

Daniel Molina	BMAT	dmolina@bmat.com

Funded by the European Commission

4 Data Summary

FuturePulse will produce several datasets during the lifetime of the project. The main purpose of these datasets is to feed the three use cases of the project: Record Labels, Live Music and Online Music Platforms[5]. The data will be both quantitative and qualitative in nature and will be analysed from a range of methodological perspectives for project development and scientific purposes.

As it is a work in progress through the first stages of the project, the first version of the FuturePulse DMP does not include data repositories or metadata about the data being produced in the project. Access to this metadata will be provided in an updated version of the DMP (D7.4, M24).

4.1 Data collection

The data consists of FuturePulse deliverables and other project documents. These documents are created by FuturePulse partners within the project, using their knowledge, raw data collected during the project, and other sources of information.

In FuturePulse, data sets and the actual FuturePulse architecture are being produced based on the requirements from FuturePulse use cases, defined in deliverable D1.2 FuturePulse requirements v1 (M6).

4.2 Data formats/size

FuturePulse datasets will be available in a variety of easily accessible formats, including Audio, Video, Post Script, Excel, Word, Power Point and images.

Given the large-scale nature of the use cases, we expect to generate a large amount of data during the project.

4.3 Origin of data

Data sources

The project will use heterogeneous data sources, ranging from audio tracks and annotated data for audio tracks or social media anonymised data. The specific list of data sources is under discussion at the moment of writing this deliverable and will be ready and included in the deliverable D2.1 – Data specifications and collection v1 by M9 (May 2018).

Public information and deliverables

Public project documentation will be available via FuturePulse web site www.futurepulse.eu.

All project deliverables are by default public, except those marked confidential in the Grant Agreement. FuturePulse documentation will be kept for later use after the end of the project, unless specified otherwise.

4.4 Data Storage and back-up

During the first stages of the project, FTP and AWS servers from BMAT have been provided to store audio tracks and annotated data. FuturePulse website is hosted and backed up by ATC partner.

[5] http://www.futurepulse.eu/use-cases

Funded by the European Commission

4.5 Surveys and questionnaires

The project will perform online surveys under the activity T1.2 FuturePulse co-design process within WP1, in order to contribute to provide a concrete impression of the envisioned capabilities and services offered by the FuturePulse platform. More information about these surveys is provided in the deliverable D1.2 FuturePulse requirements v1 (M6).

Funded by the
European Commission

5 FAIR data

FuturePulse participates in Open Research Data Pilot, which requires the policy of FAIR data (findable, accessible, interoperable and re-usable research data).

5.1 Making data findable, including provisions for metadata

Each dataset generated during the project will be recorded in an Excel spreadsheet with a standard format. The spreadsheet will be hosted at Zenodo. Search keywords will be provided when the dataset is uploaded to Zenodo which will optimise possibilities for re-use. Zenodo follows the minimum Data Cite metadata standards.

A DOI will be assigned to datasets for effective and persistent citation when it is uploaded to the Zenodo repository. This DOI can be used in any relevant publications to direct readers to the underlying dataset.

In order to clearly describe the content of the data, FuturePulse naming convention for project datasets will follow a similar approach to the deliverable naming convention described in the deliverable *D7.1 Quality Assurance and Risk Management Plan* for supporting documents:

FuturePulse_SPD_DATASET” + dataset_number + dataset_name + dataset_version

Some examples:

FuturePulse_SPD_DATASET_01_Pilot1-results_v1.xls

FuturePulse_SPD_DATASET_Annotated_audios_v1.xls

The data will be accompanied with metadata clarifying the meaning of the data and how the data has been collected. The metadata can be provided without the actual data, if requested.

Datasets metadata will follow the META−SHARE[6] schema for data sets description. META−SHARE is an open resource exchange infrastructure. The following data set description is based on the DMP template circulated by CRACKER[7].

Metadata	Description
Resource Name	Complete title of the resource
Resource Type	Conceptual resource
Media Type	The Physical Medium of the content representation, e.g. video, audio, text, numerical data, etc.
Language(s)	The language(s) of the resource content
License	The licensing terms and conditions under which the tool/service can be used
Distribution Medium	The channel used for delivery or providing access to the resource, e.g. accessible through interface, downloadable, CD/DVD., etc.

[6] http://www.meta-net.eu/meta-share

[7] http://cracker-project.eu

Funded by the European Commission

Usage	Foreseen use of the resource for which it has been produced
Size	Size of the resource with regards to a specific size unit measurement in form of a number
Description	A brief description of the main features of the dataset

Table 1: FuturePulse Dataset metadata

The specific metadata contents, formats and schema may be further refined in the future versions of the DMP.

5.2 Making data openly accessible

The consortium has chosen Zenodo, the open research repository from OpenAIRE and CERN, as the central scientific publication and data repository for the project outcomes. The repository has been designed to help researchers based at institutions of all sizes to share results in a wide variety of formats across all fields of science. Furthermore, Zenodo supports DOI versioning, allowing users to update the record's files after they have been made public and researchers to easily cite either specific versions of a record or to cite, via a top-level DOI, all the versions of a record[8].

Research data, specially those needed to validate the results of scientific publications, will be deposited in the Zenodo repository.

Zenodo enables users to:

- easily share the long tail of small data sets in a wide variety of formats, including text, spreadsheets, audio, video, and images across all fields of science
- display and curate research results, get credited by making the research results citable, and integrate them into existing reporting lines to funding agencies like the European Commission
- easily access and reuse shared research results
- define the different licenses and access levels that will be provided

Furthermore, Zenodo assigns a Digital Object Identifier (DOI) to all publicly available uploads, in order to make content easily and uniquely citable.

5.3 Making data interoperable

The data will be stored in a format readable by commonly used data management tools or office software. Direct automatic interoperability of the data with other external data sets is not sought for.

5.4 Increase data re-use (through clarifying licences)

All the research data will be of the highest quality, have long-term validity and will be well documented in order other researchers to be able to get access and understand them.

The datasets will be made available for re-use through uploads to the Zenodo community page for the project. General policies of Zenodo will apply for content, access and reuse, removal and longevity[9].

[8] http://blog.zenodo.org/2017/05/30/doi-versioning-launched

[9] http://about.zenodo.org/policies

Funded by the European Commission

As the processed data will be public, no licences are needed for re-use of the data, as long as the data source is acknowledged.

Funded by the
European Commission

6 Allocation or resources

The FuturePulse datasets stored in Zenodo will be preserved a minimum of 5 years according to the European Commission Data Deposit Policy. There are currently no costs for archiving data in this repository.

Each partner will authorise a responsible of data management who will take the responsibility to control the correct storage, management, sharing and security of the dataset.

As all social media data within FuturePulse will be aggregated and no personal information will be stored by the platform, FuturePulse will not have its own Data Protection Officer according to GPDR legislation[10]. Daniel Molina and Joaquín Luzón at BMAT are responsible for the creation, management and updates of the Data Management Plan in FuturePulse project.

[10] https://www.privacy-regulation.eu/en/article-39-tasks-of-the-data-protection-officer-GDPR.htm

Funded by the
European Commission

7 Data security

7.1 Data storage

For the duration of the project, datasets will be stored on the responsible partner's centrally provided storage, summarised in the table below.

FuturePulse partner	Data Storage
BMAT	1. Data from Vericast (accessible through the Vericast Website): - Metadata: in redundant MySQL Database - Recordings and references: in NFS storage systems 2. Playground Music Catalog: - FTP Server with NFS storage backend
ATC	1. ATC will provide a storage mechanism to save output produced by the analysis components so that it can be used as a common facility to save generated data concerning analysis of audio tracks, predictions and Media statistics that will be gathered during the project. This will be a graph-based storage which can support partners' existing solutions to a large extent but could also be accompanied by another RDBMS or NoSQL database according to the emerging needs. 2. A temporary cloud storage will be considered in order to store the actual audio files that will be used to train the components that require training as well as for new audio files that are going to be uploaded by users. A cloud option is considered for this file storage so as to have a cost-effective and easily accessible solution to make large volumes of audio files available to many components. The provision of an API to upload new files to the storage will be also taken under consideration for the selection of the appropriate cloud service.
CERTH	CERTH is currently using four workstations for the temporary storing of data: 1. Charts collected by CERTH are stored temporarily as files in a windows workstation. 2. Predictions about artists and genres from Google trends and charts are stored temporarily as files in a windows workstation. 3. Sources from social media and streaming platforms (e.g. channels, accounts, etc) and their associations with Future Pulse targeted artists are stored temporarily in a relational DBMS in a Linux workstation. The same workstation keeps events and artists from Facebook in a graph database (Neo4j). Tweets about music genres are also stored in this workstation. Data coming from the tracking social media and streaming platforms will be stored also in the same workstation in the appropriate NoSQL solution. 4. Results from Twitter data analysis (demographics predictions) are stored temporarily in a windows workstation. All aforementioned workstations are located in CERTH premises and are part of its internal network, where standard security mechanisms

Funded by the European Commission

	and policies are applied. All the above data and any new data that will appear at later stages of FuturePulse, will eventually be migrated in a linux server, also located in CERTH. These data will be exposed by a REST API with the appropriate authentication/authorisation level.
MUSIMAP	MUSIMAP is currently using dedicated private storage hosted by AWS 1. Audio files are temporary stored in a dedicated and secured S3 storage in order to be analysed. 2. Analysis results are also stored in a dedicated and secured S3 storage and/or secured database 3. Meta-data are stored in a dedicated Neo4j database All aforementioned servers are hosted on Amazon Web Services Cloud and are part of a Musimap dedicated private network, where high security mechanisms and policies are applied. All the above data and any new data that will emerge at later stages of FuturePulse, will be exposed by a REST API with the appropriate authentication/authorisation level.
IRCAM	IRCAM will not store data available for the project. IRCAM will store internally data for the purpose of developing Machine-Learning algorithm, this will not be accessible outside a restricted number of people (for copyright reasons). IRCAM web-service will not store data but will remove immediately all processed audio file.
Playground	Playground do not have data for this project in itself. Our data is in our IT-systems, and are exported and shared with the project as needed. In practise, all data is collected by the technical partners themselves through APIs.
BASS NATION	Bass Nation do not have data for this project in itself.
SYB	The SYB use case will, as I understand it, be temporarily stored by technical partners during the research process, f.e. IRCAM and Musimap audio analysis, CERTH charts analysis and BMAT broadcast analysis. For the pilots, sales data from SYB pilot partners will be stored locally on our own servers for the analysis.

Table 2: Data storage

After the completion of the project, datasets will keep stored and freely accessible in the Zenodo repository for long term preservation and curation.

7.2 Personal data

FuturePulse datasets do not include personal information. All Social media information used in the project is aggregated after its collection, and as a result no personal information will be stored by the FuturePulse platform.

Some surveys within the project (see Section 4.6) are using Cint[11], a survey platform that connects community owners to researchers, agencies and brands, for the sharing and accessing of consumer data.

Apart from Cint, and whenever applicable, the consent of the survey participant will be asked in all surveys conducted within FuturePulse. The consent will include a

[11] https://www.cint.com

Funded by the European Commission

description how and why the data is to be used. The survey participants will not include children or other groups needing a supervisor.

Funded by the
European Commission

8 Ethical aspects

Future WP8 Ethics requirements sets out the 'ethics requirements' that the project must comply with. There are two deliverables associated to this Work Package:

- D8.1: POPD – Requirement No. 1:

A report on ethical considerations in regards to social media, addressing address

ing general ethical issues regarding social media research, ethical and legal issues in regards to informed consent

by participants, ethical issues and consent in regards to children and young people.

- D8.2 H – Requirement No. 2: Providing details on the procedures and criteria used to identify/recruit research participants and informed consent procedures

FuturePulse questionnaires do not include any form of personal data. Therefore, no personal data is subject to be shared or long term preserved.

Funded by the
European Commission

9 Other issues

Use of national procedures for data management

IRCAM makes use of Archive ouverte HAL[12], the French national repository for scientific publications.

[12] https://hal.archives-ouvertes.fr

Funded by the
European Commission