New Frontiers in Music Information Processing (MIP-Frontiers)

Grant Agreement Number: 765068

D6.2 – Data management Plan

- Title — Data Management Plan
- Lead Beneficiary — QMUL
- Nature — ORDP: Open Research Data Pilot
- Dissemination level — ***Non confidential – Website version***

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Table of contents

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

1. Table of changes

Version	Date	Responsible	Changes made and comments
V 0.9	April 2nd 2019	Simon Dixon	Data management plan first version Sent to consortium for comments on April 2nd
V 0.9.1	April 6th 2019	Simon Dixon	Proof reading and edits
V 1.0	April 24th 2019	Simon Dixon	Review of the Data Management Plan

2. Acronyms

DMP	Data Management Plan
ESR	Early Stage Researcher
MIP	Music Information Processing
MIR	Music Information Retrieval
WP	Work Package
MSCA	Marie Skłodowska-Curie Action
ITN	Innovative Training Network
EU	European Union
QMUL	Queen Mary University of London
UPF	Universitat Pompeu Fabra
JKU	Johannes Kepler University
TPT	Telecom ParisTech

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

3. Summary

The aim of the data management plan is to describe a virtual environment for all researchers to access, store, manage, analyse and re-use data for research, innovation and educational purposes. This plan describes the handling of data, research data and data for research, and the guiding principles to make it Findable, Accessible, Interoperable, and Reusable (FAIR). The plan will evolve with the project life and change as the project does.

MIP-Frontiers is a project on Music Information Processing (MIP), and by its definition it involves the use of information (or data) processing methodologies. In addition, one of the aims of the project is to address an issue of MIP, that standard methods for most music information processing tasks have been developed and tested on small datasets, thus the methods tend to be neither robust nor scalable to industrial scale datasets (WP1 in particular deals with data-driven methods).

It is understandable that good data management, and a good data management plan, is a key and essential part of MIP-Frontiers and of all 15 ESR projects individually. As so:

- One of the first training topics at the first training week for the ESRs was software carpentry (Software development best practices for reproducible research) including data management (code, open data, backups…).
- All ESRs were asked to review and complete individually the European Commission template for a Data Management Plan. Their individual plans are attached to this Data Management Plan.

4. Introduction and state of the project

MIP-Frontiers will follow the guidelines given by the EC on FAIR data and each academic institution's policy on data management. It is the purpose of the EC, and also MIP-Frontiers' aim, that the results, scientific publications and the data behind the scientific outputs are open by default (with some reasons for opt-out) and thus broadly accessible. This has several advantages:

- the validation of research results (e.g. for peer review) becomes easier
- scientific breakthroughs become more visible
- research results will be more cited and therefore have a greater impact on ESR careers
- duplication of research activities will be avoided which improves the quality of results
- research data is preserved
- EC research funds are better valued and visible for society
- research is better distributed across scientific fields which helps to solve complex (social) challenges.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Data management is fundamental during any research and researcher career. It is important to understand what it involves: research data refers to information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. As well as the advantages above, the DMP will help each individual ESR project to:

- Save the ESR's time
- Ensure integrity and reproducibility of data
- Disseminate and preserve the data
- Avoid possible loss of data

At this point, March 2019, the project is at month 12; all ESRs have been enrolled for at least 3 months, and most of them for 6 months. After their Stage 0 thesis review, December 2018, they are starting to carry out research and work with data and databases.

It is the perfect moment to start their data management plan, although still some of them do not know which data or databases they will use, especially during their secondments. At their Stage 1 review, they will need to write a data summary point.

5. Updates and new versions of the DMP

The first version of the Data Management Plan (DMP) for MIP-Frontiers, which provides an individual analysis of all projects and of the main aspects to be followed by the project's data management policy was compiled in March 2019, month 12 of the project.

The DMP needs to be updated over the course of the project whenever significant changes arise. The beneficiaries will be responsible for the data management of their ESRs and communicate it to QMUL as coordinator of DMP. Significant changes that will need to be added to the DMP are (but not limited to):

- New data
- Changes in consortium policies (e.g. new innovations, patents or license applications)
- Changes in consortium composition and external factors.

Major changes will be included as they happen, and uploaded, and it is also foreseen that after the project thesis Stage 1, 2 and 3 a standard follow up review will be made, where minor changes will be incorporated.

Consortium partners will be responsible for the data management of their ESRs and communicate it to QMUL, as coordinator of DMP, as it evolves throughout the duration of the project.

Data protection and privacy, the communication approach, informed consent and relevant regulations on data security are covered in the ethics deliverables.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

6. Training and information on DMP

Data and the data management plan are crucial for any MIR project; MIR is by definition information processing. At the first training week in Paris at ISMIR 2019, a 2-hour training course on software carpentry including data management (code, open data, backups…) took place.

The software carpentry training was entitled Software Development Best Practices for Reproducible Research, and included the following:

- code, open data, backups, arguments, source control, versions, testing and reproducible research.

In Annex I there is a list of training and lecture materials on data management and data management planning that has been forward to the ESRs for training and information purposes before they completed their first data management plan.

7. Data Management

The following section for data management will include the policy and objectives of MIP-Frontiers for data management and the information that the individual data management plan needs to have, and also explain some terminology and the scope of the plan.

Each of the 15 PhD projects at MIP-Frontiers involves a specific Beneficiary and Partner with particular needs and interests, and therefore different data will be generated and used for each project. Each ESR describes their data management in their individual data management plan (attached in annex III).

a. Data Summary

This section gives, at project and individual level, an overview about the research data which is generated, collected, processed and stored during the MIP-Frontiers project. This includes the data description for different types and formats, purpose with respect to project objectives and tasks, preparation for data re-use, data origin, expected data size, and to whom it might be useful.

Data generated in MIP-Frontiers should be strictly digital. In general, the data file formats to be used shall meet the following criteria:

- widely used and accepted as best practice within the specific discipline,
- self-documenting, i.e. the digital file itself can include useful metadata,
- independent from specific platforms, hardware or software.

At the level of corpora/datasets there are open initiatives that MIP-Frontiers will use; Jamendo, Freesound and AcousticBrainz are maintained by partners of this proposal (will be used by UPF1, UPF2, UPF3). For other projects the industrial partners will make available data (DRM Score Cloud Song database will be used for QMUL1 and QMUL2).

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

At the level of software tools with which to extract musical features from audio recordings, Essentia, developed at the UPF, and Sonic Visualiser / Annotator, developed at QMUL are open tools that will be used and further developed.

b. FAIR data

F - Making data findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of making data "FAIR".

During the MIP-Frontiers life cycle, ESRs and their supervisors should ensure and try to make data and metadata findable by means of:

- Assigning (meta)data a globally unique and persistent identifier.
- Describing data with rich metadata.
- Registering or indexing (meta)data in a searchable resource.
- Specifying the metadata using standard identifiers.

A - Making data openly accessible

All data produced in the MIP Frontiers project will be openly accessible, unless there is a conflict of interest with a beneficiary or partner; in this case it will be specifically explained in the ESR's DMP attached. Data sets and metadata (anonymised) will be placed in a repository accessible from the project website, such as institutional repositories and the open data platform Zenodo.

These kinds of repositories allow researchers to store and publish both research outputs and data, while providing tools to link them. In addition, all content on the project web site will be licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, 2019.

During the MIP-Frontiers life cycle, ESRs and their supervisors should ensure and try to make data and metadata openly accessible by means of:

- Retrievable (meta)data by their identifier using a standardized communications protocol.
- Open, free, and universally implementable protocols.
- Allowing the protocol for an authentication and authorization procedure, where necessary.
- Accessible metadata, even when the data is no longer available.

MIP-Frontiers does not have a data access committee. All researchers involved in data gathering will become familiar with the guidelines of open data platforms and any data/data management query will be addressed to the supervisor, and if necessary to the project Supervisory Board for discussion.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

I - Making data interoperable

Interoperable data means that the exchange and reuse of data enabled by OA is possible. Therefore, standard data vocabulary and formats which are compliant with available open software have to be used for storing and providing the data. OA of data which can only be used with special and restricted software makes no sense. Interoperability will make data exchange between researchers, institutions or countries possible and will allow the re-combination of different datasets from different origins.

During the MIP-Frontiers life cycle, ESRs and their supervisors should ensure and try to make data and metadata openly accessible by means of:

- (meta)data using a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (meta)data using vocabularies that follow FAIR principles.
- (meta)data including qualified references to other (meta)data.

R - Increase data re-use

The MIP-Frontiers project will make its data produced available on platforms such as Zenodo (https://zenodo.org). The project will follow the platform's guidelines to optimize the possibilities for re-use. All data will follow a clear version numbering structure, if needed. For all quantitative and qualitative research in the project regarding people, non-identifiable metadata will be produced and made available on the aforementioned platform. Metadata will describe instruments used, methodologies employed and goals and target groups of the research. Metadata will be collected and appropriately stored by the researchers. All personal data will be anonymised.

All data produced in the MIP-Frontiers project will be open access, unless there are restrictions from industry partners. In case of publication (and patents), data will be made available after the final publication (and/or acceptance of patent). The data produced and used in the project will be usable by third parties, before and after the end of the project.

During the MIP-Frontiers life cycle, ESRs and their supervisors should ensure and try to make data and metadata openly accessible by means of:

- meta(data) having a plurality of accurate and relevant attributes.
- (meta)data released with a clear and accessible data usage license.
- (meta)data associated with their provenance.
- (meta)data meeting domain-relevant community standards.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

c. Allocation of resources

The data will be openly accessible as far as this is financially viable. Open access publications will be covered from the Research and Training budget of each ESR managed by their employer. MIP-Frontiers has no centrally allocated resources for this purpose. Data repositories that are free or that charge a one-off fee will be considered (so that data persists beyond the end of the project).

Decisions on what data will be kept and for how long will be made by supervisors and where necessary discussed by the project Supervisory Board.

d. Data Protection

All beneficiaries and partners are required by law to comply with the General Data Protection Regulation (GDPR) and national data protection laws.

According to the GDPR each partner is responsible for data security of the data they gather within their organisation. By means of example, for the coordinating institution, QMUL, data protection regulations can be found on their web site:

- QMUL, data protection regulations: http://www.arcs.qmul.ac.uk/governance/information-governance/data-protection/gdpr/

e. Ethical aspects

Ethical aspects are cover by the Ethics deliverables and managed by the Ethics Committee.

f. Other issues

Any foreseeable data and data management issues during secondments should be covered and described in the secondment agreement.

8. Diffusion and Communication

A publishable version of the DMP will be uploaded at the website for communication and diffusion, as a best practise for other projects. It will include when possible lessons learned and best practices.

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Annex I. Information and training for ESRs on DMP

The following Information material and links have been sent to the ESRs for information and training purposes:

- Horizon 2020 Data Management Plan (DMP) Template

  http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template

- Digital Curation Centre DMP checklist

  Checklist for a Data Management Plan. Digital Curation Centre UK

- UK Data Archive checklist

  Data Management checklist.

- **Further reading:**

  Arms, C. R., Fleischhauer, C., Murray, K. (2013). *Sustainability of digital formats: Planning for Library of Congress collections*. Library of Congress. Compilation. Last updated 24 July 2013. Retrieved from www.digitalpreservation.gov/formats

  Beagrie, N., Houghton, J. (March 2014). *The value and impact of data sharing and curation - synthesis of three recent UK studies* (v.1.0). Jisc. Retrieved from repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

  Beagrie, C. (2013). *Keeping research data safe: cost / benefit studies, tools, and methodologies focussing on long-lived data*. Retrieved 4 August 2014 from www.beagrie.com/krds.php

  Digital Curation Centre (DCC). (2010). *Data management plans*. Retrieved 4 August 2014 from www.dcc.ac.uk/resources/data-management-plans

  Drummond, C. (2009). 'Replicability is not reproducibility: nor is it good science' in *Proceedings of the Evaluation Methods for Machine Learning Workshop*, 26th ICML, Montreal, Canada. Retrieved from cogprints.org/7691

  European Commission (EC). (11 December 2013). *Guidelines on open access to scientific publications and research data in Horizon 2020* (v.1.0). Retrieved from ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

  Jones, S. (2011). *Develop a Data Management and Sharing Plan*. 8 September 2011. Retrieved from www.dcc.ac.uk/resources/how-guides/develop-data-plan

  UK Data Archive (UKDA). *Create and manage data: planning for sharing*. Retrieved 4 August 2014 from www.data-archive.ac.uk/create-manage/planning-for-sharing

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Swan, Alma (2010) The Open Access citation advantage: Studies and results to date s.n. https://eprints.soton.ac.uk/268516/

Budapest Open Access Initiative https://www.budapestopenaccessinitiative.org/

- **Training videos**

  Videos suggested in QMUL online training module for data management plans.

  Data Management: Preparing data management plans for funding applications

  https://www.youtube.com/watch?v=OoFxlpqjKV4

  MANTRA: Data management plans videos

  https://www.youtube.com/playlist?list=PL2XF5RiVI7GOyWMmvh9Lcyi1JZ0rJ-Ldo

- **Data management information on the Beneficiaries' web sites**

  UPF DMP information

  https://guiesbibtic.upf.edu/data/en/dmp

  QMUL DMP information

  https://www.library.qmul.ac.uk/research/research-data-management/planning-your-data/

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Annex II. Checklist for data management update

DATA MANAGEMENT Check list

ESR Name: Name and Last name

Project Acronym: QMUL1, TPT1

1. Data Summary			Yes	No
Is there any new data from your previous DMP check/Thesis Stage?				
· Do you have any new data collection/generation from your previous DMP review?				
· Have you made any changes in the types and formats of data that you have generated/collected?				
· Have you re-used any existing data from your previous DMP review?				
If you answer yes to any question above,				
· Have you updated your Data Management Plan?				
2. FAIR data				
Please, check that you comply with the following FAIR data principles:				
To be Findable:				
· Assign (meta)data a globally unique and eternally persistent identifier				
· Describe data with rich metadata				
· Register or index your (meta)data in a searchable resource				
· Specify the metadata data identifier				
To be Accessible:				
· Retrievable (meta)data by their identifier using a standardized communications protocol				
· Open, free, and universally implementable protocol				
· Allowing the protocol for an authentication and authorization procedure, where necessary				
· Accessible metadata, even when the data are no longer available				
To be Interoperable:				
· (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation				
· (meta)data use vocabularies that follow FAIR principles				
· (meta)data include qualified references to other (meta)data				
To be Re-usable:				
· meta(data) have a plurality of accurate and relevant attributes				
· (meta)data are released with a clear and accessible data usage license				
· (meta)data are associated with their provenance				
· (meta)data meet domain-relevant community standards				

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Annex III. ESR individual Data Management Plans

The ESRs were asked to complete the following data management plan template; it includes 7 data management topics, each one having several question that has/will help them complete the plan; if needed the questions are answered and commented on the ESR DMP. After the individual data management plan template, all ESR plans are included.

DATA MANAGEMENT PLAN

ESR Name: [name and last name]

Project Acronym: [e.g. QMUL1, TPT1]

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

What types and formats of data will the project generate/collect?

Will you re-use any existing data and how?

What is the origin of the data?

What is the expected size of the data?

To whom might it be useful ('data utility')?

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

What naming conventions do you follow?

Will search keywords be provided that optimize possibilities for re-use?

Do you provide clear version numbers?

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

What methods or software tools are needed to access the data?

Is documentation about the software needed to access the data included?

Is it possible to include the relevant software (e.g. in open source code)?

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Have you explored appropriate arrangements with the identified repository?

If there are restrictions on use, how will access be provided?

Is there a need for a data access committee?

Are there well described conditions for access (i.e. a machine readable license)?

How will the identity of the person accessing the data be ascertained?

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

How long is it intended that the data remains re-usable?

Are data quality assurance processes described?

Further to the FAIR principles, DMPs should also address the following points.

3. Allocation of resources

What are the costs for making data FAIR in your project?

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Is the data safely stored in certified repositories for long term preservation and curation?

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

6. Other issues

Do you make use of other national/funder/sectoral/departmental procedures for data management? If yes, which ones?

7. Further support in developing your DMP

The Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.

The EUDAT B2SHARE tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

Useful listings of repositories include:

Registry of Research Data Repositories

Some repositories like Zenodo, an OpenAIRE and CERN collaboration, allow researchers to deposit both publications and data, while providing tools to link them.

Other useful tools include DMP online and platforms for making individual scientific observations available such as ScienceMatters.

List of each ESR individual Data Management Plan

ID	Name	Host
QMUL1	Emir Demirel	QMUL
QMUL2	Carlos Lordelo	DoReMIR
QMUL3	Ruchit Agrawal	QMUL
QMUL4	Alejandro Delgado	Roli
QMUL5	Vinod Subramanian	QMUL
UPF1	Philip Tovstogan	UPF
UPF2	Antonio Ramires	UPF
UPF3	Furkan Yesiler	UPF
TPT1	Karim Ibrahim	TPT
TPT2	Kilian Schulze-Forster	TPT
TPT3	Giorgia Cantisani	TPT
TPT4	Ondřej Cífka	TPT
TPT5	Javier Nistal	Sony
JKU1	Luís Carvalho	JKU
JKU2	Charles Brazier	JKU

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765068.