EduGames Data Management Plan

WP 1 (Management)	Deliverable D1.1
Version	1.0
Creation Date	7 / 12 / 2022
Deliverable Type	DMP
Dissemination level	Public
Author	Andrea Franceschini (andrea.franceschini.2@unipd.it)

Funded by the European Union, HORIZON-MSCA-2021-PF-01, Grant Agreement ID 101062788

Executive summary

The EduGames Data Management Plan (DMP) describes the data management lifecycle for the data to be collected, processed and/or generated and made available by this Horizon Europe project. As part of making research data findable, accessible, interoperable, and re-usable (FAIR), this DMP includes information on:

- handling the research data during and after the end of the project;
- what data will be collected, processed and/or generated;
- which methodology and standards will be applied;
- whether and how data will be shared and made open access; and
- how the data will be curated and preserved for the duration of the project and beyond.

1 Data summary

The EduGames DMP describes the strategy for managing the data generated and collected during the project, as well as any assets contributed by or acquired from third parties for the execution of the project, and to optimize re-use. Data generated during the project include:

- Data collected from participants and other stakeholders during exploration, design, experimentation.
- Data generated by the beneficiaries, the project partners, and other stakeholders during data analysis, and experimental design and preparation.

The purpose of this data collection and generation is to help achieve the main objective of the project: to define a framework of design and evaluation practices and guidelines for the development of ludic video games and interactive media for education and skill transfer. The data generated or collected is necessary for the analysis work (usability, compliance, effectiveness in delivering the project goal) and for project management (progress monitoring reports and management plans). It is not foreseen that any existing data are going to be reused as part of this project.

Table 1 lists the data types and formats that will be generated and collected during this project.

Table 1: Data types and formats collected and generated

Data type	Format	WP
Application logs	text	2
Interviews	audio, text	2
Questionnaires	text, tabular	2
Video recordings	video, text	2
Metadata	text	2
Data coding (open coding)	text	2
Project website	text, image, video	4
Social media	text, image, video	4
Reports on Dissemination, Exploitation, and Communication	text	4
Reports on project progress	text	1
Reports on training	text	3
Reports on networking	text	4

The project generates and collects data of varying origin and provenance – e.g., participants, designers, researchers, industrial and network partners, members of the public, among others. Generations and collection methods can be automatic – e.g., application logs generated by experimental software, – semi automatic – e.g., audio and video recordings, – and manual – e.g., questionnaire responses, research notes, open coding.

The data can be useful for three main groups outside the project:

- Academics engaged in research on serious games and interactive media;

- Educators who wish to incorporate ludic and interactive media into their practice;
- Industrial partners, such as producers of video games and serious games, as well as educational institutions, such as museums and galleries who wish to incorporate effective interactive experiences to augment their educational outcomes.

The size of the data is not known at this stage, but it is estimated in an order of magnitude between tens and hundreds of GB on account of rather space-hungry data formats such as video recordings.

2 FAIR data

The EduGames project aims to make as much as possible of the data generated and collected FAIR.

2.1 Making data findable, including provisions for metadata

Non-restricted data and metadata will be published on public repositories that can provide identifiers such as DOI if none are already assigned. The Dublin Core Metadata Initiative (DCMI, also DC) provides domain agnostic terms for archiving and provenance tracking, including type, format, creator, language, rights, subject, and identifier. DC metadata will be provided to make the data findable, harvestable, and indexable. Any restricted versions of the non-restricted published data sets will not be published as it may contain sensitive data such as PII. Restricted data will be instead privately stored and tagged with the corresponding public identifiers to enable tracing and reconstruction of the anonymized data if necessary. DC includes provisions to specify keywords to make the data more findable.

It is foreseen that custom metadata will be generated through data analysis (open coding) and would become part of the design and evaluation framework that is the main objective of this project as a new metadata schema. The framework will be published as a project artifact and iteratively evaluated and updated throughout the course of the project.

2.2 Making data accessible

Repository

Data generated and collected will be deposited in trusted repositories chosen based on their characteristics and the characteristics of the data. At this stage, it is foreseen that three types of repositories will be used.

- Private, encrypted, cold storage for raw and processed data that may also contain sensitive information and PII;
- domain specific, public, versioned repositories for source code, such as GitHub; and

- institutional or third-party repositories that can provide data identifiers (DOI) for any non-restricted data sets collected or generated, including any data stored in specialized repositories that do not provide an identifier – e.g., Zenodo.

Assets created by project partners and other third parties will be assessed separately and stored as appropriate, whether in one of the repositories outlined above, or other repositories. Provenance and location will be recorded as metadata and will be published in one of the repositories outlined above as appropriate.

Data

It is expected that data collected and generated will contain some form of PII or other sensitive information, therefore the data will not be published in the first instance, instead it will be kept in private, encrypted, cold storage pending further processing. The data will then be processed for publishing to remove sensitive information and add metadata. Only then will it be made publicly available through one of the repositories outlined above, depending on origin and type. Access to restricted data will only be provided to trusted parties, pending evaluation from the supervisory team based on purpose of the request and capability for secure access and handling. Trusted parties may be required to sign a Data Handling Agreement and/or a Non-Disclosure Agreement recording metadata including the contact information of the subject accessing and handling the data, the method for accessing and handling the data, and the purpose of accessing and handling the data.

Public access repository will be chosen on the basis that the data will be accessible through standard and encrypted protocols such as HTTPS and SSH.

Metadata

Metadata will be made openly available and licensed under a public domain license – e.g., CC0. DC metadata can record access information such as hyperlinks to the data and contact information for the collecting, generating, and managing subject, thus making data accessible.

It is foreseen that the data and metadata hosted in public repositories will remain available for as long as said repositories remain available and at least for the duration of the project. Restricted access data should remain accessible to trusted parties for as long as possible with periodic management and review of the private storage, so that the non-restricted version of the data can be made available in the event of public repository failure.

2.3 Making data interoperable

The DCMI provides domain agnostic terms and allows for domain specific extensions for the purpose of making data findable, harvestable, and re-usable. It may be necessary to create new vocabularies which will be documented and published for the purpose of creating and evaluating the design and evaluation framework which is the main objective of this project.

In addition, Protocol Data Element Definitions may be useful when pre-registering experiments with human participants so the integrity of the data sets can be assessed throughout the lifetime of the project.

It is not foreseen the need for using any proprietary data or metadata format in this project. In addition, formats with the widest possible adoption and ease of maintenance will be chosen – e.g., plain text, including CSV tabular data and XML for metadata; open and/or interoperable encoding formats for audio, images, and video; and other well-documented formats and standards if necessary for certain special purposes.

2.4 Increase data re-use

It is expected that only a small portion, if any at all, of the project's output in terms of data, metadata, and assets, will not be available for third parties to re-use under permissive licenses.

Assets generated by project partners, such as artworks and code, may carry some restrictions in areas such as copyright and distribution, depending on each partner's needs and requirements. If it is not possible to publish some of this with a public domain license, this decision must be documented and provided with the assets.

3 Other research outputs

The following is a list of other types of data that can be collected and/or generated by the project and project partners.

- Research software
- Research hardware
- Requirements, documentation, materials, and assembly plans
- Experimental protocols
- Artworks (audio, images, animations, videos, other assets)
- Communication (social media, project website)
- List of contacts (industrial and academic partners)

Of these, only the list of contacts is expected to be private by default, and made public only if individual contacts desire so.

4 Allocation of resources

The following is an initial estimate of the costs for making data and other research outputs FAIR in this project.

- Storage provided by online repositories free at the point of use;

- cold storage on encrypted hard drives estimated at 500 € per lifetime of storage; and
- estimated time for curation and maintenance estimated at 150 hours per annum.

Grant allowance and monthly time budget allocation are expected to cover the costs outlined above. As this is an initial estimate, it will be periodically updated and reviewed with this document.

The beneficiary researcher Dr Andrea Franceschini will be mainly responsible for data management and curation, supported by archiving and management expertise available at the Centro di Sonologia Computazionale laboratory of the beneficiary institution University of Padova. The beneficiary researcher will also be responsible for monitoring the continued availability of the online repositories, and select alternatives as a contingency.

5 Data security

Data security and data recovery is expected to be handled as follows.

- Sensitive and restricted data to be stored only on encrypted volumes, best if using hardware encryption with encrypted backup on separate storage;
- Non-sensitive and non-restricted data can be derived from the above and will be stored in online repositories. These provide a level of redundancy against data loss. Failing this, recovery will be possible from cold storage.

6 Ethics

Foreseen ethical issues with data collected or generated by this project are assessed as follows.

- Participants and third parties may withdraw their participation and contributions.
- It is necessary to keep track of where the data and assets come from, where they are stored, and who has access to them.
- Participants may withdraw up to the point of anonymization (i.e., their restricted records can be anonymized and stored as such). Participants may expressly request that all the data generated by them is destroyed beyond anonymization. Such case should be handled if enough linking information was recorded to precisely determine the relevant data, and the reason for the data loss should be recorded.
- Third parties may withdraw all or parts of their contributions to the project and contingency planning should be in place from the onset of their involvement through a negotiated agreement.

Every party involved in data creation or sharing should provide their consent to their data being collected in this project. Parties should be informed of the ways in which

their data are going to be used and should be informed of their right to withdraw their participation at any stage, up to the point of anonymization – i.e., when it is impossible to link them to their data.

7 Other issues

No other issues are foreseen at the time of writing.