# MediaVerse

A universe of media assets
and co-creation opportunities

# D1.2

## Data Management Plan

| | |
|---|---|
| **Project Title** | MediaVerse |
| **Contract No.** | 957252 |
| **Instrument** | Innovation Action |
| **Thematic Priority** | ICT-44-2020 Next Generation Media |
| **Start of Project** | 1 October 2020 |
| **Duration** | 36 months |

| | |
|---|---|
| **Deliverable title** | Data Management Plan |
| **Deliverable number** | D1.2 |
| **Deliverable version** | V1.0 |
| **Previous version(s)** | N/A |
| **Contractual Date of delivery** | 31.01.2021 |
| **Actual Date of delivery** | 29.01.2021 |
| **Nature of deliverable** | ORDP: Open Research Data Pilot |
| **Dissemination level** | Public |
| **Partner Responsible** | CERTH |
| **Author(s)** | Nikos Sarris, Symeon Papadopoulos, Manos Schinas |
| **Reviewer(s)** | Dorien Surinx (TLX), Pau Pamplona (UAB), Estella Oncins (UAB) |
| **EC Project Officer** | Alberto Rabbachin |


| | |
|---|---|
| **Abstract** | This deliverable documents the data management plan of the project in accordance with the H2020 data management requirements. |
| **Keywords** | Data Management Plan, Data Protection, Open Data, Open Research Data Pilot |


# Copyright




MediaVerse is an H2020 Innovation Project co-financed by the EC under Grant Agreement ID: 957252.
The content of this document is © the author(s). For further information, visit mediaverse-project.eu.

## Revision History

| VERSION | DATE | MODIFIED BY | COMMENTS |
|---|---|---|---|
| V0.1 | 21/12/2020 | Nikos Sarris | Table of Contents |
| V0.2 | 11/01/2021 | Symeon Papadopoulos | Revision and first inputs |
| V0.3 | 12/01/2021 | Nikos Sarris, Symeon Papadopoulos | Further revisions and additions |
| V0.4 | 14/01/2021 | Nikos Sarris | Restructuring and revisions |
| V0.5 | 18/01/2021 | Nikos Sarris | Added more inputs from partners on datasets |
| V0.6 | 25/01/2021 | Dorien Surinx, Pau Pamplona, Estella Oncins | Peer Review Comments |
| V0.7 | 26/01/2021 | Nikos Sarris, Symeon Papadopoulos | Revised draft |
| V0.8 | 27/01/2021 | Nikos Sarris | Further improvements |
| V0.9 | 28/01/2021 | Nikos Sarris | Formatting optimization |
| V1.0 | 29/01/2021 | Nikos Sarris, Symeon Papadopoulos | Final version |

## Glossary

| ABBREVIATION | MEANING |
|---|---|
| DMP | Data Management Plan |
| EC | European Commission |
| H2020 | Horizon 2020 |
| FAIR | Findable, accessible, interoperable and re-usable |
| ORDP | Open Research Data Pilot |
| WP | Work Package |
| TBD | To be determined |

# Table of Contents

# Index of Tables

# Executive Summary

The MediaVerse Data Management Plan (DMP) will guide the MediaVerse partners in the process of recording and managing the project-related data assets and activities, while identifying and addressing potential emerging security, resource and ethical issues.

The DMP is based on the H2020 Online Manual for *Data Management Plan*[1], which describes the data management life cycle for the data to be collected, processed and/or generated by the project. Following these guidelines for making research data findable, accessible, interoperable and re-usable (FAIR), this deliverable includes information on:

- The handling of research data (during and after the end of the project)
- The types and formats of data collected (processed and/or generated)
- Methodologies and standards applied
- Data accessibility and restrictions
- How data will be curated and preserved (during and after the end of the project)

As this deliverable is being composed already in M4 of the project, when many of the technical activities are at an early stage or have not even started, it is not possible to include comprehensive and definitive details concerning all of the data management activities in the project. Therefore, the main objective for this deliverable is to set the framework and an initial listing of data management activities for the project, reflecting the status on the information currently available. The document will be revisited and updated as the implementation of the project progresses and when significant changes occur; consequently, future versions of the DMP will be more detailed.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

# 1 Introduction

The traditional media landscape has rapidly evolved: newspapers, radio and TV are now part of a landscape that includes blogs, vlogs and social media platforms. With audiences seeking more user-driven and accessible multimedia experiences, the boundaries between professional media, prosumers[2] and small creators are blurring.

MediaVerse is aiming to bridge this gap, offering a decentralized network for intelligent, automated and accessible digital asset management systems, enabling secure and traceable media exchange. Through MediaVerse, traditional stakeholders and other media owners can share, enrich, verify and monetize multimedia content.

The project Data Management Plan (DMP) aims at defining the management strategy of all data within the MediaVerse framework, and describes all activities and procedures to ensure that all data is FAIR, following the template and approach recommended by the EC. While the project is in favour of making key data assets produced within the project openly available and accessible, data sharing may also be restricted in several cases, taking into account *"the need to balance openness and protection of scientific information, and privacy concerns, security as well as data management and preservation questions"*[3].

In order to define this DMP, the baseline proposed by the EC was discussed with the Data Protection Officer of CERTH and TLX, being the legal expert in the consortium. Following this discussion, a template was composed to gather information from all partners regarding the datasets they can foresee being necessary for their work in the project. The gathered information has been consolidated in a tabular structure that provides a summarized but complete view for every dataset. The template will continue to be used during the course of the project for any new datasets that are found to be necessary and with any substantial changes the current document will also be updated.

This document consists of two main parts, Section 2 and Section 3. In Section 2, we present the general data management methodology of MediaVerse -according to H2020 Guidelines and FAIR data- along with the FAIR data financing plan, data security measures and ethical aspects. In Section 3, we present and explain the MediaVerse dataset template, and then, using this template, we document the MediaVerse datasets – up to the composition date of this deliverable.

[2] Being at the same time producers and consumers of media content, as, for example, citizen journalists.
[3] Guidelines on FAIR Data Management in Horizon 2020
https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

# 2 Data Management Methodology

The methodological approach that has been used to compile this deliverable follows the "Template for Horizon 2020 Data Management Plan (DMP)", version 1.0, released on 13.10.2016 by the European Commission. The MediaVerse DMP presented in this deliverable addresses the following aspects of MediaVerse data:

- Data summary;
- FAIR data
    - Making data findable, including provisions for metadata;
    - Making data openly accessible;
    - Making data interoperable;
    - Increase data re-use.
- Allocation of resources;
- Data security;
- Ethical aspects.

In this section, we briefly present the kind of questions associated with each of these aspects. For each question we also provide a summary of the general strategy adopted by the project consortium for handling different dataset categories. Detailed answers for each dataset are provided in Section 3.

**Updating Methodology and future versions of the MediaVerse DMP**

Generally, as no modifications are expected on the DMP methodology during the project's lifetime, future versions will mainly include updates in Section 3 of this deliverable and will be maintained in the form of a live document. In this document, each dataset owner will add new or update – when necessary – the dataset entries for which he/she is responsible for, in close cooperation with the WP1 Leader. The document will be stored in the file repository (NextCloud) of the project and will follow the Data Security measures described in section 2.7.

## 2.1 Data Summary

The Data Summary addresses the following issues:

- purpose of the collected/generated data and its relation to the objectives of MediaVerse;
- types and formats of data already collected/generated and/or foreseen for collection/generation at this stage of the project;
- reusability of existing data;
- origin of the data;
- expected size of the data;
- data utility.

This field describes the data that will be generated or collected, including references to their origin (in cases where data is collected), nature, scale, to whom it could be useful, and whether it underpins a scientific publication. With regard to the individual questions, our generic DMP approach is summarized below (detailed answers for each dataset are given in Section 3).

- **Purpose of data collection/generation and relation to the objectives of the project**

The main goal of MediaVerse is to present a set of multimedia tools embodied in a decentralised network to facilitate collaborations and enhance media exchange standards overall. For this vision to unfold, the following **types of dataset** are expected to be used, collected or generated.

  - **Requirements analysis data** (in the form of questionnaires, interviews, focus groups, etc.) will be collected by use case partners from users, to identify user needs, use case scenarios and desired software functionalities. The objective of collecting such data is to orient the design and development of the MediaVerse applications, tools and assets towards the needs of actual users, and support the MediaVerse use cases altogether.
  - **Evaluation data**, such as user activity and survey data, will be collected from the end users with the aim to assess the impact and effectiveness of the proposed set of apps/tools.
  - **Technical data** (existing or generated) will be collected by technical partners in order to develop and test the MediaVerse applications and tools. A variety of data will be necessary, including audio-visual content, and social media content.
  - **Data related to MediaVerse dissemination and communication activities**, to allow better organisation of events and offer better services to attendees. Video content and photos from participants will also be used for creating dissemination content.
  - **Contact data of MediaVerse consortium members** (e.g. name, email, organisation, etc.) used for project management activities. Selected video-conference calls may be recorded, so there is also audio-visual content of the partners involved in this category.

- **Types and formats of data collected/generated**

The project will use different types of data (video, images, audio, text, system log data, etc.), both personal and non-personal, from a variety of sources (web, partners, testers, etc.) and will also generate datasets in the process of creating the MediaVerse tools and assets.

- **Reusability of existing data**

There will be reuse of existing datasets, for instance annotated image and video corpora, in tasks related to developing and benchmarking media analysis and indexing components.

- **The origin of the data**

The data comes from various origins, including the following:

  - Individual researchers that openly share their data in open repositories such as GitHub and Zenodo or via their webpages;
  - Research and academic organisations that openly share data in open or institutional repositories;

- Use case partners that share data with the technical partners of the consortium to help them train and test their algorithms and software;
- Social media platforms;
- Web pages;
- Participants, end users or project partners, after filling out a consent form;
- Questionnaires and surveys filled in by end users (evaluation questionnaires);
- Interviews and focus groups conducted with end users;
- Audio-visual content recorded for the project needs;
- Use of MediaVerse software tools by the users (user/software analytics).

- **Expected size of the data**

Dataset sizes are discussed on a dataset basis in Section 2.6.

- **Data utility**

The datasets listed in this DMP are necessary to project partners for identifying user and technical requirements for the use cases, designing, developing and testing the MediaVerse methodologies, algorithms and tools, and assessing the effectiveness of these tools in real-life trials involving end users. It is also crucial for increasing the project outreach and achieving high dissemination impact. Technical and evaluation data may also be useful to researchers with a focus on the development of similar multimedia tools.

The following subsections about making data FAIR refer to the datasets that are produced by the project and not those that already exist and are being used by the project.

## 2.2 Making Data Findable

This point addresses the following issues:

- Are the data produced in the project discoverable and identifiable?
- What naming conventions are followed?
- Will search keywords be provided that optimize possibilities for re-use?
- Are clear version numbers provided?
- What metadata will be created?

In general, the data collected and generated by the project will be identifiable and discoverable. With regard to the individual questions, our DMP approach is summarised below.

- **Are the data produced in the project discoverable and identifiable?**

Datasets that will be made publicly available will be uploaded to open repositories like Zenodo, thus making it both easily discoverable and identifiable externally. With regard to datasets that will only be used internally in the project, either because of confidentiality reasons or due to limited value to external parties, they will only be discoverable and identifiable by consortium partners or selected institutional users involved in the processing of this data. Consequently, these datasets are not subject to the FAIR data principles.

- **What naming conventions are followed?**

A specific naming convention is suggested to identify the various MediaVerse datasets:

*MV_<serial number of dataset>_<data type>_<dataset title>*

The data type field is determined according to the categorisation presented below.

*Table 1: Naming conventions for MediaVerse main data types*

| ACRONYM | DESCRIPTION |
|---|---|
| RTD | Supporting research and technical development |
| PILOT | Resulting from pilot activities |
| DISEXP | Related to dissemination activities and exploitation planning |
| MGT | Supporting project management |

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

Keywords will be provided in the cases where this is applicable.

- **Are clear version numbers provided?**

For datasets that will be made publicly available in open repositories, versioning will be supported by appropriate naming conventions.

- **What metadata will be created?**

Depending on the nature of data and its characteristics, one or more of the following types of **metadata** may be used:

  - **Descriptive**: defining key properties of data and its resources (author, title, identifier, creation or publication date, etc.).
  - **Administrative**: presenting information that facilitates data management (location of the data or resource, how data was collected/generated, etc.).
  - **Structural**: facilitating proper navigation through data and resources.

Data discoverability will be further enhanced by associating search keywords along with the data, as well as promoting the datasets through the project's communication activities (e.g. blog posts, tweets, etc.). As a part of metadata provision, keywording must comply with the following **principles**:

  - Who, what, when, where and why: these questions must be covered.
  - Consistency among the different keyword tags needs to be ensured.
  - Keywording must be relevant, understandable and clear.

## 2.3 Making Data Openly Accessible

This point addresses the following issues:

- Which data produced in the project will be made openly available as the default?
- 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)?
- Where will the data and associated metadata, documentation and code be deposited?
- 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 licence)?
- How will the identity of the person accessing the data be ascertained?

With regard to the individual questions about data accessibility, our generic DMP approach is summarized below.

- **Which data produced in the project will be made openly available as the default?**

Many of the datasets to be used in this project (as described in Section 2.6) is open data already, made openly available by third parties. Since this data is already open, as a general policy, MediaVerse will not re-share it. Sharing such data will only be pursued in cases where the data licence allows it and when MediaVerse researchers estimate that re-sharing of the data (in some new form) provides additional benefit to third parties.

In addition to open data, there are also privately owned datasets. These are owned by the organisations involved in MediaVerse, as well as some technical and academic partners, and have been collected and created over a period of years or in the context of other projects or internal processes, independently from MediaVerse. Such data may be provided to the project for research purposes, but will not be shared openly. However, effort will be made to make this (or part of this) data openly available in cooperation with the data owners, wherever this is possible.

Data that will be collected by the project in the form of questionnaires or forms addressed to project partners (for Requirements Analysis), end-users (for Evaluation) will not be made openly accessible, since they may contain personal or confidential information. Wherever possible and in case there is added value from their sharing, such data will be anonymized before being shared (mainly with regard to the evaluation data).

In any case, the aforementioned data (whether public, private, or personal) will be used exclusively for achieving the project objectives. Where appropriate, the analysis results will be made open as part of public project deliverables and publications available in open repositories.

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

Open data will be deposited in open repositories like Zenodo but also GitHub or GitLab. The datasets will also be shared through the MediaVerse website.

Datasets destined to be used internally by project partners will be stored either on the project's file repository on CERTH's servers or in the servers of project partners.

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

Different methods and software tools will be required to access the data depending on the dataset. More details are provided in Section 3.

- **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)?**

The relevant software and its documentation will be included, where applicable.

- **Where will the data and associated metadata, documentation and code be deposited?**

Open data will be deposited in open repositories, like Zenodo, GitHub or GitLab. These adopt standard and simple procedures to allow data sharing by researchers.

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

If such cases are identified, access could be granted by regulating (e.g. through a signed agreement) and restricting access to specific users.

- **Is there a need for a data access committee?**

No such need has emerged yet.

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

Such licences will be used for the data destined to be openly available (cf. section 2.5).

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

This will be dealt with on a case-by-case basis. For the open datasets, no identification of the person accessing the data will take place. For the data that will be used only internally by project partners (which is stored on the project file repository or partners' servers), access control procedures are in place that define access rights and provide secure access with username/password credentials.

## 2.4 Making Data Interoperable

This point specifies what data and metadata vocabularies, standards or methodologies are followed in order to facilitate interoperability. It also addresses whether a standard vocabulary is used for all data types within the dataset, in order to allow interoperability. The specific issues covered are the following:

- Are the data produced in the project interoperable?
- 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 dataset, 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?

- **Are the data produced in the project interoperable?**

Effort will be made to achieve interoperability on most of the data produced in MediaVerse. More information will be provided as the project unfolds.

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

In order to ensure interoperability and maximum re-use of MediaVerse data, project partners will try to collect existing and new data in standardized formats, following well-known data representation models and metadata vocabularies.

Standard and simple data vocabularies will be adopted for different types of datasets (social media data, audiovisual data, user analytics, etc.). Additionally, we will consult the OpenAIRE Guidelines for Data Archives[4]. As the project progresses and data is identified and collected, further information on making data interoperable will be outlined in subsequent versions of the DMP.

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

Whenever possible, standardised vocabularies will be used to encourage the wide exchange of information and sharing of data.

- **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?**

This will be defined on a per case basis. In general, effort will be made towards providing such mappings.

Further actions on making data interoperable will be outlined in subsequent versions of the DMP, as the project progresses. These actions refer to revisiting data and metadata vocabularies, imposing additional standards or methodologies and optimizing interoperability overall.

## 2.5 Increase Data Re-use

This point addresses the following issues:

- How will the data be licenced 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.

[4] OpenAIRE Guidelines for Data Archives: https://guidelines.openaire.eu/en/latest/

- 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?

Individual questions are addressed below:

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

This matter will be dealt with on a per case basis, depending on the features of the dataset examined. Generally, a CC-BY 4.0 (Creative Commons Attribution 4.0 International Licence) licence will be considered, which allows open sharing but also allows keeping some control over the data (e.g. requires attribution).

Most common CC –BY 4.0 licensing types are the following:

- **Creative commons Attribution-Share Alike 4.0** (CC BY-SA 4.0): any third party can freely copy, distribute, display and modify the datasets for any purpose. Remix, transform, or built upon data, must be distributed under the same licence as the original. Third parties must give appropriate credit, provide a link to the licence, and indicate if changes were made.
- **Creative Commons Attribution 4.0 International** (CC BY 4.0): any third party can freely copy, distribute, display and modify the datasets for any purpose. Third parties must give appropriate credit, provide a link to the licence, and indicate if changes were made.
- **Creative Commons Attribution-NoDerivatives 4.0 International** (CC BY-ND 4.0): any third party can freely copy, distribute, display and modify the datasets for any purpose. Remix, transform, or built upon data, however, must not be distributed. Third parties must give appropriate credit, provide a link to the licence, and indicate if changes were made.
- **Creative Commons Attribution-NonCommercial 4.0 International** (CC BY-NC 4.0): third parties can copy, distribute, display and modify the datasets for any purpose other than commercial unless they get a permission by project partners first. Third parties must give appropriate credit, provide a link to the licence, and indicate if changes were made.
- **Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International** (CC BY-NC-ND 4.0): third parties can copy, distribute, display and modify the datasets for any purpose other than commercial unless they get a permission by project partners first. Remix, transform, or built upon data, however, must not be distributed. Third parties must give appropriate credit, provide a link to the licence, and indicate if changes were made.

Licencing will be discussed in later stages of the project with all involved parties. Alternative licence schemes may also be adopted at the discretion of the dataset owner.

- **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.**

This will be examined on a per case basis. In general, effort will be made for the data to 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.**

The openly shared datasets will be reusable after the end of the project on Zenodo and any additional platforms/outlets (GitHub, Gitlab, MediaVerse website, partners' websites.

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

The openly shared datasets will in general be perpetually reusable.

- **Are data quality assurance processes described?**

Automatic *data cleaning* techniques will be employed to improve data quality. Data cleaning consists of identifying incomplete, incorrect, inaccurate, or inconsistent parts of the data and then replacing, modifying or deleting such data. For datasets including questionnaire data, a manual quality control will be performed by partners to ensure data quality.

## 2.6 Financing of FAIR Data Implementation

Since the beginning of the design of the project, data management was taken into consideration and every partner has been allocated effort for this purpose. This is embedded into the tasks dealing with data management activities, either collecting, processing, or creating datasets. Hence, all related costs for data management are already covered by the project and no additional resources will be needed. Zenodo, which is free of charge, will be used to make available papers and datasets (Green Open Access model) under the MediaVerse community[5], which has been created for this purpose, as described in D8.1 "Initial Dissemination Report".

## 2.7 Data Security

This addresses secure storage and transfer of sensitive data as well as data recovery, including the following questions:

- Is the data safely stored in certified repositories for long-term preservation and curation?
- What provisions are in place for data security?

All software tools and data storage mechanisms used within MediaVerse are designed to safeguard collected data against unauthorized use and to comply with all national and EU regulations. Engineering best practices and state-of-the-art data security measures along with GDPR legislation will all be incorporated following their respective guidelines and principles.

As explained in section 3, MediaVerse datasets will either be openly shared (by uploading them in open repositories) or shared internally among specific partners (stored in the project file repository or partners' servers). Below, we examine the data security strategy for these options.

[5] https://zenodo.org/communities/mediaverse

**Open repositories**

Datasets to be openly shared will be deposited in repositories such as Zenodo that have in place strong mechanisms and protocols for data recovery and long-term data preservation.

**MediaVerse file repository**

CERTH is hosting all data related to project management and selected research and innovation activities in their premises, utilising a file repository based on NextCloud and a Wiki page based on Wikijs, as detailed in deliverable D1.1 "Quality and Knowledge Management Plan". All partners have access to this repository and store content that needs to be shared among the consortium and it is expected that this infrastructure will be used for most project datasets.

To ensure the security and confidentiality of these datasets, the following measures have been taken:

- The servers hosting the said services are securely located and monitored in CERTH-ITI's main building. The local network is firewall protected from all external traffic and access to CERTH's servers is granted through password protected SSH with only selected trusted employees maintaining privileged accounts.
- The most secure and production-stable versions of NextCloud and Wikijs -at the given moment- were selected for these operations to be deployed. Access to both services is encrypted with HTTPS and a user login is required to access any of their content. The registration process for both services was conducted as follows:
    1. Partners' email domains were whitelisted and all the relevant parties were asked to self-register, while the process was closely monitored to ensure that there were no suspicious or unexpected signups.
    2. After a sufficiently long grace period, self-registration was disabled and the whitelisted domains were removed. All subsequent accounts are manually created by the repository administrators (CERTH).

The aforementioned structure guarantees data security as well as data recovery: stored instances can be used to restore and access data in case of systems failure or human errors.

**Partners' servers**

MediaVerse partners have significant experience in data handling and protection both in the context of their institutional operation as well as in the context of their participation in other H2020 projects. As a result, the beneficiaries already have in place operational policies regarding potential ethics issues as well as privacy and security guidelines for data protection, adhering to national and EU regulations. Ultimately, each partner is responsible for the data protection and security mechanisms in their own servers.

## 2.8 Ethical Aspects

MediaVerse research will comply with ethical principles and applicable law, guaranteeing that the rights of research participants are ensured and that research methodologies do not result in discriminatory practices or unfair treatment. Special attention will also be paid to the privacy, data protection, data

management, and health and safety of participants. Every project team needs to plan in advance every action involving users that needs to be performed in order to develop an ethically sound and integral research in all aspects of its process.

UAB is the partner appointed as ethical manager (EM) and is responsible for the management of all ethical aspects. UAB has an Ethics Committee on Animal and Human Experimentation (CEEAH)[6], which guarantees that any research complies with EU's ethical requirements.

A detailed description with guidelines for the project and complete report on requirements to obtain Ethical Certification for testing with end-users, taking special care of vulnerable groups will be provided in D1.3 "Ethical Requirements", due in month 8.

**Confidentiality and archiving**

The MediaVerse project expects the development of a set of qualitative information collecting activities. In particular, surveys, interviews and questionnaires are planned, most of them in WP2 and WP7. The importance of consent arises as both personal data and potentially confidential information might be collected. Therefore, two issues become crucial from an ethical perspective: the confidentiality of the information and the anonymisation of personal data. Specific guidelines have been agreed between the Ethical Manager (EM) at the UAB and the Data Protection Officer (DPO) at CERTH. These have been approved by the CCEAH at UAB and provided to the MediaVerse partners. General rules that partners must apply when designing and conducting their research in relation to data management are:

- Information gathered from the participants should be kept confidential.
- Information gathered should be anonymised and used only for the purpose for which it was collected.
- Participants must be given, in a clear and transparent manner, the opportunity to withdraw from the activity at any time, or to modify or delete the information provided.

In addition, MediaVerse will follow the Data Minimisation principles according to the current EU regulations[7].

The original consent forms will be sent to UAB and kept safe at UAB premises, and they will be destroyed after 5 years of the end of the project, as requested by H2020.

CERTH, as data server host and responsible for Project Management, maintains reporting documents on data handling and structure, according to national law. Consequently, each organization of the Consortium will be responsible for managing any type of dataset to adhere to their respective national legal framework.

[6] https://www.uab.cat/web/ethics-committee-on-animal-and-human-experimentation-1345735628829.html

[7] https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/how-much-data-can-be-collected_en

# 3 MediaVerse Datasets

This section includes information on all datasets that can be foreseen as necessary, at the time of writing this deliverable. Every table provides information on the dataset, along with explanation on whether and how this dataset will be FAIR and secure, as far as the datasets produced by the project are concerned. In Table 2 we present the structure of the table along with explanations for every field contained. This table template also includes the partner responsible for data collecting and maintaining the dataset, along with an indication on whether these data will be based on existing datasets.

*Table 2: Template for documenting a dataset used in MediaVerse*

| MV_<SERIAL NUMBER OF DATASET>_<DATA TYPE>_<DATASET TITLE> | |
|---|---|
| Dataset summary | Responsible partner: Partner responsible for producing and/or using the specific dataset<br><br>Purpose: Short description of data. Also, what is the purpose of data collection/generation (and its relation to project objectives) in the context of MediaVerse?<br><br>Type/format: What is the type/format of the data?<br><br>Re-use of existing data: Are existing datasets reused and how?<br><br>Data origin: What is the origin/source of the data?<br><br>Expected size: What is the expected data/dataset size?<br><br>Data utility: To whom will this data be useful and how? (inside the project and also to third parties, if applicable) |
| Findability | Is data discoverable: Are the data produced 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)?<br><br>Search keywords: Will search keywords be provided that optimize possibilities for re-use?<br><br>Versioning: Will clear version numbers be provided?<br><br>Metadata creation: Specify standards for metadata creation (if any). If there are no standards in your discipline, describe what type of metadata will be created and how. |
| Accessibility | Data openly accessible: Will data produced in the project 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.<br><br>How it will be accessible: How will the data be made accessible (e.g. by deposition in an open repository)?<br><br>Methods/software tools to access data: What methods or software tools are needed to access the data? Also, 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)?<br><br>Repository: 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.<br><br>Restrictions on access: If there are restrictions on use, how will access be provided? |
| Inter-operability | Interoperability: 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)?<br><br>Data and metadata vocabularies: Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability<br><br>Use of standard vocabularies: Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability?<br><br>Mappings to commonly used vocabularies: 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? |
| Reusability | Licence: Specify how the data will be licenced to permit the widest reuse possible. E.g. Open Data Licence (Creative Commons CC0 Licence, Creative Common Attribution Licence-CC-BY v4.0, etc.).<br><br>Availability for re-use: When will data be made available for re-use. If applicable, specify why and for what period a data embargo is needed<br><br>Usable by third parties after end of project: Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.<br><br>Re-use timeframe: Specify the length of time for which the data will remain re-usable<br><br>Data quality assurance process: Describe data quality assurance processes |
| Security | Security measures: Security measures implemented for data protection (incl. controlled access, user authentication, firewalls, VPNs, encryption, back-ups, etc.) |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: Are there any ethical or legal issues that can have an impact on data sharing?<br><br>Is informed consent for data sharing and long term preservation given: Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? |
| Other issues | Refer to other national/funder/sectorial/departmental procedures for data management that you may be using (if any) |

The datasets are categorized under four sections depending on their purpose, including a) supporting research and technical development, b) resulting from pilot activities, c) related to dissemination activities and exploitation planning and d) supporting project management. According to this classification we also codify their naming under RTD, PILOT, DISEXP and MGT.

## 3.1 Datasets for the Development of Technologies

*Table 3: MV_001_RTD_Media-understanding-test-content*

| MV_001_RTD_MEDIA-UNDERSTANDING-TEST-CONTENT | |
|---|---|
| Dataset summary | Responsible partner: CERTH<br><br>Purpose: This is a set of datasets that include multimedia content and will be used for developing, training, testing and demonstrating algorithms on new media understanding and annotation, which are being developed under T3.1. Most are existing datasets but new content may be also collected or synthetically produced.<br><br>Type/format: Multimedia files in various formats<br><br>Re-use of existing data: ModelNet40, ShapeNet, 3dscan, ScanNet, Google's Conceptual Captions, Emoji sentiment data (Kaggle), ImageNet, MSCOCO<br><br>Data origin: Publicly available on the Web or synthetically produced<br><br>Expected size: Many GB per dataset<br><br>Data utility: The data is useful to WP3 partners for training and testing algorithms on media analysis. Generated data could potentially be made publicly available. |
| Findability | Is data discoverable: The existing datasets are discoverable through the Web.<br><br>Search keywords: N/A<br><br>Versioning: N/A<br><br>Metadata creation: TBD |
| Accessibility | Data openly accessible: The existing datasets are already openly accessible.<br><br>For potentially newly collected content and the synthetically generated images it will be examined how and when to make them accessible.<br><br>How it will be accessible: Through the Web at the following addresses:<br><br>• ModelNet40: https://modelnet.cs.princeton.edu/<br>• ShapeNet: https://www.shapenet.org/<br>• 3dscan: http://redwood-data.org/3dscan/<br>• ScanNet: http://www.scan-net.org/<br>• Google's Conceptual Captions: https://ai.google.com/research/ConceptualCaptions<br>• Emoji sentiment data https://www.kaggle.com/thomasseleck/emoji-sentiment-data<br>• ImageNet: http://www.image-net.org/<br>• MSCOCO: https://cocodataset.org/#home<br><br>Methods/software tools to access data: Web browser<br><br>Repository: N/A<br><br>Restrictions on access: N/A |
| Inter-operability | Interoperability: N/A<br><br>Data and metadata vocabularies: N/A<br><br>Use of standard vocabularies: N/A<br><br>Mappings to commonly used vocabularies: N/A |

| | |
|---|---|
| Reusability | Licence: To be decided for the newly collected and synthetically generated data.<br>Availability for re-use: N/A<br>Usable by third parties after end of project: N/A<br>Re-use timeframe: N/A<br>Data quality assurance process: N/A |
| Security | Security measures: The datasets will be stored in the project file repository, which is on a dedicated web server hosted in CERTH's premises and protected as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: N/A<br>Is informed consent for data sharing and long term preservation given: N/A |
| Other issues | No |

*Table 4: MV_002_RTD_ Deepfake-detection-challenge-dataset*

| MV_002_RTD_ DEEPFAKE-DETECTION-CHALLENGE-DATASET | |
|---|---|
| Dataset summary | Responsible partner: CERTH<br>Purpose: The Deepfake Detection Challenge dataset (DFDC) consists of more than 124k videos. The DFDC has enabled experts from around the world to come together, benchmark their deepfake detection models, try new approaches, and learn from each other's work. The dataset contains real videos and videos that have been manipulated with eight facial modification algorithms<br>The dataset was created by Facebook with paid actors who entered into an agreement to the use and manipulation of their faces in the creation of the dataset.<br>Type/format: Video (mp4)<br>Re-use of existing data: Yes<br>Data origin: Kaggle/AWS<br>Expected size: ~500GB<br>Data utility: This dataset is useful for someone that wants to train deep learning models for facial manipulation detection with a large-scale dataset and will be used internally by T5.1. |
| Findability | Is data discoverable: Data is hosted on Kaggle and AWS.<br>Search keywords: N/A<br>Versioning: N/A<br>Metadata creation: TBD |
| Accessibility | Data openly accessible: Yes<br>How it will be accessible: Data is hosted on Kaggle and AWS.<br>Methods/software tools to access data: Creation of Kaggle account and AWS account with an IAM user and Access Keys.<br>Repository: Data is hosted on Kaggle and AWS.<br>Restrictions on access: The user should accept licence agreement first (cf. Reusability below). |
| Inter-operability | Interoperability: Yes |

| | |
|---|---|
| | Data and metadata vocabularies: Videos are in mp4 format and are accompanied with a metadata file that contains information about the authenticity of a particular video. Also, for manipulated videos there is information regarding the original video that was used to produce it.<br><br>Use of standard vocabularies: N/A<br><br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: To download the dataset from Kaggle the user had to agree to the Challenge rules. The guidelines and licence for the DFDC dataset are listed in section 7. COMPETITION DATA.<br><br>Availability for re-use: N/A<br><br>Usable by third parties after end of project: N/A<br><br>Re-use timeframe: N/A<br><br>Data quality assurance process: N/A |
| Security | Security measures: The datasets will be stored in CERTH's premises and protected as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: N/A<br><br>Is informed consent for data sharing and long-term preservation given: N/A |
| Other issues | No |

*Table 5: MV_003_RTD_Face-forensics++*

| MV_003_RTD_FACE-FORENSICS++ | |
|---|---|
| Dataset summary | Responsible partner: CERTH<br><br>Purpose: FaceForensics++ is a video forensics dataset consisting of 1000 original video sequences that have been manipulated with five automated face manipulation methods: Deepfakes, Face2Face, FaceSwap, FaceShifter and NeuralTextures. The data has been sourced from 977 youtube videos and all videos contain a trackable mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries. Dataset is available in 3 different video qualities.<br><br>Type/format: Videos (mp4)<br><br>Re-use of existing data: Yes<br><br>Data origin: youtube.com<br><br>Expected size: ~400GB (all video qualities)<br><br>Data utility: This dataset is useful for someone that wants to train deep learning models for facial manipulation detection with a large-scale dataset and will be used internally by T5.1. |
| Findability | Is data discoverable: Data is discoverable in the authors' GitHub repository: https://github.com/ondyari/FaceForensics<br><br>Search keywords: N/A<br><br>Versioning: N/A<br><br>Metadata creation: N/A |
| Accessibility | Data openly accessible: Yes<br><br>How it will be accessible: Through the above GitHub repository.<br><br>Methods/software tools to access data: Authors provide a download script. |

| | |
|---|---|
| | Repository: N/A<br><br>Restrictions on access: The user should accept the terms of use (cf. Reusability). |
| Inter-operability | Interoperability: The file structure makes the use of the dataset easy. Original videos are in a separate folder from manipulated. Each manipulation method also appears in a separate folder.<br><br>Data and metadata vocabularies: N/A<br><br>Use of standard vocabularies: N/A<br><br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: The data is released under the FaceForensics Terms of Use, and the code is released under the MIT licence.<br><br>Availability for re-use: N/A<br><br>Usable by third parties after end of project: N/A<br><br>Re-use timeframe: N/A<br><br>Data quality assurance process: N/A |
| Security | Security measures: The datasets will be stored in CERTH's premises and protected as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: The dataset is already accessible through the authors' GitHub repository under a given licence.<br><br>Is informed consent for data sharing and long-term preservation given: N/A |
| Other issues | No |

*Table 6: MV_004_RTD_Content-adaptation-test-media-content*

| MV_004_RTD_CONTENT-ADAPTATION-TEST-MEDIA-CONTENT | |
|---|---|
| Dataset summary | Responsible partner: ATOS<br><br>Purpose: This dataset will include multimedia content that will be gathered and used for developing, testing and demonstrating the MediaVerse content adaptation services.<br><br>Type/format: Media content formats, coding specs and media containers<br><br>Re-use of existing data: TBD<br><br>Data origin: Partner and public media content sources<br><br>Expected size: Hundreds or thousands of MB<br><br>Data utility: The data is useful to WP3 partners for training and testing algorithms on content adaptation in T3.3. Generated data could potentially be made publicly available. |
| Findability | Is data discoverable: N/A<br><br>Search keywords: N/A<br><br>Versioning: N/A<br><br>Metadata creation: TBD |
| Accessibility | Data openly accessible: TBD<br><br>How it will be accessible: TBD<br><br>Methods/software tools to access data: TBD |

| | |
|---|---|
| | Repository: TBD<br>Restrictions on access: N/A |
| Inter-operability | Interoperability: TBD<br>Data and metadata vocabularies: TBD<br>Use of standard vocabularies TBD<br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: N/A<br>Availability for re-use: N/A<br>Usable by third parties after end of project: N/A<br>Re-use timeframe: N/A<br>Data quality assurance process: N/A |
| Security | Security measures: The dataset will be stored in the ATOS premises as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: N/A<br>Is informed consent for data sharing and long term preservation given: N/A |
| Other issues | No |

*Table 7: MV_005_RTD_Content-adaptation-test-media-content-metadata*

| MV_005_RTD_CONTENT-ADAPTATION-TEST-MEDIA-CONTENT-METADATA | |
|---|---|
| Dataset summary | Responsible partner: ATOS<br>Purpose: This dataset will include multimedia content metadata that will be gathered and used for designing and demonstrating the MediaVerse content model, which are being developed under T3.4<br>Type/format: Media content formats, coding specs and media containers<br>Re-use of existing data: TBD<br>Data origin: Partners, standardisation bodies and public media sources<br>Expected size: In principle, hundreds of KB or some MB<br>Data utility: The data is useful to WP3 partners for training and testing algorithms on content adaptation. Generated data could potentially be made publicly available. |
| Findability | Is data discoverable: N/A<br>Search keywords: N/A<br>Versioning: N/A<br>Metadata creation: TBD |
| Accessibility | Data openly accessible: TBD<br>How it will be accessible: TBD<br>Methods/software tools to access data: TBD<br>Repository: TBD<br>Restrictions on access: N/A |

| | |
|---|---|
| Inter-operability | Interoperability: TBD<br>Data and metadata vocabularies: TBD<br>Use of standard vocabularies TBD<br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: N/A<br>Availability for re-use: N/A<br>Usable by third parties after end of project N/A<br>Re-use timeframe: N/A<br>Data quality assurance process: N/A |
| Security | Security measures: The dataset will be stored in the ATOS premises as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: N/A<br>Is informed consent for data sharing and long term preservation given: N/A |
| Other issues | No |

*Table 8: MV_006_RTD_ Social-media-analytics*

| MV_006_RTD_ SOCIAL-MEDIA-ANALYTICS | |
|---|---|
| Dataset summary | Responsible partner: ATC<br>Purpose: This dataset will include social media content that will be gathered and used for testing and demonstrating the MediaVerse social analytics engine.<br>Type/format: Media content formats, coding specs and media containers<br>Re-use of existing data: TBD<br>Data origin: This dataset will include social media content that will be gathered and used for testing and demonstrating the MediaVerse social analytics engine.<br>Expected size: In principle, hundreds of KB or some MB<br>Data utility: The dataset is useful for consortium partners and pertinent MediaVerse users (e.g. administrators of social media pages/accounts) and is related to T5.4. |
| Findability | Is data discoverable: N/A<br>Search keywords: N/A<br>Versioning: N/A<br>Metadata creation: TBD |
| Accessibility | Data openly accessible: TBD<br>How it will be accessible: N/A<br>Methods/software tools to access data: N/A<br>Repository: N/A<br>Restrictions on access: N/A |
| Inter-operability | Interoperability: Using the relevant social network data representation model |

| | |
|---|---|
| | Data and metadata vocabularies: Relevant social network data and metadata vocabularies<br><br>Use of standard vocabularies: No<br><br>Mappings to commonly used vocabularies: No |
| Reusability | Licence: N/A<br><br>Availability for re-use: N/A<br><br>Usable by third parties after end of project: N/A<br><br>Re-use timeframe: N/A<br><br>Data quality assurance process: N/A |
| Security | Security measures: TBD but generally datasets will be stored on the cloud and in ATC premises. For cloud storage, appropriate security measures will be in place (firewall, security plugins, etc.). |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: Sharing should comply with the Terms of Service/Use of the respective social network providers, as well as with the GDPR and any pertinent personal data protection laws.<br><br>Is informed consent for data sharing and long term preservation given: N/A |
| Other issues | No |

*Table 9: MV_007_RTD_Test-content-for-video-similarity*

| MV_007_RTD_TEST-CONTENT-FOR-VIDEO-SIMILARITY | |
|---|---|
| Dataset summary | Responsible partner: CERTH<br><br>Purpose: This is a set of datasets that include multimedia content that will be used for developing, training, testing and demonstrating algorithms on video similarity.<br><br>Type/format: The video files have been downloaded from popular video platforms or provided by the authors of the datasets, and stored in various video formats. The annotations are in JSON or plain text format.<br><br>Re-use of existing data: Yes. The following datasets are reused: FIVR-200K, VCDB, CC_WEB_VIDEO, SVD.<br><br>Data origin:<br>• FIVR-200K: http://ndd.iti.gr/fivr/<br>• VCDB: http://www.yugangjiang.info/research/VCDB/index.html<br>• CC_WEB_VIDEO: http://vireo.cs.cityu.edu.hk/webvideo/<br>• SVD: https://svdbase.github.io/<br><br>Expected size:<br>• FIVR: 225,960 videos with 16,209 annotations<br>• VCDB: 100,532 videos with 9,000 video copies<br>• CC_WEB_VIDEO: 13,129 videos with 24 query sets<br>• SVD: 562,013 videos with 34,020 annotations<br><br>Data utility: The data is useful to T4.4 partners for training and testing algorithms on video similarity, but could also be useful to other Tasks dealing with video content. |
| Findability | Is data discoverable: The existing datasets are discoverable through the Web. The datasets will be stored in the project file repository and will only be accessible by registered MediaVerse users. |

| | |
|---|---|
| | Search keywords: N/A<br>Versioning: N/A<br>Metadata creation: TBD |
| Accessibility | Data openly accessible: For FIVR-200K, the dataset's annotations are publicly available on GitHub, and the video files can be downloaded from the video links (if still online). For the remaining datasets, the original authors must be contacted to request access.<br>How it will be accessible: Through the Web.<br>Methods/software tools to access data: Web browser<br>Repository: N/A<br>Restrictions on access: N/A |
| Inter-operability | Interoperability: N/A<br>Data and metadata vocabularies: N/A<br>Use of standard vocabularies: N/A<br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: N/A<br>Availability for re-use: N/A<br>Usable by third parties after end of project: N/A<br>Re-use timeframe: N/A<br>Data quality assurance process: N/A |
| Security | Security measures: The datasets will be stored in CERTH's premises and protected as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: N/A<br>Is informed consent for data sharing and long term preservation given: N/A |
| Other issues | No |

## 3.2 Datasets for Piloting Activities

*Table 10: MV_008_PILOT_Evaluation-data*

| MV_008_PILOT_EVALUATION-DATA | |
|---|---|
| Dataset summary | Responsible partner: UAB<br>Purpose: Structured questionnaires will be developed by WP7 partners for the evaluation of the developed tools in the context of the various use cases during the pilot trials. The questionnaires will include questions that cover issues such as usefulness, usability, visualisation and interaction, learnability, encountered problems and future expectations, etc. as well as user demographics. This dataset includes questionnaires filled by the end users to assess the tools developed. Data collected through the questionnaires is used exclusively for analysis and statistical purposes.<br>Type/format: Word documents containing questions and user responses.<br>Re-use of existing data: No |

| | |
|---|---|
| | Data origin: Questionnaires filled by end-users in the context of pilot evaluation<br><br>Expected size: A few KBs per questionnaire. A few MBs in total.<br><br>Data utility: This data will be used in the context of WP7 to evaluate MediaVerse technologies. The evaluation results of the first pilot phase will be used by the technical partners to improve and extend the functionalities of the developed tools during the next development phase. The final evaluation results will help partners to improve these tools as part of further development and commercial exploitation activities. |
| Findability | Is data discoverable: No<br><br>Search keywords: N/A<br><br>Versioning: N/A<br><br>Metadata creation: TBD |
| Accessibility | Data openly accessible: Raw data is considered internal working material. Hence, interviews, questionnaires, assessments, etc. will be considered confidential and will only be accessible by UAB staff. After aggregation and processing, analysis results based on this data will be shared with the consortium and included in relevant deliverables. In case of a report or paper submitted for publication, all research findings will be integrated into the report or paper. Datasets will not be added to the publication.<br><br>How it will be accessible: N/A<br><br>Methods/software tools to access data: N/A<br><br>Repository: N/A<br><br>Restrictions on access: N/A |
| Inter-operability | Interoperability: N/A<br><br>Data and metadata vocabularies: N/A<br><br>Use of standard vocabularies: N/A<br><br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: N/A<br><br>Availability for re-use: N/A<br><br>Usable by third parties after end of project: N/A<br><br>Re-use timeframe: N/A<br><br>Data quality assurance process: N/A |
| Security | Security measures: The datasets will be stored in the premises of the respective pilot leaders protected as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: These datasets may contain personal information of end-users.<br><br>Is informed consent for data sharing and long term preservation given: An Informed Consent Form will be prepared for the participation to the use case trials. The evaluation questionnaires will include a Privacy Notice that specifies that the treatment of the data is confidential, complies with GDPR and is carried out exclusively for analysis and statistical purposes. |
| Other issues | No |

*Table 11: MV_009_PILOT_Evaluation-content*

| MV_009_PILOT_EVALUATION-CONTENT | |
|---|---|
| Dataset summary | Responsible partner: UAB<br><br>Purpose: This dataset will include content that will be created during pilot evaluation sessions. This will include multimedia and VR content along with associated metadata which will be created during test session by users of MediaVerse technologies. This content will be used for evaluation purposes and demonstration of the results of the project.<br><br>Type/format: Multimedia files with associated metadata.<br><br>Re-use of existing data: No<br><br>Data origin: Content created by end-users in the context of pilot evaluation<br><br>Expected size: TBD<br><br>Data utility: This data will be used in the context of WP7 to evaluate MediaVerse technologies, understand how useful the tools can be to users and demonstrate the capabilities of the project technologies. |
| Findability | Is data discoverable: No<br><br>Search keywords: N/A<br><br>Versioning: N/A<br><br>Metadata creation: TBD |
| Accessibility | Data openly accessible: Selected items of high quality may be promoted through the project web site and communication channels.<br><br>How it will be accessible: N/A<br><br>Methods/software tools to access data: N/A<br><br>Repository: N/A<br><br>Restrictions on access: N/A |
| Inter-operability | Interoperability: N/A<br><br>Data and metadata vocabularies: N/A<br><br>Use of standard vocabularies: N/A<br><br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: N/A<br><br>Availability for re-use: N/A<br><br>Usable by third parties after end of project: N/A<br><br>Re-use timeframe: N/A<br><br>Data quality assurance process: N/A |
| Security | Security measures: The content will be stored in the premises of the respective pilot leaders protected as described in section 2.7 |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: Data will be anonymised to allow sharing according to GDPR legislation. |

| | |
|---|---|
| | Is informed consent for data sharing and long term preservation given: An Informed Consent Form will be prepared for the participation to the use case trials. The evaluation questionnaires will include a Privacy Notice that specifies that the treatment of all personal data is confidential, complies with GDPR and is carried out exclusively for analytical and statistical purposes. Consent will be asked to use content created for demonstration purposes. |
| Other issues | No |

## 3.3 Datasets for Exploitation and Dissemination Planning

*Table 12: MV_010_DISEXP_Exploitation-plans*

| MV_010_DISEXP_EXPLOITATION-PLANS | |
|---|---|
| Dataset summary | Responsible partner: LINKS<br><br>Purpose: This dataset will include exploitation planning data from partners. Every partner will draw plans on how they expect to exploit the results of their work in the project, whether this is academic or commercial exploitation. These plans may include market and competition analysis in the industrial areas that the exploitable outcomes could potentially be positioned. Pricing and business models may also be drawn either for outcomes of individual partners or for results of group collaborations. This data will be used to produce exploitation plans for several potential exploitable outcomes of the project which will be documented in the relevant deliverables.<br><br>Type/format: Documented reports and spreadsheets<br><br>Re-use of existing data: Probably existing market data<br><br>Data origin: Content provided by partners<br><br>Expected size: Several reports in the order of 40-50 pages<br><br>Data utility: This data will be used in the context of WP8 to draw the final exploitation plans for MediaVerse technologies. |
| Findability | Is data discoverable: No<br><br>Search keywords: N/A<br><br>Versioning: N/A<br><br>Metadata creation: N/A |
| Accessibility | Data openly accessible: No.<br><br>How it will be accessible: N/A<br><br>Methods/software tools to access data: N/A<br><br>Repository: N/A<br><br>Restrictions on access: N/A |
| Inter-operability | Interoperability: N/A<br><br>Data and metadata vocabularies: N/A<br><br>Use of standard vocabularies: N/A<br><br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: N/A |

| | |
|---|---|
| | Availability for re-use: N/A<br>Usable by third parties after end of project: N/A<br>Re-use timeframe: N/A<br>Data quality assurance process: N/A |
| Security | Security measures: The content will be stored in the premises of LINKS protected as described in section 2.7 |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: N/A<br>Is informed consent for data sharing and long term preservation given: N/A |
| Other issues | No |

*Table 13: MV_011_DISEXP_Dissemination-plans*

| MV_011_DISEXP_DISSEMINATION-PLANS | |
|---|---|
| Dataset summary | Responsible partner: DW<br>Purpose: This dataset will include dissemination planning data from partners. Every partner will draw plans on how they expect to disseminate the results of their work in the project, whether this is academic or commercial. These plans may include participation in conferences or exhibitions, either with physical or remote presence. They can also include publication opportunities in relevant academic journals, or commercial forums in printed or electronic form. This data will be used to produce dissemination plans for individual outcomes of the project or collaborative results and will be documented in the relevant deliverables that will be publicly available.<br>Type/format: Documented reports and spreadsheets<br>Re-use of existing data: Probably existing data of relevant sectorial events and publications<br>Data origin: Content provided by partners<br>Expected size: Several reports in the order of 30-50 pages<br>Data utility: This data will be used in the context of WP8 to draw the dissemination strategy and plans for MediaVerse technologies. Lists of relevant events and publications can also be interesting to external parties of relevant communities. |
| Findability | Is data discoverable: Yes, as dissemination reports.<br>Search keywords: #dissemination #events, plus event-specific relevant keywords, e.g. naming the areas of scientific interest<br>Versioning: Yes<br>Metadata creation: N/A |
| Accessibility | Data openly accessible: Yes, as dissemination reports.<br>How it will be accessible: At the project website<br>Methods/software tools to access data: Web browser, PDF Viewer<br>Repository: N/A<br>Restrictions on access: No |
| Inter-operability | Interoperability: N/A<br>Data and metadata vocabularies: N/A |

| | |
|---|---|
| | Use of standard vocabularies: N/A<br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: None<br>Availability for re-use: Yes<br>Usable by third parties after end of project: Yes<br>Re-use timeframe: No limitations<br>Data quality assurance process: N/A |
| Security | Security measures: The confidential content will be stored in the premises of DW protected as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: No<br>Is informed consent for data sharing and long term preservation given: No |
| Other issues | No |

## 3.4 Datasets for Project Management

*Table 14: MV_012_MGT_Partner-admin-data*

| MV_012_MGT_PARTNER-ADMIN-DATA | |
|---|---|
| Dataset summary | Responsible partner: CERTH<br>Purpose: This dataset will include partner administrative data that will be necessary for the management of the project. This may include contact data of partners and individuals as well as administrative information for partners.<br>Type/format: Tabular data, reports and spreadsheets<br>Re-use of existing data: No<br>Data origin: Content provided by partners<br>Expected size: Kbytes<br>Data utility: This will be used in the context of WP1 to carry out project management activities. Some will be necessary to partners of the consortium, some only to the coordinator and some to the EC. |
| Findability | Is data discoverable: Only within the consortium<br>Search keywords: N/A<br>Versioning: N/A<br>Metadata creation: N/A |
| Accessibility | Data openly accessible: No<br>How it will be accessible: N/A<br>Methods/software tools to access data: N/A<br>Repository: N/A<br>Restrictions on access: N/A |
| Inter-operability | Interoperability: N/A |

| | |
|---|---|
| | Data and metadata vocabularies: N/A<br>Use of standard vocabularies: N/A<br>Mappings to commonly used vocabularies: N/A |
| Reusability | Licence: N/A<br>Availability for re-use: N/A<br>Usable by third parties after end of project: N/A<br>Re-use timeframe: N/A<br>Data quality assurance process: N/A |
| Security | Security measures: The content will be stored in the wiki, file repository and private premises of CERTH protected as described in section 2.7. |
| Ethical aspects | Possible ethical and legal aspects preventing sharing: May include personal data and can therefore not be shared outside the consortium<br>Is informed consent for data sharing and long term preservation given: N/A |
| Other issues | No |

MediaVerse is an H2020 Innovation Project co-financed by the EC under Grant Agreement ID: 957252.