Data Management Plan

Created with DMPonline.be

Filippo Carnovalini - DINF - Vrije Universiteit Brussel
Supervisor: Geraint A. Wiggins
Version 1.0 – 19/03/2024
Project CALIOPE - Grant agreement ID 101108690
Deliverable D4.1 – Work Package 4 (Management)
Dissemination Level: Public

Funded by the European Union, under action HORIZON-MSCA-2022-PF-01

Versioning
1.0. (19/03/24) Initial Version

Summary

The Data Management Plan describes the data that is anticipated to be generated during the project, and how it will be managed, stored, and preserved.

1. Data summary

1.1 Will you re-use any existing data and what will you re-use it for?

I will reuse music scores available in the public domain to build a dataset of musical exercises. It is important to use widely used exercises, so that their pedagogical value is established, and thus it is important to reuse data for this.

1.2 What types and formats of data and other research outputs will the project generate or re-use?

#	DATA TYPE	FORMAT	New/reused
1	Annotated Music Scores	text (xml)	reused/redacted, and new
2	Metadata	text	new
3	Interviews, questionnaires	text, audio	new
4	Informed consent	text	new
5	Code	text	new
6	Project website	text, audio, image, video	new
7	Social media	text, audio, image, video	new
8	Reports	text	new

1.3 What is the purpose of the data generation or re-use and its relation to the objectives of the project?

1 and 2 are needed for the design of the music generation system and for automated difficulty analysis of music.
3 and 4 are related to the definition of a Learner model through the use of focus groups of music teachers.
5 is the computational result of both of the above
6 and 7 are needed for effective communication and dissemination
8 is required by the funding programme.

1.4 What is the expected size of the data that you intend to generate or re-use?

Given the lightweight formats used, 1-5 are expected to be within the order of 10GB, with the audio form 3 representing the most size-consuming.
8 is similarly easily within 1GB.
6 and 7 may contain more multimedia, which can make their size comparatively bigger but still within the order of GBs, as these data are meant for web usage which requires lightweight data.

1.5 What is the origin/provenance of the data, either generated or re-used?

1 will require bibliographical research of available music-education related literature, and will probably include data from different sources which will require editing and revision.
The remaining data will either be produced by the researcher directly or by the expert collaborators involved in the research.

1.6 To whom might your data be useful ('data utility'), outside your project?

Other researchers interested in computer-assisted education may find interest in 1, 2, 5, and 3 (to the extent to which the interviews will be made available).
Researchers interested in music generation can also find 1, 2, and 5 useful.
Music students and teacher may find 1 useful as it will be a relatively large collection of musical exercises.
6 and 7 will be useful to the general public as a way to be informed on state-of-the art technology.

2. FAIR data

2.1 FAIR data: Making data findable, including provisions for metadata

2.1.1 Will data and other research outputs be identified by a persistent identifier?

- Yes: describe below

The main dataset of musical scores and their metadata will be assigned a DOI.

2.1.2 Will rich metadata be provided to allow discovery?
What metadata will be created?
What disciplinary or general standards will be followed?
In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
Metadata regarding the dataset will describe both the bibliographical data about the original sources from which the scores come from, and additional musicological information.
This information will be described by the use of Dublin Core.

2.1.3 Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

- Yes: describe below

Appropriate keywords will be selected to make the data easily findable by all the categories of interested users (see 1.6).

2.1.4 Will metadata be offered in such a way that it can be harvested and indexed?

- Yes: describe below

Through the use of an ontology the metadata will be easily indexed.

2.2 FAIR data: Making data accessible

2.2.1 Will the data and other research outputs be deposited in a trusted repository?

- Yes: describe below

Public data will be deposited and made available on Zenodo.
Restricted data will be stored in VUB's institutional repositories, namely Sharepoint.
Code will be stored privately in appropriate repositories for development, and then made available either through code repositories.

2.2.2 Have you explored appropriate arrangements with the identified repository where your data and other research outputs will be deposited?

- No

Given the standard requirements of this project, there is no need for special arrangements.

2.2.3 Does the repository ensure that the data and other research outputs are assigned an identifier? Will the repository resolve the identifier to a digital object?

Zenodo provides DOIs to appropriately identify digital objects.

Code repositories usually do not employ DOIs but provide URLs to identify code releases and milestones, which will be employed by this project.

2.2.4 Will all data and other research outputs be made openly available?

- No, certain datasets cannot be shared openly for the following reasons:

The data from the interviews to experts will contain some personal data, although the large amount of data will be non-sensitive. Therefore it will be divided into two parts, having sensitive data not be made publicly available. The results of the interviews and the data in elaborated, aggregated form will be made available through scientific publications.

Some parts of the scores dataset may be subject to copyright. While a reasonable effort to only include data from the public domain will be made, it may be possible that some of the scores will need to have restricted access.

2.2.5 Is an embargo applied to give time to publish or seek protection of the intellectual property (e.g. patents)?

- No

2.2.6 If an embargo is applied (see question 2.2.5), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

N.A.

2.2.7 Will the data and other research outputs be accessible through a free and standardized access protocol?

- Yes: describe below

An appropriate license will be provided (e.g. CC0) on all the data publicly shared.

2.2.8 If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

Personal data, in its raw form, will not be given access to other researchers unless necessary for the project.

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

Sensitive data will be stored solely on paper under lock, in VUB buildings. No other person other than the data manager will have access.

2.2.10 Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

- No

As access to personal data will not be granted outside the scope of the project, this will not be necessary.

2.2.11 Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why.

- Yes

2.2.12 Will metadata contain information to enable the user to access the data?

- Yes

2.2.13 How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?
The public dataset will remain available as long as possible through Zenodo. In case of force majeur causes that make it unavailble there, another suitable repository will be chosen.
Restricted data will be kept for 10 years after the end of the project on paper, under lock.

2.2.14 Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?
Documentation about the dataset will be provided and stored along with the dataset and its metadata.
The source code of the developed system will be released as open source (e.g. MIT license) in appropriate repositories (e.g. GitHub).

2.3 FAIR data: Making data interoperable

2.3.1 What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines?
Will you follow community-endorsed interoperability best practices? Which ones?
The metadata will be encoded using Dublin Core.
Additionally, in compliance with VUB regulations, the metadata will also be registered in Pure following the FOSB metadata schema.

2.3.2 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?
Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
This is not expected to be necessary.

2.3.3 Will your data and other research outputs include qualified references to other data (e.g. other data from your project, or datasets from previous research)?
- No

This is not expected to happen.

2.4 FAIR data: Increase data re-use

2.4.1 How will you provide documentation needed to validate data analysis and facilitate data re-use?
The data package deposited in the trusted repository will include README files with information on methodology, as well as links to either notebooks or code repository used for data cleaning and analyses.

2.4.2 Will your data and other research outputs be made freely available in the public domain to permit the widest re-use possible?
Will your data and other research outputs be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?
The dataset of music scores will be made available to the extent that is possible depending on the included sources.
The code used, as well as additional data, will be made available with open licences (eg CC0, MIT) except for the parts that have restricted or closed access.

2.4.3 Will the data and other research output produced in the project be useable by third parties, in particular after the end of the project?

- Yes

Given the open licences used on the dataset and the code, it will be easy to reuse the produced outputs.
See 1.6 for example of interested third parties.

2.4.4 Will the provenance of the data and other research outputs be thoroughly documented using the appropriate standards?

- Yes

2.4.5 Describe all relevant data quality assurance processes.
All structured data (musicXML) as well as metadata will be validated against relevant format specifications.

3. Other research outputs

3.1 Do you have any additional information, that was not addressed in the previous sections, which you wish to provide regarding other research outputs that are generated or re-used throughout the project?

All the produced code will be made available on open code repositories (eg GitHub) at the end of the project or earlier, including the code for website and possible demonstrators.
While in progress, code will be stored in private repositories, such as the AI Lab's GitLab. (https://gitlab.ai.vub.ac.be)

4. Allocation of resources

4.1 What will the costs be for making data and other research outputs FAIR in your project?

The costs for FAIR approach is related to having online and offline storage, as well as personnel time involved for making data FAIR. Open access publications also have a cost that can be seen as related to making all research outputs FAIR.

4.2 How will these be covered?

Given the available resources available at VUB, as well as the free availability of repositories such as Zenodo and GitHub, the costs for the online storage are already covered. The personnel time is also covered via the main researcher's personnel time, which has already time allocated for management. Offline storage remains to be paid, as well as open access publications. These can be covered by Grant funding.

4.3 Who will be responsible for data management in your project?

During the research project, the main researcher (Filippo Carnovalini) is going to be the responsible for data management. After the end of the project, the supervisor (Geraint A. Wiggins) is most indicated to be the responsible given his permanent position.

4.4 How will long term preservation be ensured?

VUB RDM policy requires data to be preserved for at least 10 years after the end of the project. This will be ensured through the use of trusted repositories (e.g. Zenodo and VUB's Pixiu) and through redundancy via a secured hard drive. See Section 5 for more information. Sensitive data will be stored on paper for that period of time under lock, and then destroyed.

5. Data security

5.1 What provisions are or will be in place for data security?

Public data published on Zenodo will be considered safe and secure. However, the dataset will also be kept on a hard drive for at least ten years since the end of the project.
Sensitive data will be stored on paper under lock in VUB buildings for ten years since the end of the project, and then destroyed.

5.2 Will the data be safely stored in trusted repositories for long term preservation and curation?

- Yes

6. Ethics

6.1 Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?

- Yes

The use of music scores composed outside the scope of the project makes it so that there may be concerns regarding Copyright. We will aim to only use sources that are in the public domain. If that is impossible or leads to unsatisfactory results, we will include additional copyrighted works but respect whatever license we can obtain on those. If necessary, two versions of the dataset will be constructed: one containing only public domain data and the other also including further material.

6.2 Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

- Yes

7. Other issues

7.1 Do you, or will you, make use of other national/funder/sectorial/ departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

- No (Not expected outside of the procedures already described above.)

Signature of Fellow Signature of Supervisor