

Deliverable

Deliverable name

D1.4 – Data Management Plan



Project Acronym:	IMMERSIFY
Grant Agreement number:	762079
Project Title:	Audiovisual Technologies for Next Generation Immersive Media

Revision:	0.5
Authors:	Maciej Strozyk, PSNC
Reviewers:	Mauricio Alvarez-Mesa , Spin Digital
Delivery date:	M6
Dissemination level (Public / Confidential)	Public

Abstract

The Data Management Plan of Immersify provides an overview on the general procedures and approaches for good datasets management in the project, conforming to FAIR principles which ensuring research data findable, accessible, interoperable and re-usable. This document describes the strategy followed by Immersify project regarding data management and also a list of already identified datasets.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 762079.

REVISION HISTORY

Revision	Date	Authors (Entity)	Description of changes
0.1	5 Mar 2018	Maciej Strozyk, PSNC	Initial content and structure
0.2	21 Mar 2018	Maciej Strozyk, PSNC	Updates and corrections
0.3	23 Mar 2018	Maciej Glowiak, PSNC	Review and corrections
0.4	29 Mar 2018	Mauricio Alvarez-Mesa, SD	General review
0.5	30 Mar 2018	Maciej Strozyk, PSNC	Final version

CONTRIBUTORS

Contributor's name	Entity	Contact e-mail
Maciej Stróżyk	PSNC	mackostr@man.poznan.pl
Maciej Glowiak	PSNC	mac@man.poznan.pl
Mauricio Alvarez-Mesa	SD	mauricio@spin-digital.com

TABLE OF CONTENTS

Introduction	4
Data Summary	5
FAIR data	6
Making data Findable	6
Making data openly Accessible	6
Making data Interoperable	7
Increase data Re-use	7
Allocation of resources	7
Data Security	8
Ethical aspects	8
Other issues	8

LIST OF ACRONYMS

Acronym	Description
DMP	Data Management Plan

1 Introduction

Immersify is part of the Open Research Data pilot and therefore has an obligation to prepare a Data Management Plan (DMP). Proper DMP is an important part of the project strategy to achieve specific dissemination objectives and ease the access of the consortium and the general public to the data produced during the execution of Immersify project. All partners will identify what kind of data will be generated and how they can be managed. As a general rule, all datasets will be kept and managed by the partner coordinating the action and gathering the data. Preservation will be done according to what kind of content is managed (research publications, video footage, software, etc.) and the country where the partner coordinating the action belongs.

The repository of documents (e.g. *Nextcloud*) will be used to store and share among the partners all preliminary documentation and materials required to produce the final versions of publicly available documents covering results of the research tasks. Additionally, *Git*, an open source distributed version control system, will be used for software management and *Seafile*, file sync and share platform, will serve all the video datasets (raw and encoded).

As Immersify is an innovation project, only limited research activities will be conducted, and therefore only a small set of research data will be produced. In addition, some of the research results will be classified as confidential and will not be published as open data. In particular, the results of the optimization of the video codec developed in WP3 “Immersive Display” and WP4 “Encoding and Streaming” are classified as confidential. The main reason for this decision is that those results include confidential information related to commercial products of the companies involved in the Project

Results produced as part of the work on WP5 “Creative Development and Demonstration” will be included in public deliverables, and, in particular, the work performed in Task 5.3 “Quality Assessment and Content Preparation Guidelines” will include guidelines documents that will be openly published and distributed.

The Immersify ensures also the adequate protection of privacy and non-disclosure or confidentiality agreements for sensitive information regarding individuals, project partners and external stakeholders that participate in the project. All private information will be anonymized before release to the public. Research publications on, e.g. business case studies, will in particular be required to respect this rule, anonymising individual and/or company names if requested and/or setting embargo periods (so that publications appear only when their content is not commercially sensitive).

Beyond that, and in respect of privacy and confidentiality legislation, all other data generated by project activities will be freely available.

The Immersify Data Management Plan is a living document and will be subject to change and evolve to be adapted to needs and requirements that may arise during the project life cycle. As a minimum the document will be updated in time with the RV1 - First project review (M14) and the final version will be provided by month 30 (RV2 - Final review).

In the following sections it can be found the ways of data management and procedures to access to the datasets, both for public and non-public categories.

2 Data Summary

The overall objective of Immersify is to advance the tools required for the distribution and exhibition of the next generation of immersive media. In order to satisfy this general goal Immersify will create new research datasets in scope of the project as well as it will collect existing data from partners and it will also buy third parties content. All the sets of data can be divided into four main categories:

- Software tools
- Video content (raw and encoded)
- Deliverables, reports, guidelines and dissemination materials
- Articles and scientific publications

The creative industries targeted in Immersify are those connected to the Film, TV, Video, Media Art, and related areas. As the project aims to enhance tools for delivery and exhibition of VR and immersive content, it addresses several industries across the value chain including: content producers and creators, developers and users of post-production tools, providers of VoD services, developers and users of tools for VR devices, and film/cinema and media art exhibitors.

The table below contains a description of the datasets that partners have identified during the initial 6 months of project.

Type of dataset	Format	Size	Description
Recorded raw video content	Video (Raw, TIFF, PNG)	> 30 TB	Different types of high quality immersive and interactive video content (8K, 8K 3D, 8K 360, 16K) will be recorded or captured in different way (e.g. laser scanning and CG rendering) during the project lifetime. The content will be used as a test sequences and reference materials for VR and HEVC related software tools developed in scope of the project.
Encoded video content	Video (.mp4, H.265/ HEVC)	~ 1TB	Final versions of the raw content encoded with the H.265/HEVC codec with various parameters and settings.
Making-of videos and pictures	Pictures / video	~500GB	During the recording sessions, pictures and/or video will be taken in order to document the whole process itself. This content will be mainly used for guide preparation as well as, communication, dissemination and reporting purposes.
Articles, presentations	Documents (.pdf, .pptx)	~1GB	Different types of communication materials including research articles, conference publications, press releases and presentations from the conferences.
Deliverables, reports	Documents (.pdf)	~1GB	Documents reporting project progress, achievements, and final results.
Software tools	Source code and binary files	~500MB	All software tools developed in scope of the project including encoding/decoding tools, media player, streaming server, etc.

Table 1: Definition of datasets

All repositories of the datasets used in the project are collected in Table 2.

Repository	Datasets	Public
Nextcloud	Deliverables, financial reports, sensitive project data	No
Google drive	Preliminary materials, working documents, promotional content, dissemination materials, meetings summaries	No
Web page	Project information, dissemination materials, preview of video content	Yes
SeaFile	Raw and encoded video content	No (on special request)
Git	Software source code	No

Table 2: Definition of datasets

3 FAIR data

The Immersify consortium will attempt to maximize the visibility and exploitation of the project and its long-term impact, by providing as many as possible publicly available results that can be easily discovered and re-used. Specifically, the appropriately identified datasets (videos, publications, software releases, etc.) will be generated and collected as a main project outcome. With this purpose, the project will identify which datasets can be made public, and which could be only available for project partners. This will be treated on a case-by-case basis.

3.1 Making data Findable

As not all datasets are not fully specified regarding the type and format, it is a bit too early to provide detailed information about all metadata standards, naming conventions and clear versioning procedures that will be applied in scope of the project. However, an important subset of the data is identified and will be described with basic metadata in order to make it easy findable. The biggest part of the shared datasets is video content produced by the project partners which will be described in well-defined and structured way (web page template) and additionally a proprietary metadata file linked with the video file will provide basic technical information about it (e.g. file format, resolution, audio standard, rights, etc.). The final audio-video content metadata format will be selected (EIDR membership is considered), when the complete set of video content will be produced and made available. The video content will be published in the dedicated section of the project website as well as in popular video services like YouTube or Vimeo with self-describing titles to make it findable with typical search engines (e.g. google).

3.2 Making data openly Accessible

In general, data generated by the project can be divided into two categories. First category includes **internal data** (e.g. financial reports, confidential software modules) accessible only for project partners and any sharing of this data is not planned. Second group is composed of **datasets which will be openly accessible** for research society and anyone interested in project research and the results. Basic information about the shared datasets will be listed on the public website of the project. There will be dedicated section for communication materials including research articles, press releases and presentations as well as dedicated sub-page with all information about the video content will be publicly available. The content itself will be published on the well-known video services (e.g. YouTube, Vimeo) in appropriate formats and codecs ready for distribution. This part of

video datasets will be easy accessible without any specialized software, only standard web browsers and media players will be required. However, some part of the video content will be prepared for specific visualisation installations such as caves, dome theaters or HMD devices and will not be accessible without specialized software tools developed in scope of the project (e.g. media player with HEVC and VR support) and dedicated hardware. As the software tools will be a regular commercial product, it will be delivered with full user documentation and SDK for researchers and developers. Selected public data (e.g guidelines documents, content descriptions) will be accessible via project web page and the content will be stored in data centre provided by PSNC. In case of specific cooperation with other project or researchers some raw materials or internal data can be also shared on the bi-directional agreement basis.

The consortium intends to keep the Immersify brand and the webpage even after the project end, so the created datasets which are publicly available will be preserved. Project partners have agreed to register Immersify as an international trademark.

3.3 Making data Interoperable

Data formats for most data generated by the Immersify project will be typical and universal (e.g. media files, office files) and therefore elaborate metadata is not required for all datasets, in some cases simple text description or basic metadata will be sufficient. Multimedia datasets, the most significant part of the shared content, generated in scope of the project will be provided in a typical and standardized audio/video formats such as .avi or .mp4, encoded using HEVC/H.265 and for all the movie clips the basic metadata file will be attached. In general, Immersify will make sure that suitable standards will be chosen wherever possible to ease interoperability.

3.4 Increase data Re-use

The public research data produced by the project partners will be made available under the *Creative Commons: Attribution-Share Alike* (CC BY-SA) licence but some parts of the video content will be purchased from the third parties companies (e.g. BBC, ScanLab) and in this case the licence type will be negotiated on a case-by-case basis. Further licensing details have to be specified during the coming project months.

The datasets will be available as soon as they will be fully reviewed and prepared for publication, no embargo is envisaged and Immersify will not make any restrictions regarding the duration of their re-use. In case of long-term interest data, they will be deposited in community repositories that are expected to have a long lifetime (e.g. Zenodo repository). Before publication all datasets will be validated according to the internal quality assessment procedures defined in scope of Task 5.3. An optimal configurations of the software tools will be tuned for maximizing the content quality and subjective tests will be prepared in order to validate the improvements in terms of data quality. Some datasets as well as parts of them will be used in research publications and these datasets can be potentially directly re-used to repeat the experiment or conduct further research. In that case, detailed terms and conditions for its re-use will specified.

4 Allocation of resources

The cost for making data FAIR will be, for the lifetime of the project, covered by the Immersify consortium. All data management actions and procedures are in scope of WP1 (Management and Coordination). Regarding long-term preservation, the resources will be determined during the project life-time based on the actual cost.

5 Data Security

PSNC (Poznan Supercomputing and Networking Center) provides secure and reliable storage (*Nextcloud, Seafile, Git instance*) for different type of datasets collected in the project. PSNC has a data centre with whole infrastructure necessary to provide redundancy of storage resources, power supply, network connections and cooling systems. Additionally, the data centre is equipped with active fire protection system, automatic gas extinguishing system, access control and CCTV monitoring. Also, the construction of the whole building in which the data centre is located complies with appropriate standards.

6 Ethical aspects

In the case of data where personal information is captured, the data will be anonymized before being made available publicly. No other ethical issues have been identified for the project at the moment.

7 Other issues

Probably the new EU regulation - *“REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)”* will have be taken into consideration in the next update of the DMP Deliverable because it enters into force on 25 May 2018.

No other issues have been identified so far.