D1.2

Initial Data Management Plan

Project Title	AI4Media - A European Excellence Centre for Media, Society and Democracy
Contract No.	951911
Instrument	Research and Innovation Action
Thematic Priority	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) / ICT-48-2020 - Towards a vibrant European network of AI excellence centres
Start of Project	1 September 2020
Duration	48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Deliverable title	Initial Data Management Plan
Deliverable number	D1.2
Deliverable version	1.0
Previous version(s)	-
Contractual date of delivery	28 February 2021
Actual date of delivery	01 March 2021
Deliverable filename	AI4Media_D1.2_Data_Management_Plan_final.docx
Nature of deliverable	ORDP: Open Research Data Pilot
Dissemination level	Public
Number of pages	172
Work Package	WP1
Task(s)	T1.1
Partner responsible	CERTH
Author(s)	Filareti Tsalakanidou (CERTH), Yiannis Kompatsiaris(CERTH), Vasileios Mezaris (CERTH), Symeon (Akis) Papadopoulos (CERTH)
Editor	Filareti Tsalakanidou (CERTH)
EC Project Officer	Evangelia Markidou

Abstract	This deliverable determines the strategy for data management within the AI4Media project and provides a detailed description of the datasets that will be collected, processed or generated. It describes the handling of data during and after the project lifetime and discusses how they will be curated and preserved. It also specifies which datasets will be openly accessible and how they will be shared, while also presenting the methodology and standards used to increase data interoperability.
Keywords	Artificial intelligence, data management, data collection, research data, non-research data, FAIR data, metadata, open data, discoverability, accessibility, re-usability, interoperability, data security, ethical & legal aspects, open repositories

Copyright

Contributors

NAME	ORGANISATION
Filareti Tsalakanidou	CERTH
Yiannis Kompatsiaris	CERTH
Vasileios Mezaris	CERTH
Symeon (Akis) Papadopoulos	CERTH
Kostas Votis	CERTH
Noémie Krack	KUL
Lidia Dutkiewicz	KUL
Georgios N. Yannakakis	UM
Antonios Liapis	UM
Lucile Sassatelli	3IA-UCA
Miguel Romero	3IA-UCA
Adrian Popescu	CEA
Wilfried Runde	DW
Samuel Almeida	F6S
Killian Levacher	IBM
Alberto Messina	RAI
Fulvio Negro	RAI
François Schnitzler	IDF
Ioannis Patras	QMUL
Bogdan Ionescu	UPB
Vasileios Mygdalis	AUTH
Ioannis Mademlis	AUTH
Fabrizio Sebastiani	CNR
Giuseppe Amato	CNR
Andrea Esuli	CNR
Rémi Mignot	IRCAM
Maritini Kalogerini	ATC
Stratos Tzoannos	ATC
Danae Tsabouraki	ATC
Hannes Fassold	JR
Tobias Blanke	UvA
Luca Cuccovillo	FHG-IDMT
Jakob Abesser	FHG-IDMT
Artur Garcia Saez	BSC
Jesse de Vos	NISV
Candela Bravo	LOBA
Georgi Kostadinov	IMG
Daniele Gravina	MODL

Peer Reviews

NAME	ORGANISATION
Danae Tsabouraki	ATC
Henning Müller	HES-SO

Revision History

VERSION	DATE	REVIEWER	MODIFICATIONS
0.1	14/01/2021	Filareti Tsalakanidou, Yiannis Kompatsiaris, Symeon (Akis) Papadopoulos	First draft sent to partners for contributions
0.2	03/02/2021	Filareti Tsalakanidou, Vasileios Mezaris	Updated version including inputs from KUL, CERTH, CEA, F6S, DW, RAI, ATC, JR in sections 3 & 4
0.3	15/02/2021	Filareti Tsalakanidou, Symeon (Akis) Papadopoulos	Updated version including inputs from UM, UCA, F6S, IBM, RAI, IRCAM, FhG, QMUL, IDF, UPB, AUTH, CNR, and BSC in sections 3 & 4
0.4	17/02/2021	Filareti Tsalakanidou, Vasileios Mezaris	Splitted section 4 in three sections (4,5,6). Added input from ATC in Section 4.
0.5	22/02/2021	Filareti Tsalakanidou	Small updates in sections 4,5,6. Ready for internal review.
0.6	28/02/2021	Filareti Tsalakanidou, Symeon (Akis) Papadopoulos, Vasileios Mezaris	Updated version based on internal review comments. Input from NISV, IDF, IMG and MODL for section 4 and LOBA for section 6.
1.0	01/03/2021	Filareti Tsalakanidou, Yiannis Kompatsiaris	Final version.

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Table of Acronyms and Abbreviations

Acronym	Meaning
2FA	2 Factor Authentication
AI	Artificial Intelligence
API	Application Programming Interface
ASV	Automatic Speaker Verification
AWS	Amazon Web Services
BSD-2	Berkeley Software Distribution 2
CC	Creative Commons
DB	Database
DFDC	DeepFake Detection Challenge
DMP	Data Management Plan
DoA	Description of Action
DOI	Digital Object Identifier
DPO	Data Protection Officer
DW	Deutsche Welle
EC	European Commission
EEA	European Economic Area
EEG	ElectroEncephaloGraphy
ENF	Electrical Network Frequency
EU	European Union
EULA	End User License Agreement
FAIR	Findable, Accessible, Interoperable, Reusable
FFHQ	Flickr-Faces-HQ
FoR	Fake-or-Real
FSTP	Financial Support to Third Parties
GAN	Generative Adversarial Networks
GDPR	General Data Protection Regulation
HDF5	Hierarchical Data Format 5
HTTPS	HyperText Transfer Protocol Secure
IAIDA	International AI Doctoral Academy
IAM	Identity and Access Management
ID	Identity
ILSVRC2012	Large Scale Visual Recognition Challenge 2012
ISO	International Organization for Standardization
ISO/IEC	Iternational Organization for Standardization/ International Electrotechnical Commission
IT	Information Technology
JAMS	JSON Annotated Music Specification
JSON	JavaScript Object Notation
JWT	JSON Web Token
LM	Language Model
MSD	Million Song Dataset
NEC	Non European Countries
NN	Neural Network

Acronym	Meaning
OCR	Optical Character Recognition
OTP	One Time Password
PII	Personally Identifiable Information
POPD	Protection of Personal Data
RDBMS	Relational DataBase Management System
SALAMI	Structural Analysis of Large Amounts of Music Information
SCC	Standard Contractual Clause
SME	Small-Medium Enterprise
SNR	Signal-to-Noise Ratio
SotA	State of the Art
SSL	Secure Sockets Layer
TF	Tensorflow
TSV	Tab-Separated Values
TTS	Text to Speech
UC	Use Case
UGC	User Generated Content
URL	Uniform Resource Locator
UTC	Coordinated Universal Time
UTF-8	8-bit Unicode Transformation Format
VPN	Virtual Private Network
WIPO	World Intellectual Property Organization
WNID	WordNet ID
WP	Work Package
XAI	Explainable Artificial Intelligence
XML	eXtensible Markup Language

Index of Contents

Index of Tables

Index of Figures

1. Executive Summary

Various datasets will be used, collected or generated during the lifetime of the AI4Media project in order to pursue the project's research agenda and accomplish the project's objectives, as described in the DoA. A variety of data (videos, images, audio, text, social media, user profile data, questionnaires, contact info, etc.) will be used, collected or generated aiming to:

- define the user requirements and use case scenarios;
- develop cutting-edge Artificial Intelligence (AI) technologies for specific media-related fields as well as for human-centered and society-centered AI, in the context of seven use cases;
- assess the effectiveness of the developed AI4Media technologies in a series of trials involving end-users;
- establish and support a PhD programme on AI;
- establish an AI4Media Network by attracting and involving AI researchers and SMEs through open call procedures.

More specifically, we can distinguish among the following types of research data that will be collected at different stages of the project:

- First, the seven use case partners in AI4Media will collect **use case requirements**, which might involve research with end users involving questionnaires, interviews, or similar methods to identify user needs with regard to the functionalities of the tools that will be developed for each use case. The objective of collecting such data is to define the user requirements and guide the technical development of the AI4Media tools.
- Then, in order to **develop the AI methodologies and tools** that will be used to support the seven AI4Media use cases, multiple datasets (existing or new ones) will be collected by technical partners in the context of WP3-WP6. To cover the needs of the different use cases (AI for social media, news, vision, games, social sciences and humanities, human co-creation, and automatic content moderation) heterogeneous datasets will be used or generated, including **video, images, audio, text, social media posts, user activity data,** etc. To this end, several existing datasets will be deployed, either owned by AI4Media technical and use-case partners or openly shared by third parties (usually academic or research institutions). In addition, new datasets will also be created in the context of the project, e.g., datasets including tweets for disinformation detection related to specific news events.
- Finally, as part of the **use cases evaluation**, data will be collected from the end users of the proposed AI solutions, including both user activity data generated during the use of the various AI4Media tools as well as survey data (e.g., questionnaires) aiming to assess the impact and effectiveness of the proposed AI technologies.

In addition to the research data described above, non-research datasets that will help us establish the operation of the International AI Doctoral Academy and the AI4Media Network and Open Calls will also be collected. These include:

- Personal data of lecturers, researchers and post-graduate students will be collected as part of registration processes and course attendance or instruction in the context of the **International AI Doctoral Academy** (IAIDA).
- Moreover, **applicant data** will also be collected by the platform that will be used for submitting applications to get funding from AI4Media's official mechanism for Financial Support to Third Parties (FSTP). Data for the selected **sub-granted projects** and information of **internal experts and evaluators** involved in this process will also be gathered.
- In addition, information about AI4Media's **Associate Members** will be collected to facilitate their involvement in the project. Also, contact information of people participating in AI4Media's dissemination activities.

The aforementioned (research and non-research) data requires a clear plan on how it is going to be managed, i.e., stored, accessed, shared, protected against unauthorized or improper use, etc. The main goals of AI4Media's Data Management Plan (DMP) are to:

1. Outline the datasets already used/collected/generated and/or foreseen for use/collection/generation at this stage of the project, including the context and procedures of the use/collection/generation, as well as the degree of privacy and confidentiality of the data;
2. Outline the procedures for FAIR (findable, accessible, interoperable, reusable) data;
3. Outline the measures that are foreseen for the adequate management of the data from an ethical and a security point of view.

The scope of the DMP is to describe the data management life cycle for all datasets to be used or collected in all Work Packages (WP) during the course of the 48 months of the AI4Media project. In accordance with the European Commission Directorate-General for Research & Innovation "Guidelines on FAIR Data Management in Horizon 2020" v.3.0, the AI4Media partners have already started to collect, analyse, and generate a series of datasets for the development of AI4Media's technologies in the context of the seven use cases as well as for the definition of user requirements. This deliverable has been compiled through the collaborative work of the project coordinator and the consortium partners who are involved in data collection, production, and processing. Each consortium partner received a call-to-action to identify the datasets most relevant to their respective deliverables.

This document is the initial version of the DMP. It offers information on the general data management policy that will be followed by the project using the "Template for HORIZON 2020 DATA MANAGEMENT PLAN (DMP)" (version 1.0, released on 13.10.2016 by the European Commission) and answers explicitly the questions listed there. Moreover, it lists the datasets that have already been collected or used by the partners, as well as datasets that project partners aim to use or generate during the course of the project. For each dataset, explicit information is provided with regard to data use, data collection, data discoverability, data accessibility, data interoperability and metadata, data re-usability and open data, data security, and relevant ethical & legal aspects of data processing.

This document reflects on the current state of Consortium agreements on the datasets that are collected, produced, and managed. Data management will be an active and evolving process throughout the lifetime of the project. The DMP will persist as a living document and will be updated with new datasets as they are generated or identified. The final version of the DMP will be delivered in M48.

2. Introduction

This deliverable presents the Data Management Plan (DMP) of the AI4Media project. The purpose of this deliverable is to provide a detailed description of the datasets that will be collected, processed or generated during the course of the AI4Media project. It describes the handling of research and non-research data during and after the project lifetime and discusses how they will be curated and preserved. It also provides some initial information about which datasets will be openly accessible and how they will be shared, while also presenting standards used to increase data interoperability.

The DMP provides an analysis of the main elements of the data management policy that will be adopted by the AI4Media consortium with regard to all the datasets that will be used in or generated by the project. It reflects the consortium's comprehensive approach towards data management. The DMP is a living document which will evolve during the whole lifespan of the project. The current document is the first of two versions to be delivered during the AI4Media project duration. The second and final version will be delivered in M48 to reflect changes including but not limited to the use of new data for the AI4Media pilots, changes in consortium policies (e.g. new innovation potential, decision to file for a patent, etc.), external factors (e.g. change of the license of data collected from a third party), etc.

The remainder of the deliverable is structured as follows.

Section 3 describes the methodology followed for drafting the DMP and provides a general overview of the project's data management policy.

Section 4 summarizes the management plan for the various research datasets created within AI4Media, following the aforementioned methodology.

Section 5 addresses the management of research datasets that are used within AI4Media but have been created by third parties, e.g. public or private research datasets used by WP3, 4, 5, and 6 partners to develop and test new AI methodologies, algorithms and tools.

Section 6 describes how we handle non-research data collected within AI4Media, e.g. data collected from open calls, in the context of IAIDA, etc.

Finally, Section 7 concludes the deliverable.

3. Data management methodology

The methodological approach that has been used for the compilation of the D1.2 follows the "*Template for HORIZON 2020 DATA MANAGEMENT PLAN (DMP)*", version 1.0, released on 13.10.2016 by the European Commission. Taking into account the proposed methodology, the AI4Media DMP addresses the following points on a dataset-by-dataset basis:

- 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;
- Other issues.

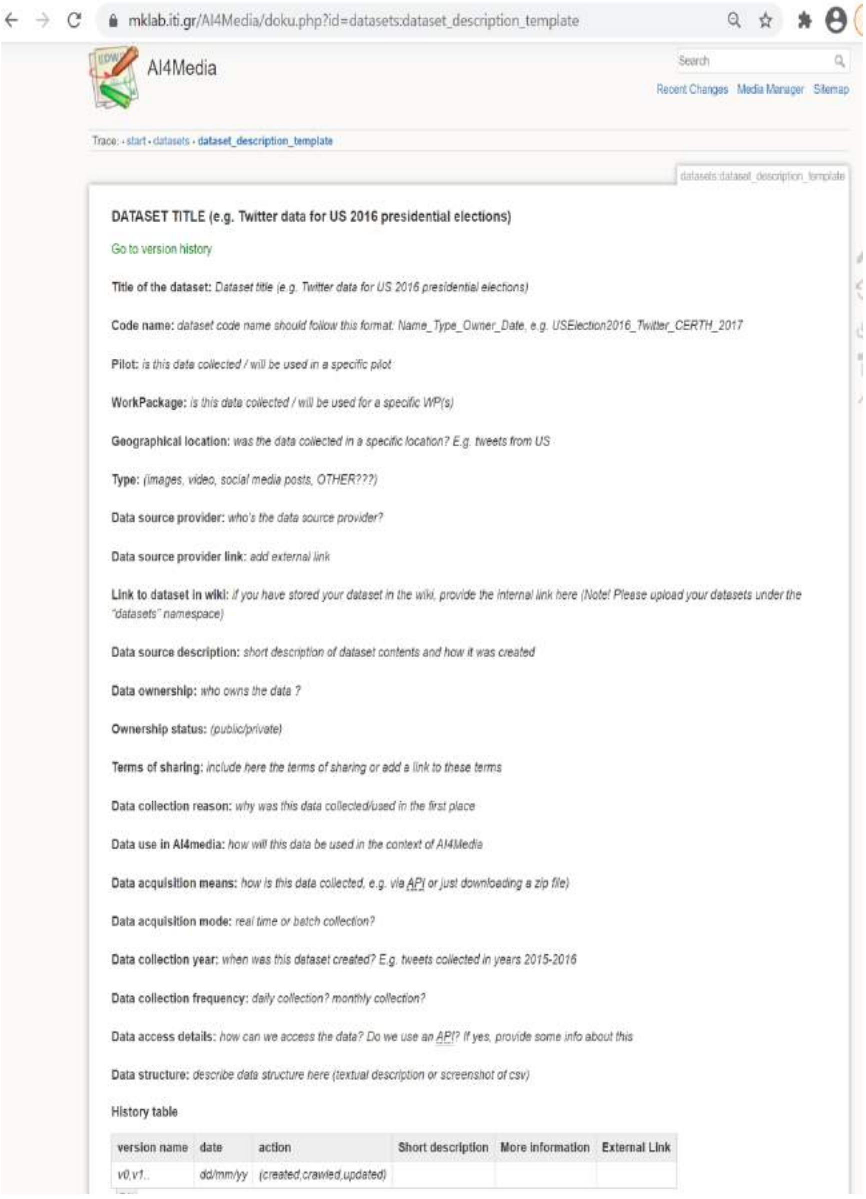
In the following subsections of section 3, we briefly present the kind of questions associated with each point in this list. Moreover, for each question we provide **a summary of the general strategy** adopted by the project consortium for handling different dataset categories. **Detailed answers to these questions on a dataset-by-dataset basis** (i.e. for each identified dataset individually) are provided in sections 4, 5 and 6.

A dataset catalogue is being maintained in the project wiki[1], which describes the datasets used in AI4Media. Each dataset has a dedicated wiki page were the dataset is described using the information defined in a relevant template[2] created by CERTH (see Figure 1). Any changes regarding the dataset are registered in a version history table. This table contains information such as when the data is added to AI4Media's catalogue, when the information is modified or if is not needed anymore in AI4Media. The aim is to always have an up-to-date version of the data catalogue. This requires a systematic process for gathering, confirming, and describing the data sources needed in each pilot.

CERTH is the partner responsible for maintaining the catalogue in the wiki and periodically checking and confirming the suitability of the data sources with the corresponding technical and use-case partners. A list including the latest confirmed entries in the data catalogue is provided to partners, so they can inform CERTH whether these datasets are still in use or not. Moreover, as mentioned above, partners fill in a template questionnaire (see Figure 1) every time a new dataset is identified. CERTH checks that the information received for the new datasets is complete and that the data sources are in fact available, while it verifies that the information for previous datasets remains still valid.

[1] https://mklab.iti.gr/AI4Media/doku.php?id=info:datasets

[2] https://mklab.iti.gr/AI4Media/doku.php?id=datasets:dataset_description_template


mklab.iti.gr/AI4Media/doku.php?id=datasets:dataset_description_template
AI4Media
Search
Recent Changes
Media Manager
Sitemap
Trace: · start · datasets · dataset_description_template
datasets:dataset_description_template

DATASET TITLE (e.g. Twitter data for US 2016 presidential elections)

Go to version history

Title of the dataset: *Dataset title (e.g. Twitter data for US 2016 presidential elections)*

Code name: *dataset code name should follow this format: Name_Type_Owner_Date, e.g. USElection2016_Twitter_CERTH_2017*

Pilot: *is this data collected / will be used in a specific pilot*

WorkPackage: *is this data collected / will be used for a specific WP(s)*

Geographical location: *was the data collected in a specific location? E.g. tweets from US*

Type: *(images, video, social media posts, OTHER???)*

Data source provider: *who's the data source provider?*

Data source provider link: *add external link*

Link to dataset in wiki: *if you have stored your dataset in the wiki, provide the internal link here (Note! Please upload your datasets under the "datasets" namespace)*

Data source description: *short description of dataset contents and how it was created*

Data ownership: *who owns the data ?*

Ownership status: *(public/private)*

Terms of sharing: *include here the terms of sharing or add a link to these terms*

Data collection reason: *why was this data collected/used in the first place*

Data use in AI4media: *how will this data be used in the context of AI4Media*

Data acquisition means: *how is this data collected, e.g. via API or just downloading a zip file)*

Data acquisition mode: *real time or batch collection?*

Data collection year: *when was this dataset created? E.g. tweets collected in years 2015-2016*

Data collection frequency: *daily collection? monthly collection?*

Data access details: *how can we access the data? Do we use an API? If yes, provide some info about this*

Data structure: *describe data structure here (textual description or screenshot of csv)*

History table

version name	date	action	Short description	More information	External Link
v0,v1..	*dd/mm/yy*	*(created,crawled,updated)*			

Figure 1: Template for dataset description as part of the data catalogue creation process in the project wiki

3.1 Data summary

The Data Summary addresses the following issues:

- Outline the purpose of the collected/ generated data and its relation to the objectives of the AI4Media project;
- Outline the types and formats of data already collected/ generated and/ or foreseen for generation at this stage of the project;
- Outline the reusability of the existing data;
- Outline the origin of the data;
- Outline the expected size of the data;
- Outline the data utility.

In this field, the data that will be generated or collected is described, 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 sections 4, 5 and 6).

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

The main goal of AI4Media is to become a centre of excellence and a wide network of researchers across Europe and beyond, with a focus on delivering the next generation of core AI advances to serve the key sector of Media, to make sure that the European values of ethical and trustworthy AI are embedded in future AI deployments, and to re-imagine AI as a crucial beneficial enabling technology in the service of Society, Democracy and Media. This goal will be achieved through six AI4Media pillars:

- The **European Media AI Observatory**, which will set and maintain a research and innovation agenda for media AI, while anticipating the social and economic disruptive potential of emerging technologies (WP2).
- An **intensive research and innovation plan** in core areas of Media AI where Europe has or can acquire a competitive advantage, generating technologies which will enrich the AI4EU platform (WP3, 4, 5, 6).
- A **portfolio of use-cases** in close coordination between academia, industry, and user groups, aimed to provide direct application of the AI4Media technologies developed within WP3-WP6 and made available through the AI4EU platform to strengthen the competitiveness of European businesses in the broader media sector and the European society (WP8).
- A **targeted programme of cascade funding** to increase engagement of actors outside the consortium and build an ecosystem around the network, in turn benefiting from it and bringing innovation to the market (WP10).
- The **International AI Doctoral Academy (IAIDA)** which will foster a new generation of talent, provide links with the industry, and ensure young skilled researchers remain in Europe (WP9).
- The **AI4Media Virtual Center of Excellence**, in close communication with the AI4EU network which will function as a portal and network nexus for all Media AI research and innovation activities in Europe (WP7, WP11).

To achieve the project’s main goal and relevant objectives the following types of datasets are expected to be used, collected, or generated:

- **User requirements data** (in the form of questionnaires, interviews, focus groups, etc.) is collected from partners involved in use cases in the context of WP8 to identify user needs, use-case scenarios, and desired software functionalities. The objective of collecting such data is to define the user-based system requirements and guide the design and development of the various AI methodologies, tools, and demonstrators that will support the AI4Media use cases. Details on these datasets can be found in section 4.4. Compliance with the legal obligations for processing personal data of users in user research activities will be ensured. For more information about this, please have a look at Ethics deliverables D12.1 and D12.2.

- **User survey data** will be collected in the context of WP2 to analyse EU AI policy, create an AI roadmap, assess the social/economic/political impact from future advances in media AI technology, and draft relevant recommendations. Details on these datasets can be found in section 4.1.

- **Media-related datasets** (existing or new ones) are used or collected by technical partners in the context of WP3-WP6 in order to develop and test the AI algorithms and software tools that will be used to support the needs of the seven AI4Media use cases, summarized below:
    - *AI for Social Media*: AI technology for detecting disinformation in social media, aiming to support journalistic fact-checking and verification workflows in news organisations;
    - *AI for News*: Smart News Assistant AI solutions to help journalists ensure that published content is both relevant for its audience and a trustworthy source of information, avoiding errors and misinformation;
    - *AI in Vision*: AI algorithms and tools for high quality video production and video content automation, including video quality enhancement, deep fake checking, data obfuscation etc.;
    - *AI for Social Sciences and Humanities*: AI technologies and tools that allow researchers and journalists to sift, connect, and analyze various data and media collections in search of factual responses to broad societal research questions;
    - *AI for Games*: AI algorithmic innovations for agent control, computer vision, and content generation to advance game design and development process, focusing on i) improved music analysis and synthesis for games, and ii) natural game interaction through edge analysis of camera stream;
    - *AI for Human Co-creation*: AI-based audio generation methods to help music composers create music;
    - *AI for (re-)Organisation and Content Moderation*: AI-based advanced content moderation solutions for media companies to allow them to effectively organize vast digital archives and collections of images and videos.

  To develop algorithms and tools for the different needs of each use case, heterogeneous datasets will be used or generated, including video, images, audio, text, social media posts, user profiles, user/user group activity data, etc. Several existing datasets will be deployed and re-used, either owned by AI4Media technical and use-case partners or openly shared by third parties (usually academic or research institutions). Details on these datasets can be found in sections 5.1-5.4. In addition, new datasets will also be created and used in the context of the project, e.g., datasets including tweets for disinformation detection on specific subjects selected in the context of the first use case. Details on these datasets can be found in sections 4.2-4.3.

- **Personal data (e.g., name, email, position, affiliation, etc.) of lecturers, researchers and post-graduate students** will be collected in the context of the IAIDA (WP9) as part of their registration process and course attendance (for students) or instruction (for lecturers). Details on these datasets can be found in section 6.3.

- **Personal data of participants** in AI4Media's dissemination (WP11) and community building events may also be collected where necessary to allow better organization of these events and better services to attendees. Data of associate members will also be collected in the context of WP11 via online questionnaire forms. Details on these datasets can be found in section 6.5.

- **Data of applicants** (individual researchers and organizations) that will use the platform set up by F6S in the context of WP10 to submit applications to get funding from AI4Media's cascade funding programme. Data for the selected sub-granted projects and information of internal experts and evaluators will also be gathered. Details on these datasets can be found in section 6.4 .

- **Evaluation data** will be collected from the end users of the proposed AI solutions in the context of WP8, including i) user activity data automatically generated during the use of the various AI4Media tools, ii) data collected during the use cases implementation to assess the evolution of the users, and iii) user research data (e.g. from questionnaires, focus groups, etc.) aiming to assess the impact and effectiveness of the proposed AI technologies and also to guide industry partners in harmonising AI research with industrial needs. Details on these datasets can be found in section in section 4.4.

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

As mentioned above, the project will use different types of data (video, images, audio, social media posts, system log data, questionnaires, applicant personal data, etc.) from various sources and will also generate datasets including data of the aforementioned types. The data will be in various formats (e.g., json and csv files for social media data; mpeg and avi files for video; excel and doc files for questionnaire data; etc.) and diverse databases (MySQL, Mongo DB etc.). Again, detailed answers for each dataset are given in sections 4, 5 and 6.

Will you re-use any existing data and how?

The datasets presented in section 5 (media-related datasets used in WP3-WP6) include existing public or private datasets available from project partners, third-party research/academic organizations, or third-party media companies. The data will be re-used in AI4Media to develop and test innovative AI algorithms, methodologies and software tools, aiming to support the needs of the seven pilots, as described above.

The data presented in section 4 is new research data generated by the project.

Beyond direct project purposes, all project data will be used for scientific publications (except from the personal data collected from participants or applicants involved in various AI4Media education, dissemination or open call activities presented in section 6; such data have no scientific value) .

What is the origin of the data?

The data originates from various sources:

- Individual researchers that openly share their data in open repositories such as GitHub and Zenodo or via their webpages.
- Research and academic organizations that openly share data in open repositories or institutional repositories.
- Use case partners (media organizations) that share existing datasets with the technical partners of the consortium to help them train and test their algorithms and software.

- Social media APIs (Twitter, Facebook, Instagram, etc.).
- Web (e.g., news articles & comments).
- Online forms filled in by applicants or participants in the context of AI4Media education, dissemination and open call activities.
- Questionnaires filled in by project partners (e.g., user requirements questionnaires), by end-users of AI4Media tools (e.g., evaluation questionnaires), etc.
- Use of AI4Media software tools by the end users during the AI4Media use case trials (this includes automatically collected user/software analytics).

What is the expected size of the data?

Dataset sizes are discussed on a dataset-by-dataset basis in sections 4, 5 and 6.

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

The aforementioned data is useful to project partners for identifying user and software requirements for the various use cases; designing, developing, and testing (and improving) the AI4Media methodologies, algorithms, and tools; and assessing the effectiveness of these tools in real-life trials involving end users. Also, it is useful for the better design, organization, and execution of activities related to IAIDA, open calls for third-party funding, outreach and dissemination.

Media-related datasets and evaluation data may also be useful to researchers with a focus on the development of AI technologies for the media in general or specific aspects of media in particular (e.g., music for games, Twitter-based disinformation analysis, or content moderation, just to name a few). The data can also be useful to social scientists that want to examine the impact of AI on media or the impact of media on the society (e.g., in the context of elections).

3.2 Making data findable, including provisions for metadata

This point addresses the following issues:

- Are the data produced and/ or used 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, most of the data produced or used by the project is or will be identifiable and discoverable. With regard to the individual questions, our generic DMP approach is summarized below (again, detailed answers for each dataset are given in in sections 4, 5 and 6):

Are the data produced and/ or used in the project discoverable and identifiable?

Publicly available third-party datasets that we will re-use to develop and test our AI technologies are already easily discoverable and identifiable from the original sources.

Datasets that will be created within AI4Media and will be made publicly available by partners will be uploaded to open repositories like Zenodo, thus making this data both easily discoverable and identifiable from the outside (since they will be assigned a DOI). These datasets will also be shared through the AI4EU platform. Some datasets wil be shared through AI4Media partners' institutional (open) repositories.

With regard to datasets that will only be used internally in the project, some of these will be stored on the project wiki in CERTH's servers. This data will only be discoverable and identifiable from registered wiki users, using simple queries with keywords.

Other internal data such as user requirement data and evaluation data will be stored on user partners' servers and access will be provided only to selected institutional users involved in the processing of this data.

What naming conventions are followed?

A specific naming convention will be used to identify the various AI4Media datasets:

AI4Media_Data_<serial number of dataset>_<WPno>_<data type>_<dataset title/ID>_v<version no>

- The <*serial number of each dataset*> is assigned manually in the order of presentation in this deliverable.
- The <*WPno*> reveals the WP in the context of which this data is collected or generated and processed.
- The <*data type*> takes one of the following values: *SOCIALMEDIA, VIDEO, AUDIO, EMAIL, TEXT, EEG, EMAIL, QUESTIONNAIRE, INTERVIEW, USER-RESEARCH, ACTIVITY-LOG, DEMOGRAPHIC*[3]*, SURVEY …*
- The <*dataset title*> is a descriptive title showing what kind of data is included in the dataset, e.g. "*Deepfake-Detection-Challenge-Dataset*", "*UseCase1-UserReq2021*".
- The <*version no*> is the dataset version. Different updated versions of the same dataset may be generated during the project lifetime.

For example, a dataset including questionnaires from the first evaluation phase of the second AI4Media use case can be named as *AI4Media_Data_10_WP8_QUESTIONNAIRE_UseCase2-EvaluationForms_v2*. For a third-party dataset which is named by its creators SumMe[4] and includes videos and relevant annotation data, we can use the following name: *AI4Media_Data_15_WP5_VIDEO_SumMeGycli14_v1*.

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

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

[3] *DEMOGRAPHIC* refers to personal data of applicants, participants, attendees, etc. collected through online forms or docs in the context of our education, dissemination and open call activities.

[4] https://gyglim.github.io/me/vsum/index.html#benchmark

Are clear version numbers provided?

For datasets that will be made publicly available by the AI4Media partners in open repositories, versioning is supported by these repositories.

Versioning is also supported by the project wiki.

What metadata will be created?

For datasets that will be shared via open repositories, the metadata standards used by these repositories will be used.

Metadata for data uploaded at the project wiki is also supported.

3.3 Making data openly accessible

This point addresses the following issues:

- Which data produced and/ or used 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? Have you explored appropriate arrangements with the identified repository?
- If there are restrictions on use, how will access be provided?
- Is there a need for a data access committee?
- Are there well-described conditions for access (i.e. a machine-readable license)?
- How will the identity of the person accessing the data be ascertained?

With regard to the individual questions about data accessibility, our generic DMP approach is summarized below (again, detailed answers for each dataset are given in sections 4, 5 and 6:

Which data produced and/ or used in the project will be made openly available as the default?

Many of the datasets used in the project is open data already made openly available by third parties (see section 5). More specifically,

- by individual researchers or research/academic organizations (e.g. benchmark video datasets, audio datasets, etc.);
- by media companies (e.g. archive of news articles from a news organization like Reuters etc.).

Since this data is already open, as a general policy, we will not re-share it. Sharing some of this data will be handled on a case-by-case basis, and will only be pursued in cases where the data license allows it and AI4Media researchers estimate that re-sharing of the data (in some new form) provides some additional benefit for the research or industrial community. In any case, we intend to provide open software tools shared on platforms like GitLab or GitHub that will allow other researchers to easily crawl and collect data from all the open data sources used in AI4Media.

With regard to datasets created within the project (see section 4), at this point, we plan to openly share the following datasets:

- a dataset including face pairs for personalities included in Wikipedia, created by partner CEA (see section 4.2.1);
- a collection of audio files with or without traces of Electrical Network Frequency, created by partner FhG (see section 4.3.4);
- a collection of audio files with traces of discontinuous Electrical Network Frequency, created by partner FhG (see section 4.3.5);
- an audio format for music production from the Demonstrator of AI co-creation in WP8, created by partner BSC (see section 4.4.10).

In addition to open data, there are also privately owned datasets. These archives are owned by the media companies and news organizations involved in AI4Media as well as by some technical and academic partners and have been usually collected and created over a period of years or in the context of other projects or internal processes. Such data (e.g. RAI's video dataset of Italian monuments – see section 5.3.3) will be provided to the project for research purposes, will not be shared openly, and will be only used internally by project partners. However, effort will be made to make some of this (or part of this) data openly available in cooperation with the data owners, wherever this is legally possible.

Data that will be collected by the project in the form of questionnaires or other survey methods addressed to project partners (e.g. to collect user requirements, see section 4.4), end users (to evaluate the AI4Media technologies, see section 4.4), and other parties (e.g. applicant data requesting AI4Media funding, see section 6.4), will not be made openly accessible since they may contain personal or sensitive information. Where possible and in case there is added value from their sharing (mainly for evaluation data), data will be anonymized before being shared.

In all cases, the aforementioned data (whether public, private, or personal) will be processed and analysed aiming to achieve the project objectives. Where appropriate, the analysis results will be made open as part of public project deliverables or publications available in open repositories.

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

Data to be openly shared will be deposited in open repositories like Zenodo but also possibly on GitHub or GitLab. The datasets will also be shared through the AI4EU platform. Some datasets will also be shared on AI4Media partners' institutional repositories. However, in this case, we will make sure that a link to these datasets is also included on the project's Zenodo repository.

Datasets that will be only used internally by project partners will be stored either on the project wiki 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 as will be explained in sections 4, 5 and 6 (e.g. web-browser, API). *In some cases, to collect the data from the original sources, dedicated crawlers will be developed by partners, which will be publicly shared in open repositories whenever possible. These can be openly used by other researchers to download the same data from the original data sources.*

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 this is applicable, the relevant software and its documentation will be included.

Where will the data and associated metadata, documentation and code be deposited? Have you explored appropriate arrangements with the identified repository?
Data to be openly shared will be deposited in open repositories like Zenodo as well as in AI4EU platform which can be accessed by registered users. These are widely used repositories adopting standard and simple procedures to allow data sharing by researchers. No need for appropriate arrangements is foreseen.

If there are restrictions on use, how will access be provided?
If such cases are identified, access could be provided either through use of consent and anonymisation or by regulating and restricting access to specific users.

Is there a need for a data access committee?
Not at this point.

Are there well-described conditions for access (i.e. a machine-readable license)?
Such licenses will be used for the data we plan to make openly available.

How will the identity of the person accessing the data be ascertained?
This will be dealt on a case-by-case basis. For the datasets we plan to share, open access will be granted. For the data that will be used only internally by project partners (which is stored on the project wiki or partners' servers), access control procedures will be in place that define access rights and provide secure access with username/password credentials.

3.4 Making data interoperable

This point describes data interoperability specifying what data and metadata vocabularies, standards or methodologies are followed in order to facilitate interoperability. Moreover, it addresses whether a standard vocabulary is used for all data types present in the dataset in order to allow inter-disciplinary interoperability. Specifically, it addresses the following issues:

- 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?

AI4Media uses data coming from diverse sources. To be able to easily integrate, analyse, and share these diverse types of data, mechanisms for data harmonization and integration will be adopted wherever possible aiming to ensure data interoperability. With regard to the individual questions about data interoperability, our generic DMP approach is summarized below (again, detailed answers for each dataset are given in sections 4 and 5):

Are the data produced in the project interoperable?
Effort will be made to make most of the data produced by the project interoperable. This will be pursued in the context of Task T7.1 *Publication of AI resources to the AI-on-demand platform*, which will develop mechanisms that will allow the smooth integration and sharing of AI4Media resources (including datasets) in the AI4EU platform. More information on interoperability will be provided in future deliverables D7.1, D7.2 and D7.4.

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 AI4Media data, project partners will try to collect existing and new data in standardized formats, following well-known data representation models and metadata vocabularies. Effort shall be made so that all datasets use the same standards for data representation (to the extent possible) and metadata creation, based on the guidelines provided by WP7. Standard data vocabularies will be adopted for different types of datasets (social media data, audiovisual data, user analytics, etc.) while a common approach will be used for metadata creation based on OpenAIRE guidelines[5]. 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. More information can be found in sections 4 and 5.

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

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

To facilitate the exchange of information and sharing of data, we will try to rely on accepted and widely used standards whenever this is possible.

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 examined on a case-by-case basis but in general effort will be made to provide such mappings.

3.5 Increase data re-use (through clarifying licenses)

This point addresses the following issues:

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

With regard to the individual questions about increasing data re-use, our generic DMP approach is summarized below (again, detailed answers per dataset are given in sections 4, 5 and 6):

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

This will be examined on a case-by-case basis depending on the dataset. Our general approach can be summarized as follows:

- In case of data coming from external open sources or in cases where the data comes with a license on its own, the data will be shared under the same licence (if we decide to reshare it).
- For other cases, a CC-BY 4.0 (Creative Commons Attribution 4.0 International License) license will be mainly selected, which allows open sharing but also allows keeping some control over the data (e.g. requires attribution).

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 case-by-case basis (see sections 4, 5 and 6). 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 useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
This will be examined on a case-by-case basis (see sections 4, 5 and 6). The datasets that we will openly share will be re-usable after the end of the project through the AI4EU platform and open repositories like Zenodo.

How long is it intended that the data remains re-usable?
This will be examined on a case-by-case basis (see sections 4, 5 and 6). The datasets that we will openly share will be re-usable at least for a few years after the end of the project.

Are data quality assurance processes described?
Automatic data cleaning techniques will be employed during data collection. Data cleaning consists of identifying incomplete, incorrect, inaccurate, or inconsistent parts of the data and then replacing, modifying, or deleting such data. This is necessary for improving data quality and producing a clean, uniform, and consistent dataset for integration; the quality of the data reflects directly upon the quality and accuracy of the data analysis results. For datasets including questionnaire data, a manual quality control will be performed by partners to ensure consistency of replies. These quality assurance processes, including data cleaning will be discussed in sections 4, 5 and 6. The data quality of openly shared datasets will be ensured through a publication process developed as part of Task 7.1 and will be described in D7.1.

3.6 Allocation of resources

This point addresses the following issues:

- Estimate the costs for making the data FAIR and describe the method of covering these costs;
- Identify responsibilities for data management in the project;
- Describe costs and potential value of long term preservation.

With regard to the individual questions, our generic DMP approach is summarized below (again detailed answers for each dataset are given in sections 4, 5 and 6):

Estimate the costs for making the data FAIR and describe the method of covering these costs.
Costs for publications are covered by the project budget. Other costs for making the data FAIR will be covered by the individual partners that will share the data. Open data sharing will be achieved by depositing the data in open repositories like Zenodo, the AI4EU platform or partners institutional open repositories where no costs for data sharing are foreseen. Resources to make the data interoperable are already foreseen in the DoA as part of the work performed in Task 7.1 "*Publication of AI resources to the AI-on-demand platform*".

Identify responsibilities for data management in the project.
A data manager role has been established in the project to ensure that data processing actions within AI4Media are in line with the law. CERTH has been appointed as the beneficiary responsible for data management and has cooperated with technical and pilot partners to draft a detailed data management plan that clearly identifies how each dataset used or created by the project will be handled. CERTH has appointed a Data Protection Officer (DPO) that will be responsible for closely monitoring the execution of the data management plan and ensuring that project partners handle project datasets appropriately. Mr Ioannis Chalinidis has been assigned to this role. Mr Chalinidis works in the central administration of CERTH and has significant experience in data management and data protection as the DPO of many H2020 projects coordinated by CERTH. The work to be done in T4.1 by KUL, the project's legal and ethical expert, on identifying the applicable legal frameworks helps with the process of data management. Moreover, ethics deliverables D12.1 and D12.2 clarify a lot of issues with regard to human participation, informed consent procedures, and protection of personal data.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?
In the following months and as more datasets are identified, the partners of the AI4Media consortium will discuss in more detail long-term preservation of project data. In principal, data used in or generated by the project will be preserved in the wiki or partners' servers for at least 3 years beyond the lifetime of the project. After this period, the data will either be deleted or preserved for a longer period of time in dedicated repositories based on the internal agreements between partners. Open data will continue to be available in open repositories after the project ends.

3.7 Data security

This point addresses data recovery, as well as secure storage and transfer of sensitive data. Specifically, this point addresses 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 developed within AI4Media will be 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 will be incorporated as well as GDPR considerations, and respective guidelines and principles.

As explained above, AI4Media datasets will either be openly shared (by uploading them in open repositories) or shared internally among specific partners (stored in the project wiki or partners' servers). In addition, some datasets will be stored in third-party cloud servers. Below, we examine the data security strategy for the aforementioned data storage options.

Open repositories

Datasets to be openly shared, will be deposited in certified repositories like Zenodo that have in place strong mechanisms and protocols for data security and long-term data preservation. The same stands for the AI4EU platform.

AI4Media wiki

The data is stored on the project wiki[6], which is hosted on a dedicated web server in CERTH's premises. The wiki web site uses for its domain an SSL certificate enabling the SHA256RSA signature algorithm and forces all visits to use HTTPS to ensure the traffic is secure. The wiki is restricted only to registered users while registration is possible only by invitation. Access to the wiki requires username/password authentication. CERTH pays special attention to security and respects the privacy and confidentiality of the users' personal data by fully complying with the applicable national (Greek), European and International framework, and the European Union's General Data Protection Regulation (GDPR) 2016/679. The AI4Media wiki uses a file-based RDBMS to enhance security as no ports for separate DB-instances are open. Web server and file-based DB are running on a Linux encrypted partition, which conforms to the data-at-rest GDPR guidelines. Moreover, at a higher level in CERTH's data center, state-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk. Regular rolling daily backups are scheduled to minimize the risk of data loss.

The data will be preserved in the wiki until the end of the project and for three years after that and will then be deleted.

Partners' servers

AI4Media 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. To that end, 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. In the following, we list some regulations and rules with regard to data protection, which can be followed by partners:

- Compliance with the internationally recognized and globally accepted and recently adopted standard, ISO/IEC 27701:2019 Security technique, which is an extension to the ISO/IEC 27001:2013 15 (Information technology - Security techniques - Information security management systems – Requirements).
- Compliance with the requirements of the ISO/IEC 27017:201516 (Information technology - Security techniques - Code of practice for information security controls based on ISO/IEC 27002 for cloud services), which provides controls and implementation guidance for cloud service providers as well as cloud service customers.
- Adherence to the internationally recognised and globally accepted standard, ISO 2701817 (Information technology - Security techniques - Code of practice for

[6] https://mklab.iti.gr/AI4Media/doku.php

protection of personally identifiable information (PII) in public clouds actions as PII processors). The standard is designed for user privacy protection. The certification combines legal requirements for data processing with technical criteria for information security systems. The goal of ISO 27018 is to provide a set of uniform security controls to public cloud computing service providers who act as personal data processors. It implements measures to protect Personally Identifiable Information (PII).

- Adherence to the General Data Protection Regulation (2016/679/EU, GDPR).
- Appropriate strong access control mechanisms (at a server level, virtual private network (VPN) level, or Virtual Machine level) to provide only the necessary level of access to specific users.
- Robust encryption of personal data at-rest and in-transfer, but also of non-personal data wherever necessary or possible.
- (Pseudo-)anonymization of personal data according to the GDPR.
- Schedule of regular (daily or weekly) backups to enable rollback in case of significant hardware storage failure and thus minimize data loss.

The data will be preserved in partner servers until the end of the project and for at least three years after that and will then be deleted.

Third-party cloud servers

In the framework of WP8, ATC will use Truly Media, a web-based platform for collaborative verification of User Generated Content (UGC), as the main demonstrator for Use Case 1. In order to facilitate the operability and functioning of Truly Media, ATC works with third-party service providers for hosting the service. The following third-party contractors are used for data storage in Truly Media:

- Amazon Web Services (Amazon S3), Inc., P.O. Box 81226, Seattle, WA 98108-1226; used for object storage.
- MongoDB, Inc., 660 York St, Ste 101, San Francisco, CA 94110, United States; used as project database.

It is noted that ATC has bound its data processors with data processing agreements concluded pursuant to Article 28 of the GDPR. In cases where third-party service providers are used, we have ensured that – even when these providers reside outside of the European Economic Area (EEA) – all data is stored in servers located in the EEA and there is no data transfer outside the EEA.

Appropriate and detailed security policies, rules, and technical measures are implemented to protect data that are used by the Truly Media platform and are stored on the platform from improper or unauthorized access, including use of firewalls where appropriate. Security measures also include 2FA (2 Factor Authentication) with OTP (One Time Password) for extra security during login, as well as Auth2.0 and JWT for authentication and authorisation. End-to-end encryption protects from man-in-the-middle attacks and data theft. All ATC employees and data processors, who have access to and are associated with the processing of personal data, are obliged to respect the confidentiality of the stored personal data. Moreover, ATC's development team has received training from external auditors for security awareness and security best practices to avoid vulnerabilities in source code. External auditors have performed black-box penetration testing to ensure that the platform is fully secure. ATC's Data Protection Officer ensures that all processes we follow are fully compliant with the GDPR provisions.

More details on the security measures adopted by each partner to protect the various datasets collected or generated by the project are provided in sections 4, 5 and 6.

3.8 Ethical & legal aspects

This point covers any ethical or legal issues that can have an impact on data sharing, including references to ethics deliverables and the ethics section (i.e. Section 5) of the DoA. Specifically, it addresses the following issues:

- Are there any ethical or legal issues that can have an impact on data sharing?
- Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

When a dataset cannot be shared, the reasons for this will be outlined (e.g. ethical restrictions, rules governing privacy and personal data protection, intellectual property, and commercial sensitivity).

With regard to the individual questions, our generic DMP approach is summarized below (again, detailed answers for each dataset are given in sections 4, 5 and 6):

Are there any ethical or legal issues that can have an impact on data sharing?
Addressing legal and ethics challenges is an important part of the AI4Media work plan. As already indicated in section 5 "*Ethics and Security*" of the DoA, special attention has been paid to these issues since the very beginning of the project. A legal partner (KUL) forms part of the consortium, providing guidance and relevant expertise. In addition, the organizational structure of AI4Media includes an external Ethics Advisory Board, which will be established after the project starts. This Board will advise on ethics issues as well as on the data processing procedures adopted in AI4Media and offer expertise on the matter. A dedicated task deals specifically with such issues: Task 1.3 "*Ethical issue management"*, led by KUL. More specifically, one of the goals of T1.3 is to deliver an Ethics Management Plan (D1.3, D1.5) that will include a set of standard procedures to assess and monitor the design, implementation and exploitation of AI-based tools and resources (incl. data collection and generation). Moreover, D1.3 and D1.5 will include the responses to the ethics requirements, as requested by the EC ethics review in relation to protection of personal data, human participants, misuse of research findings, and participation of non-EU countries. Moreover, in the context of task T2. 1 *"Analysis of the EU policy on AI and the forthcoming Commission's legislative proposal on AI regulation*" and as part of deliverable D2.1, KUL will provide an overview of existing and upcoming policy frameworks and an analysis of the ensuing requirements, offering clarifications with regard to "privacy and data governance", among others. In addition to the aforementioned tasks, WP12 "*Ethics requirements*" will also deal with specific ethics and legal requirements. These include the design of consent forms and information sheets for the collection of personal data from human participants (D12.1. - *H - Requirement No. 1*), requirements for the collection and protection of personal data (D12.2 - *POPD - Requirement No. 2*), risk assessment to prevent misuse of research findings including generated data (D12.3 - *M - Requirement No. 3*), and procedures to ensure data transfer to/from non-EU countries as required by national/EU legislation (D12.4 - *NEC - Requirement*

No. 4). All these requirements will be duly met by relevant partners and will define our decisions with regard to data sharing.

Handling of ethics, legal, and privacy issues is one of the building blocks of AI4Media. Special focus has been given to privacy rights and data protection regime under the GDPR. In the framework of AI4Media research activities, the consortium will be processing personal data or anonymized data collected during other previous professional experiences external to the research project. In other words, the consortium may happen to process data already stored in its partners' datasets or data initially collected by third parties. Article 5(1)b GDPR stipulates that personal data shall be 'collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes'. Article 5(1)b GDPR, as complemented by Recital 50 GDPR, allows further processing where the new purposes are compatible with the initial ones. Furthermore, it specifies that further processing should be considered to be compatible when is carried out for 'archiving purposes in the public interest, scientific or historical research purposes or statistical purposes' (emphasis added). When personal data are further used for compatible purposes, such as in the case of further processing of personal data for research purposes, 'no legal basis separate from that which allowed the collection of the personal data is required'.

AI4Media is a research project receiving funding from the European Union's Horizon 2020 research and innovation programme. In the light of the information provided by other partners, in the AI4Media project the partners will indeed use data that got previously collected for an initial purpose, for research purposes. Such further processing is assumed compatible with the initial purpose within the meaning of the GDPR under two conditions: (i) the data were initially collected based on a lawful basis and (ii) that appropriate technical and organisational measures are in place to safeguard the rights of the data subjects (Art. 89 GDPR).

Where relevant and in respect with our research obligations, anonymisation and pseudonymisation techniques will be used to safeguard the personal data used as part of the DMP. Where relevant, personal data will be pseudonymised according to the current state of the art, and the additional information necessary for re-identifying the individual will be kept separately (according to Art. 4(5) of the GDPR). Pseudonymisation is expressly mentioned as a measure within Art. 89 GDPR.

Collected personal data will also include processing of special categories of data (so-called 'sensitive data'), such as data revealing political opinions. More specifically, in the context of WP6 one of the research objectives is enhancing opinion mining performance in politically charged texts (e.g. tweets), especially in the presence of implicitly expressed opinions, sarcasm and metaphors. According to Article 9 GDPR, the processing of such special categories of personal data is prohibited. However, according to Article 9(2)(j) GDPR this prohibition does not apply when the processing is necessary for scientific or research purposes in accordance with Article 89(1) GDPR. The safeguards of Article 89(1) GDPR include the use of specific 'technical and organizational measures to ensure data minimisation, such as pseudonymisation, or the anonymisation of data where possible'.

The AI4Media Consortium will respect and fully comply with the GDPR provisions. Any processing of personal data in AI4Media is covered by the appropriate legal ground (e.g. informed consent or legitimate interest; for more, see D12.1 "*H - Requirement No. 1*" and

D12.2 “*POPD - Requirement No. 2*”).

Moreover, for personal data processed in the context of AI4Media use cases, relevant controller-processor agreements for data sharing will be concluded between end users (media partners) and technology partners, if applicable. Pilot participants will be provided with informed consent forms and information sheets where personal data has been collected and processed for research purposes. Considering the ethical aspect along with the legal requirements on consent as set out under the GDPR, informed consent is a key condition for autonomous decision-making, which demonstrates the notion of control over one’s personal data. By providing comprehensive information for the envisaged purposes, the aim is to sufficiently inform the person concerned about the use of his or her data in order to provide control over how the data is being managed. As indicated, the data involved has been pre-processed in a pseudonymised and confidential manner. Only anonymised (through aggregation) research results may be scientifically exchanged or disseminated.

Finally, in the context of AI4Media, we also foresee transfer of data, including personal data, between EU and non-EU countries (this issue will be addressed in D12.4 – “*NEC - Requirement No. 4”).* The AI4Media consortium includes partners that are based in Switzerland (HES-SO, IDIAP) and the United Kingdom (QMUL, F6S). Thus, the data processing activities that will involve these partners will include the import/export of personal data to these countries. As regards Switzerland, the European Commission has so far recognised this country as providing adequate protection with the Decision 2000/518/EC. According to Article 96 GDPR, international agreements involving the transfer of personal data to third countries which were concluded prior to May 2016 shall remain in force until amended, replaced or revoked. Thus, this decision is still valid. The effect of this decision is that personal data can flow from the EU to Switzerland without any further safeguard being necessary. In others words, transfers to Switzerland will be assimilated to intra-EU transmissions of data, as the European Commission’s website clearly explains.[7] As regards the United Kingdom, the UK Government has already made clear its intention to enable data to flow from the UK to EU countries without additional measures. However, transfers of personal data from the EU to the UK will be affected and standard contractual clauses (SCCs) can assist the AI4Media consortium to maintain the flow of personal data, according to Article 46 GDPR. However, since the lifetime of the project will be 48 months, the issuing of an adequacy decision to the UK by the European Commission should be anticipated within this timeframe.

Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?
Consent has been identified as a lawful basis for the processing of data gathered during AI4Media use case trials. Accordingly, the end users that will participate in the trials will be provided with informed consent forms and information sheets (see D12.1 “*H - Requirement No. 1*” and D12.2 “*POPD - Requirement No. 2*”). Consent is a key tenet of the new data protection legislation (GDPR) and can only be obtained when providing the individual control

[7] https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en

over the processing of their personal data, as only lawfully obtained consent ensures that the processing is fair and transparent to the data subject.

AI4Media will seek informed consent of all research participants. In order to obtain informed consent, the consortium will provide prospective participants sufficient opportunity to consider whether or not to participate and this under circumstances that minimise the possibility of coercion or undue influence. Pilot participants will be provided with informed consent forms and information sheets where personal data has been collected and processed for research purposes. Considering the ethical aspect along with the legal requirements on consent as set out under the GDPR, informed consent is a key condition for autonomous decision-making, which demonstrates the notion of control over one's personal data. By providing comprehensive information for the envisaged purposes, the aim is to sufficiently inform the person concerned about the use of his or her data in order to provide control over how the data is being managed. As indicated, the data involved has been pre-processed in a pseudonymised and confidential manner. Only anonymised (through aggregation) research results may be scientifically exchanged or disseminated.

The collected data will be used solely for the research purposes of AI4Media, will not be transferred to any third Parties (as specified earlier) and will be deleted three years after the end of the project.

The consortium guarantees that all personal data collected during the project will be kept secure and unreachable by unauthorized persons. The data will be handled with appropriate confidentiality and technical security, as required by law in the individual countries and EU laws and recommendations, mainly the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/67[8] of the EU. All AI4Media partners have in place their own data privacy and security policies, which are compliant with EU regulations.

Before obtaining written consent, information concerning the data processing operations will be handed to trial participants. The specific information requirements are laid down in Art. 13 and 14 GDPR. Accordingly, in order to provide information to the data subjects in a clear manner and to give the individual participants a genuine choice with regard to the envisaged data processing, the information sheets give research participants information about, inter alia:

- Purposes of data collection, data processing and data analysis;
- Types of personal data processed;
- Transfer of their personal data between the use case leader and the relevant technical partner(s), involved in the trials;
- The rights they have as data subjects, and information on how to exercise them;
- The period for which the data will be stored.

The participation at the research is entirely voluntary and the participants have the right to withdraw from the research at any time without any adverse consequences.

The Ethics Advisory Board (and where they exist, the Ethics Committees of project partners) will provide guidance with regard to the organization of trials with end users, including procedures for providing consent.

[8] https://www.eugdpr.org/eugdpr.org.html

3.9 Other issues

Other issues refer to other national/ funder/ sectoral/ departmental procedures for data management used in the project.

As mentioned above, all consortium partners (media organizations as well the research organizations and SMEs/industry that participate in the project) have in place their own data privacy and security policies, which are compliant with EU regulations and especially the GDPR.

4. Data management plan for research datasets created within AI4Media

This section presents the research datasets that partners of the AI4Media consortium will create during the project lifetime. The list is not exhaustive and represents the current status; more datasets may be created in the future, depending on the needs and opportunities that may be identified during the lifetime of the project, especially after the first evaluation phase.

In the following sub-sections, we present the DMP plan for the different research datasets that will be created within AI4Media, organized per WP. The DMP information for each dataset is provided in the form of a Table with specific fields, following the methodological approach described in section 3. The dataset presentation template is shown below. The field *DMP component* refers to the dataset reference name (i.e. dataset id).

Table 1: Template for the presentation of the data management plan for a specific dataset

DMP component	**AI4Media_Data_DatasetNo_WPX_TypeofData_DatasetTitle_Version** **Partner: Short nameof partner processing this data**
Data Summary	Purpose: Short description of data + What is the purpose of data collection/generation (and it relation to project objectives) in the context of AI4Media? A Type/format: What is the type/format of the data? Re-use of existing data: Are you re-using an existing dataset and how? Data origin: What is the origin/source of the data? Expected size: What is the expected data/dataset size? Data utility: To whom will this data be useful and how? (inside the project and also to third parties, if applicable)
Making data findable, incl. provisions for metadata	Is data discoverable: Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)? Search keywords: Will search keywords be provided that optimize possibilities for re-use? Versioning: Do you provide clear version numbers? 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.
Making data openly accessible	Data openly accessible: Will data produced and/or used 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. How it will be accessible: How will the data be made accessible (e.g. by deposition in an open repository)? 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)? 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 Restrictions on access: If there are restrictions on use, how will access be provided?
Making data interoperable	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)? Data and metadata vocabularies: Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability 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? 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?
Increase data re-use	Licence: Specify how the data will be licensed to permit the widest reuse possible. E.g. Open Data License (Creative Commons CC Zero License, Creative Common Attribution License-CC-BY v4.0, etc.). 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 Usable by third parties after end of project: Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why. Re-use timeframe: Specify the length of time for which the data will remain re-usable Data quality assurance process: Describe data quality assurance processes
Allocation of resources	Costs for making data FAIR: Estimate the costs for making your data FAIR. Describe how you intend to cover these costs Costs for long-term preservation: Describe costs and potential value of long term preservation
Data 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? 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? Discuss
Other Issues	Refer to other national/funder/sectorial/departmental procedures for data management that you may be using (if any)

Up to this point (month 6 of the project), we have identified 19 research datasets that will be created within AI4Media. Below, we provide a Table that briefly summarizes these datasets and offers a glance at the structure of this section and its subsections.

Table 2: Summary of research datasets created within AI4Media

DMP component	WP	Short summary	Relevant sub-section
Data collected in WP2 (European AI Vision, Policy and Common Research Agendas)			4.1
AI4Media_Data_01_WP2_QUESTIONNAIRE_AI-tech-roadmap-2021_v1	WP2	Questionnaires for AI technology roadmap	4.1.1
AI4Media_Data_02_WP2_SURVEY_AI-tech-impact2021_v1	WP2	Survey data for AI technology impact	4.1.2
Data collected in WP3 (New Learning Paradigms & Distributed AI)			4.2
AI4Media_Data_03_WP3_IMAGE_ FaVCI2D _v1	WP3	FaVCI2D image dataset for demographically diversified face verification	4.2.1
Data collected in WP6 (Human- and Society-centred AI)			4.3
AI4Media_Data_04_WP6_SOCIALMEDIA_GreekTwitterPolitics_v1	WP6	Greek Politics Twitter dataset	4.3.1
AI4Media_Data_05_WP6_SOCIALMEDIA_Covid19Twitter_v1	WP6	Covid-19 Twitter dataset	4.3.2
AI4Media_Data_06_WP6_SOCIALMEDIA_TextTwitter_v1	WP6	Twitter text dataset	4.3.3
AI4Media_Data_07_WP6_Audio_ENF-Presence_v1	WP6	ENF-Presence audio dataset	4.3.4
AI4Media_Data_08_WP6_Audio_ENF-Discontinuity_v1	WP6	ENF-Discontinuity audio dataset	4.3.5
Data collected in WP8 (Use cases & demonstrators in media, society and politics)			4.4
AI4Media_Data_09_WP8_USER-RESEARCH_UseCase1_DW_v1	WP8	Data from user research activities in Use Case 1	4.4.1
AI4Media_Data_10_WP8_QUESTIONNAIRE_UseCase3-UserReqCollection2021_v1	WP8	Questionnaires for the collection of user requirements for Use Case 3	4.4.2
AI4Media_Data_11_WP8_QUESTIONNAIRE_UseCase3Evaluation_v1	WP8	Questionnaires for the evaluation of Use Case 3	4.4.3
AI4Media_Data_12_WP8_USER-RESEARCH_UseCase4_NISV_v1	WP8	Data from user research activities in Use Case 4	4.4.4
AI4Media_Data_13_WP8_QUESTIONNAIRE_VIDEO_UseCase7-UserFeedbackData_v1	WP8	Data from user research activities in Use Case 7	4.4.5
AI4Media_Data_14_WP8_Text_ThruthNestDataset-UseCase1-ATC_v1	WP8	TruthNest Twitter dataset	4.4.6
AI4Media_Data_15_WP8_Text_TrulyMediaDataset-UseCase1-ATC_v1	WP8	TrulyMedia social media +web dataset	4.4.7
AI4Media_Data_16_WP8_Text_ UI_pose-UC5-IDF_v1	WP8	UI pose dataset for Use Case 5	4.4.8
AI4Media_Data_17_WP8_Video_GameGlitches-UC5-MODL_v1	WP8	Game glitches dataset for Use Case 5	4.4.9
AI4Media_Data_18_WP8_Audio_RAW compositions_v1	WP8	Musical production for AI co-creation dataset	4.4.10
AI4Media_Data_19_WP8_ QUESTIONNAIRE_AI-IndustrialNeeds-T8.4_v1	WP8	User survey data for AI industrial needs (for T8.4)	4.4.11

4.1 Datasets collected in the context of WP2

4.1.1 Questionnaires for AI technology roadmap

DMP component	AI4Media_Data_01_WP2_QUESTIONNAIRE_AI-tech-roadmap-2021_v1 Partner: CERTH
Data Summary	Purpose: A structured questionnaire will be developed by CERTH to collect information on the evolving landscape of AI technologies for the media sector. The questionnaire will include questions on existing key AI technologies and tools for media, application areas, relevant research papers, datasets, future trends, ground-breaking new areas of implementation, etc. The online questionnaire will be filled by project partners and associate members of the consortium. The collected questionnaires will then be analyzed and the results of this analysis will be used for the development of a roadmap for AI technologies and applications in the media sector (D2.3) in the context of T2.3. Type/format: Word documents containing questions and user responses. Re-use of existing data: No. Data origin: Questionnaires filled by project partners and associate members of the consortium. Expected size: A few KBs per questionnaire. ~2 MB in total. Data utility: It is useful to WP2 partners for the development of the roadmap for AI in the media sector. It could also be useful to other researchers that work on the same field (AI surveys) but also to policy makers dealing with AI-related issues.
Making data findable, incl. provisions for metadata	Is data discoverable: The questionnaires will be stored in the project wiki. They will be discoverable by registered wiki users using appropriate keywords. No DOI will be assigned. Search keywords: Appropriate keywords to help wiki users identify the dataset (e.g. wp2, questionnaire, roadmap, etc.). Versioning: The wiki supports versioning. Metadata creation: Date, Filename, File size, References, History
Making data openly accessible	Data openly accessible: We do not intend to make this data openly available since they are used as part of an internal exercise that will facilitate the authoring of D2.3. However, a summary of this data has will be presented in D2.3, which will available on the project website and Zenodo. 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. How it will be accessible: Stored in the wiki, it is only internally accessible by project partners. Methods/software tools to access data: Web-browser Repository: Project wiki. Restrictions on access: Shared among project partners with a wiki account. Access Control List: All partners (R), CERTH (W)
Making data interoperable	Interoperability: N/A

	Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will not be shared. Availability for re-use: This data is not expected to be re-used. It has been used by WP2 partners to develop the AI roadmap. It will remain on the project wiki for three years after the end of the project. Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data will be stored on the project wiki, which is on a dedicated web server hosted in CERTH's premises. The wiki web site uses for its domain an SSL certificate enabling the SHA256RSA signature algorithm and forces all visits to use HTTPS to ensure the traffic is secure. The wiki is restricted only to registered users while registration is possible only by invitation. Access requires username/password authentication. CERTH fully complies with the applicable national, European and International framework, and the GDPR. The wiki uses a file-based RDBMS to enhance security. Web server and file-based DB are running on a Linux encrypted partition, which conforms to the data-at-rest GDPR guidelines. Moreover, state-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk. Regular rolling daily backups are scheduled to minimize the risk of data loss. The data will be preserved there for three years after the end of the project and will then be deleted.
Ethical aspects	Possible ethical and legal aspects preventing sharing: No. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

4.1.2 Survey data for AI technology impact

DMP component	**AI4Media_Data_02_WP2_SURVEY_AI-tech-impact2021_v1** **Partner: UvA**
Data Summary	Purpose: A structured questionnaire might be developed by UvA to collect information on the social and economic impact of AI for media technologies in the context of T2.4. The structured questionnaire will be followed by dedicated interviews. The collected questionnaires and the interviews will be analyzed and the results of this analysis will be used for drafting a white paper on the social, economic, and political impact of media AI Technologies (D2.2). Type/format: Online survey and interviews. Re-use of existing data: No.

	Data origin: Surveys by project partners and associate members of the consortium. Expected size: <2MB Data utility: Only internal for WP2 research. The collected data will be used for the authoring of D2.2.
Making data findable, incl. provisions for metadata	Is data discoverable: No Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: We do not intend to make this data openly available. A summary of this data has will be presented in D2.2, which will be available on the project website and Zenodo. No personal information will be published. How it will be accessible: Stored in the project wiki, will only be internally accessible by project partners. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: Shared among project partners with a wiki account. Access Control List: All partners (R), CERTH (W)
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data will be stored on the project wiki, which is on a dedicated web server hosted in CERTH's premises. The wiki web site uses for its domain an SSL certificate enabling the SHA256RSA signature algorithm and forces all visits to use HTTPS to ensure the traffic is secure. The wiki is restricted only to registered users while registration is possible only by invitation. Access requires username/password authentication. CERTH fully complies with the applicable national, European and International framework, and the GDPR. The wiki uses a file-based RDBMS to enhance security. Web server and file-based DB are running on a Linux encrypted partition, which conforms to the data-at-rest GDPR guidelines. Moreover, state-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file

	system (to prevent illegitimate access from inside) are the countermeasures against this risk. Regular rolling daily backups are scheduled to minimize the risk of data loss. The data will be preserved there for three years after the end of the project and will then be deleted.
Ethical aspects	Possible ethical and legal aspects preventing sharing: No. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

4.2 Datasets collected in the context of WP3

4.2.1 FaVCI2D image dataset for demographically diversified face verification

DMP component	AI4Media_Data_03_WP3_IMAGE_ FaVCI2D _v1 Partner: CEA
Data Summary	Purpose: This dataset includes face pairs for personalities included in Wikipedia. Focus is put on having a demographically diversified sample of persons (age, origin, gender, professions) and on the inclusion of difficult pairs of images. Its creation is necessary because most existing datasets are biased on at least one important demographic dimension. As a consequence, their use does not allow a fair evaluation of face verification algorithms. FAVCI2D will be used in WP3 (T3.3) to evaluate transferability of face analysis models and in WP4 (T4.6) as a contribution to improved benchmarking of AI systems. Type/format: JPEG images and accompanying metadata in json format. Re-use of existing data: No, the dataset is created within AI4Media. Data origin: https://commons.wikimedia.org and Bing Image Search Expected size: ~1 GB Data utility: It is useful to WP3 partners to benchmark face verification algorithms in a setting which is close to realistic conditions.
Making data findable, incl. provisions for metadata	Is data discoverable: Data will be made available as a subproject of the CEA's GitHub account: https://github.com/cea-list-lasti Search keywords: N/A Versioning: GitHub supports versioning Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data will be openly accessible via GitHub at https://github.com/cea-list-lasti How it will be accessible: The data can be downloaded from an online archive after completing a form. Methods/software tools to access data: N/A Repository: GitHub Restrictions on access: The user should accept the terms of use.
Making data	Interoperability: The file structure makes the use of the dataset easy.

interoperable	Data and metadata vocabularies: The dataset is structured around identities, with a JSON file including necessary attributes such as: wiki ID, name, age, origin, gender, professions and two pairs of images associated to it. The first pair includes two images of the person, while the second includes an image of the target person and an imposter image. The imposter image belongs to another identity which is visually similar to the target person. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is released under the FaVCI2D Terms of Use, and the code is released under the CC license. Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The full dataset (including images and non-anonymized metadata) will be hosted on CEA's servers. CEA fully complies with the applicable national, European and International framework, and the GDPR. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The version of the dataset which is shared publicly includes data minimization, in compliance with art. 9 of GDPR. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.3 Datasets collected in the context of WP6

4.3.1 Greek Politics Twitter dataset

DMP component	**AI4Media_Data_04_WP6_SOCIALMEDIA_GreekTwitterPolitics_v1** **Partner: AUTH**
Data Summary	Purpose: The dataset will include Greek-language tweets with political content, published during the decade 2010-2020 from users in Greece. The tweets will be collected using the Twitter Stream API and a set of appropriate keywords and hashtags and/or from public datasets. This data is to be used in Task 6.4 "*AI for Healthier Political Debate*" to investigate political public opinion mining and/or monitoring. The dataset is to be used by AUTH for training and testing the new algorithms developed within T6.4. Type/format: txt & csv files Re-use of existing data: We will use data mined from Twitter and, potentially, data

	from public datasets. Data origin: twitter.com https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5663401/?fbclid=IwAR3GJ4RtUHWEff31i2UrydxryLnZQbQtGhp8bJV9YtU7Z74lGPLNi95T5zM https://link.springer.com/article/10.1007/s10579-018-9420-4 https://www.ihu.edu.gr/tjortjis/A%20Hybrid%20Method%20for%20Sentiment%20Analysis%20of%20Election%20Related%20Tweets.pdf Expected size: ~ 10 GB Data utility: The data will be useful to WP6 partners that process Greek-language tweets. The data will also be useful to other social media researchers, but also to social or political scientists that study public opinion in Greece.
Making data findable, incl. provisions for metadata	Is data discoverable: The data comprising the dataset (i.e. the tweets) will be discoverable by searching the Twitter feed with selected keywords. The dataset will be stored at an internal AUTH server and will only be accessible by interested AI4Media partners, after they have signed a Joint Controller / Data Sharing Agreement with AUTH, if so required, according to internal AUTH procedures and CA provisions. It will not be discoverable from outside the consortium. Search keywords: N/A Versioning: N/A Metadata creation: Date, Filename, File size, References, History
Making data openly accessible	Data openly accessible: The dataset will not be openly accessible. Redistribution of data (Twitter posts) is against the terms of service of Twitter APIs. 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. How it will be accessible: N/A Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: Using Twitter's standard data representation model. Data and metadata vocabularies: The following metadata vocabulary will be used: created_at: UTC time when Tweet was created. id: Unique identifier for Tweet. id_str: The string representation of Tweet unique identifier. text: UTF-8 text of the status update. source: Utility used to post the Tweet.

	truncated: Indicates whether the value of the text parameter was truncated. geo: The geo object of the status. coordinates: The coordinates of the status. in_reply_to_status_id: Original Tweet's ID if the Tweet is a reply. in_reply_to_status_id_str: Original Tweet's ID if the Tweet is a reply. in_reply_to_user_id: Tweet's author ID If the Tweet is a reply. in_reply_to_user_id_str: Tweet's author ID If the Tweet is a reply. in_reply_to_screen_name: Screen name of original Tweet's author if the Tweet is a reply. user: The user who posted this Tweet. coordinates: Geographic location of Tweet. place: Tweet is associated with a place. quoted_status_id: Tweet ID of the quoted Tweet. quoted_status_id_str: Tweet ID of the quoted Tweet. is_quote_status: Quoted Tweet indicator. quoted_status: Tweet object of the original Tweet that was quoted. retweeted_status: Original Tweet that was retweeted. quote_count: Number of quotes. reply_count: Number of replies. retweet_count: Number of retweets. favorite_count: Number of likes. geo: Contains place details in GeoJSON format. place_type: Specified the particular type of information represented by this place information, such as a city name, or a point of interest. Use of standard vocabularies: Standard vocabularies as used in the Twitter APIs. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will not be shared. Twitter data is used by complying to Twitter's Developer Agreement and Policy (https://developer.twitter.com/en/developer-terms/agreement-and-policy.html) Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: Data quality assurance will be ensured through a data

	pre-processing step, including data cleaning and harmonization as part of data crawling.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data will be stored at an internal AUTH server in a pseudonymized form. A Data Protection Impact Assessment will be performed before storage, according to GDPR provisions, in order to identify proper data security measures.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Twitter data cannot be redistributed because of the terms of service of Twitter APIs. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

4.3.2 Covid-19 Twitter dataset

DMP component	**AI4Media_Data_05_WP6_SOCIALMEDIA_Covid19Twitter_v1** **Partner: BSC**
Data Summary	Purpose: The dataset includes COVID-19 related tweets. The tweets were collected in real time using Twitter's stream API and a set of appropriate keywords and hashtags. The collection started from March 24th, 2020 and continues in real time. This data is used in Task 6.4 "AI for Healthier Political Debate" mainly to research the properties of COVID-19 related discussions and healthy discussions on social media in general. The dataset is used to train and test the new algorithms developed within T6.4. Type/format: MongoDB database composed of JSON files (one file per tweet) Re-use of existing data: The original data was stored by Twitter. This dataset was originally collected for several research projects including AI4Media. Data origin: Twitter API Expected size: Size at 09/02/2021: compressed: ~2 TB, uncompressed ~ 6.5 TB Data utility: The data will be used by T6.4 partners for research on topics including, but not limited to discussion healthiness estimation, argumentation mining and analysis, deep fake video analysis, long text analysis and sentiment analysis. Dataset will be available to other partners within AI4Media to conduct research in other topics as well.
Making data findable, incl. provisions for metadata	Is data discoverable: The tweets within the dataset are discoverable via twitter.com and via twitter API. The dataset is stored on the BSC servers and will be available to AI4Media partners on demand. The dataset will not be made discoverable outside of the consortium. Search keywords: N/A Versioning: N/A Metadata creation: The data is in the raw format provided by the Twitter Streaming

	API, with no modifications nor metadata added.
Making data openly accessible	Data openly accessible: The dataset will not be openly accessible. Redistribution of data (Twitter posts) is against the terms of service of Twitter APIs. 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. How it will be accessible: N/A Methods/software tools to access data: The data is accessed via MongoDB API, either directly via secure connections or from code (for instance, via pymongo package for Python) Repository: N/A Restrictions on access: The access is provided to AI4Media partners on demand.
Making data interoperable	Interoperability: Using Twitter's standard data representation model. Data and metadata vocabularies: The following metadata vocabulary is used: *_id*: MongoDB unique identifier for the Tweet. *created_at*: UTC time when Tweet was created. *timestamp_ms*: Unix timestamp in miliseconds. *id*: Unique identifier for Tweet. *id_str*: The string representation of Tweet unique identifier. *text*: UTF-8 text of the status update. *source*: Utility used to post the Tweet. *truncated*: Indicates whether the value of the text parameter was truncated. *in_reply_to_status_id*: Original Tweet's ID if the Tweet is a reply *in_reply_to_status_id_str*: Original Tweet's ID if the Tweet is a reply *in_reply_to_user_id*: Tweet's author ID If the Tweet is a reply *in_reply_to_user_id_str*: Tweet's author ID If the Tweet is a reply *in_reply_to_screen_name*: Screen name of original Tweet's author if the Tweet is a reply *user*: The user who posted this Tweet. *coordinates*: Geographic location of Tweet *place*: Tweet is associated with a place *quoted_status_id*: Tweet ID of the quoted Tweet. *quoted_status_id_str*: Tweet ID of the quoted Tweet. *is_quote_status*: Quoted Tweet indicator *quoted_status*: Tweet object of the original Tweet that was quoted. *retweeted_status*: Original Tweet that was retweeted.

	quote_count: Number of quotes *reply_count*: Number of replies *retweet_count*: Number of retweets *favorite*_count: Number of likes *entities*: Entities which have been parsed out of the text of the Tweet *extended_entities*: Entities which have been parsed out of the text of the Tweet if there is media content *favorited*: Indicates whether this Tweet has been liked by the authenticating user *retweeted*: Indicates whether this Tweet has been Retweeted by the authenticating user *possibly_sensitive*: URL contained in the Tweet may contain content or media identified as sensitive *filter_level*: The maximum value of the filter_level parameter to still stream this Tweet *lang*: Machine-detected language of the Tweet text *matching_rules*: Provides the id and tag associated with the rule that matched the Tweet *geo*: [deprecated] Coordinates in [lat, long] format Use of standard vocabularies: Standard vocabularies as used in the Twitter APIs, plus 2 additional database fields. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will not be shared. Twitter data is used by complying to Twitter's Developer Agreement and Policy (https://developer.twitter.com/en/developer-terms/agreement-and-policy.html) Availability for re-use: N/A Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: Twitter data is collected as originally provided by Twitter API. Additional post-processing of data is left for the researchers.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data will be stored in BSC servers. Access will be allowed only via ad hoc credentials created for each user.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Twitter data cannot be redistributed because of the terms of service of Twitter APIs. Is informed consent for data sharing and long term preservation given: N/A

Other Issues	N/A

4.3.3 Twitter Text dataset

DMP component	**AI4Media_Data_06_WP6_SOCIALMEDIA_TextTwitter_v1** **Partner: BSC**
Data Summary	Purpose: The dataset contains reduced versions of COVID-19 related tweets (mostly the text field and attachments) from the COVID-19 Twitter dataset described in section 4.3.2. It uses the Solr database functionality to provide efficient text search throughout the whole collection. The tweets were collected in real time using Twitter's stream API and a set of appropriate keywords and hashtags. The collection started from March 24th, 2020 and continues in real time. These tweets were than parsed, relevant fields processed and extracted, and then added to the Solr DB. This data is used in task 6.4 "AI for Healthier Political Debate" mainly to research the properties of COVID-19 related discussions and healthy discussions on social media in general. The dataset is used to train and test the new algorithms developed within T6.4. Type/format: Solr database Re-use of existing data: The original tweets were stored by Twitter. This dataset is based on the dataset described in section 4.3.2, originally collected by BSC. Data origin: Twitter API Expected size: ∽ 160 GB Data utility: The data will be used by T6.4 partners for research on topics including, but not limited to discussion healthiness estimation, argumentation mining and analysis, deep fake video analysis, long text analysis and sentiment analysis. Dataset will be available to other partners within AI4Media to conduct research in other topics as well.
Making data findable, incl. provisions for metadata	Is data discoverable: The tweets within the dataset are discoverable via twitter.com and via twitter API. The dataset is stored on the BSC servers and will be available to AI4Media partners on demand. The dataset will not be made discoverable outside of the consortium. Search keywords: N/A Versioning: N/A Metadata creation: The data is in the raw format provided by the Twitter Streaming API, with no modifications nor metadata added.
Making data openly accessible	Data openly accessible: The dataset will not be openly accessible. Redistribution of data (Twitter posts) is against the terms of service of Twitter APIs. 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. How it will be accessible: N/A Methods/software tools to access data: The data is accessed via Solr API, either directly via secure connections or from code (for instance, via pymongo package for Python)

	Repository: N/A Restrictions on access: The access is provided to AI4Media partners on demand.
Making data interoperable	Interoperability: If the same data schemas are used, different datasets can be merged. Data and metadata vocabularies: The following metadata vocabulary is used: *id*: Unique Tweet id as provided by Twitter *timestamp_ms*: Unix timestamp (in miliseconds) of when the Tweet was created_at *text*: Text of the Tweet. If the Tweet is a retweet or quote, text of the original tweet is appended to this field *lang*: Machine-detected language of the Tweet text (provided by Twitter) *attachment_photos*: JSON list of links to the photos attached to the Tweet *attachment_videos*: JSON list of links and metadata of the videos attached to the Tweet, for versions of different quality *attachment_texts*: Tesseract character recognition algorithms was applied to photos attached to the Tweet. If more than 40 symbols is extracted, the extracted text is contained in this field (and is fully searchable) *attachment_gifs*: JSON list of links and metadata of the GIFs attached to the Tweet, as videos of different quality *attachment_urls:* If the Tweet contained a URL link to anything except a Tweet, it is contained here. Use of standard vocabularies: When applicable, Twitter API field names were used. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will not be shared. Twitter data is used by complying to Twitter's Developer Agreement and Policy (https://developer.twitter.com/en/developer-terms/agreement-and-policy.html) Availability for re-use: N/A Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: Twitter data is collected as originally provided by Twitter API. Text extraction from photos is performed via Tesseract OCR, a top-end character recognition and extraction module.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data will be stored in BSC servers. Access will be allowed only via ad hoc credentials created for each user.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Twitter data cannot be redistributed because of the terms of service of Twitter APIs. Is informed consent for data sharing and long term preservation given: N/A

Other Issues	N/A

4.3.4 ENF-Presence audio dataset

DMP component	AI4Media_Data_07_WP6_Audio_ENF-Presence_v1 Partner: FhG-IDMT
Data Summary	Purpose: The ENF-presence dataset is going to be a collection of audio files with or without traces of Electrical Network Frequency (ENF). The dataset can be used to train classifiers for ENF detection. The dataset is going to be composed by ENF tracks corrupted with different kind of noises with varying SNRs, to prepare a common set for benchmarking ENF detection and extraction. Type/format: Audio (wav) Re-use of existing data: We foresee the re-use of pre-existing background noise files. Data origin: Own recordings Expected size: ~4GB (four different variants) Data utility: The dataset is useful in the context of T6.2 for the detection of ENF traces, by means of which it is possible to identify the presence of content manipulations, as well as to determine the time of a recording.
	Is data discoverable: We plan to make the data discoverable by means of a related publication. A Publication would allow for indexing on common search engines, and on search by keywords. Search keywords: N/A Versioning: N/A Metadata creation: We plan to release not only ENF audio files, but also an accompanying metadata files with ground-truth information on its trajectory
Making data openly accessible	Data openly accessible: We plan to make the data openly accessible and re-distributable. How it will be accessible: We plan to host the data on an open-access repository, and to let it be shared also through the AI4Media platform, to maximize visibility. Methods/software tools to access data: Web-browser to download the data as zip file. Repository: Zenodo or equivalent Restrictions on access: N/A
Making data interoperable	Interoperability: We plan to use basic formats such as wav and csv/xml files, with accompanying instructions and a full description in a related publication. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is going to be released under a copyleft Creative Commons

	Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license Availability for re-use: Yes, if the license terms and conditions are fulfilled. Usable by third parties after end of project: Data is going to be publicly shared also after the end of the project. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: We plan to rely on open repositories with several measures in place for preventing data loss and corruption by unauthorized authors.
Ethical aspects	Possible ethical and legal aspects preventing sharing: None. No person-related data nor otherwise sensitive data are going to be included. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.3.5 ENF-Discontinuity audio dataset

DMP component	**AI4Media_Data_08_WP6_Audio_ENF-Discontinuity_v1** **Partner: FhG-IDMT**
Data Summary	Purpose: The ENF-discontinuity dataset is going to be a collection of audio files with traces of discontinuous Electrical Network Frequency (ENF). The dataset can be used to train classifiers for ENF discontinuity detection, robust to different kind of manipulations. The dataset is going to be composed by discontinuous ENF tracks corrupted with different kind of noises with varying SNRs, to prepare a common set for benchmarking ENF discontinuity analysis. Several kind of discontinuities are going to be considered – e.g., sudden appearance, sudden disappearance, phase discontinuities, frequency discontinuities. Type/format: Audio (wav) Re-use of existing data: We foresee the re-use of pre-existing background noise files. Data origin: Own recordings Expected size: ~16GB (four different variants) Data utility: The dataset is useful in the context of T6.2 for the detection of ENF traces, by means of which it is possible to identify the presence of content manipulations, as well as to determine the time of a recording.
	Is data discoverable: We plan to make the data discoverable by means of a related publication. A publication would allow for indexing on common search engines, and on search by keywords. Search keywords: N/A Versioning: N/A

	<u>Metadata creation</u>: We plan to release not only ENF audio files, but also an accompanying metadata files with ground-truth information on the location of the discontinuities
Making data openly accessible	<u>Data openly accessible</u>: We plan to make the data openly accessible and re-distributable. <u>How it will be accessible</u>: We plan to host the data on an open-access repository, and to let it be shared also through the AI4Media platform, to maximize visibility. <u>Methods/software tools to access data</u>: Web-browser to download the data as zip file. <u>Repository</u>: Zenodo or equivalent <u>Restrictions on access</u>: N/A
Making data interoperable	<u>Interoperability</u>: We plan to use basic formats such as wav and csv/xml files, with accompanying instructions and a full description in a related publication. <u>Data and metadata vocabularies</u>: N/A <u>Use of standard vocabularies</u>: N/A <u>Mappings to commonly used vocabularies</u>: N/A
Increase data re-use	<u>Licence</u>: The data is going to be released under a copyleft Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license <u>Availability for re-use</u>: Yes, if the license terms and conditions are fulfilled. <u>Usable by third parties after end of project</u>: Data is going to be publicly shared also after the end of the project. <u>Re-use timeframe</u>: N/A <u>Data quality assurance process</u>: N/A
Allocation of resources	<u>Costs for making data FAIR</u>: N/A <u>Costs for long-term preservation</u>: N/A
Data security	<u>Security measures</u>: We plan to rely on open repositories with several measures in place for preventing data loss and corruption by unauthorized authors.
Ethical aspects	<u>Possible ethical and legal aspects preventing sharing</u>: None. No person-related data nor otherwise sensitive data are going to be included. <u>Is informed consent for data sharing and long term preservation given</u>: N/A
Other Issues	N/A

4.4 Datasets collected in the context of WP8

4.4.1 Data from user research activities in Use Case 1

DMP component	**AI4Media_Data_09_WP8_USER-RESEARCH_UseCase1_DW_v1** **Partner: DW**
Data Summary	<u>Purpose</u>: In order to realise Use Case 1 in WP8, partner DW will design various research activities with *professional* users, who work in media or fact-checking

	organisations (journalists, verification experts or managers). This research is required for the final set of user requirements to be provided at M12 as well as the user evaluation of two planned demonstrators (journalism tools for professional users that contain new AI-functionalities) - starting from M19. At this early point in time (M6), the following, mainly qualitative research methods are planned with small, selected target groups: tailored questionnaires, interviews, focus groups or demo/evaluation sessions with single users. User research of this kind identifies the users' needs and their opinion regarding new functions/tools provided by the R&D project (aspects such as usefulness, usability, business and journalistic value, ethics or performance). By conducting this research with professional users, DW will generate a dataset containing *Participant Personal Data* and *User Responses*. This data will be stored internally on DW's systems/servers and it is not planned - at this point in time - to share this data with external organisations or AI4Media project partners. Participant Personal Data is likely to include information such as Name, Company Email, Company, Position, and – if necessary – Area of Work and Additional Contact Details. The data will be used for internal analysis purposes by DW and the production of aggregated results summaries in Deliverables D8.1 (M12), D8.3 (M24) and D8.6 (M48). Type/format: Documents containing Participant Personal Data and documents containing questions/user responses from user research activities. Re-use of existing data: No. Data origin: Internal research to identify potential user research participants, questionnaires or other research methods (focus groups, demo sessions or interviews). Expected size: A few KBs per document - a few MBs in total. Data utility: This data will be used in the context of WP8 to define the final set of user requirements and to evaluate the upgraded demonstrators (see above). Analysed, aggregated results will be used by use case and technical partners to develop/improve functionalities, the demonstrators and the plans for commercial exploitation.
Making data findable, incl. provisions for metadata	Is data discoverable: No, because the dataset described will be stored on DW's internal corporate systems and there are no plans for data sharing. The deliverables containing aggregated, analysed results are classified as confidential (D8.1, D8.3 and D8.6) and therefore only available to the consortium. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data will not be openly accessible since it contains personal information. How it will be accessible: It is planned that the data will only be accessible by project partner DW. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A

	Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset described will be stored on internal corporate systems/servers of DW. Data handling and protection follows standard DW operations and national/EU guidelines. IT security measures mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The data will not be shared since it contains personal information of end users. Is informed consent for data sharing and long-term preservation given: An *Informed Consent Form* or similar methods to obtain consent will be prepared for potential participants in the user research activities described above. This will include information about the purpose of the research, the level of anonymisation of personal information provided, how responses are used/reported and data treatment/compliance.
Other Issues	No

4.4.2 Questionnaires for the collection of user requirements for Use Case 3

DMP component	**AI4Media_Data_10_WP8_QUESTIONNAIRE_UseCase3-UserReqCollection2021_V1** **Partner: RAI**
Data Summary	Purpose: A structured questionnaire will be developed by RAI to collect user opinions about end user requirements proposed by use case 3. The questionnaire will include questions about each presented feature/epic/user story; it will try to understand if and how each feature can be useful in a production workflow and to highlight a possible development priority. The questionnaire will be filled by RAI journalist/editorial staff people/archivists and possibly by people with same profiles belonging to other broadcasters. The collected questionnaires will then be analysed and the results of this analysis will be used to drive use case 3 features development. The results and related analyses will then be presented in D8.1. Type/format: Text documents containing questions and user responses. Re-use of existing data: No Data origin: Questionnaires filled by end-users in the context of use case 3 during the year 1 of the project. Expected size: A few KBs per questionnaire. A few MBs in total. Data utility: This data will be used in the context of WP8 to evaluate requirements

	proposed by use case 3. The consortium will use requirements evaluation results to define a development plan for use case 3 in the project lifetime.
Making data findable, incl. provisions for metadata	Is data discoverable: The user requirement evaluation questionnaires for use case 3 will be stored on RAI servers. A summary of the requirements evaluation results will be available in D8.1. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The raw data will not be openly accessible since it contains personal information. It is protected and shared only among project partners and a summary of this data will be presented in D8.1. In case of a report or paper submitted for publication, all findings will be integrated into the report or paper. Datasets will not be added to the publication. How it will be accessible: The dataset will be stored on RAI servers; the questionnaires are only accessible by project partners. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: N/A Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: Data quality has been assured by asking end users to fill out the questionnaire in their own languages. Feedback collected has been translated to English, in order to ensure accurate data collection and analysis.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data will be stored on local repositories inside RAI's premises, subject to the same security measures already used for IT infrastructure in RAI. These include network isolation from external internet accesses, firewalling, account-based access control management to the storage where the data copies are located. RAI fully complies with the applicable national, European data security frameworks, and the GDPR.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The daset will not be shared since it may contain personal information of end-users. Is informed consent for data sharing and long term preservation given: An Informed

	Consent Form will be prepared for the participation to the requirements collection activity in case treatment of personal data will be needed. In that case, the 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 analytical and statistical purposes.
Other Issues	No

4.4.3 Questionnaires for the evaluation of Use Case 3

DMP component	**AI4Media_Data_11_WP8_QUESTIONNAIRE_UseCase3Evaluation_V1** **Partner: RAI**
Data Summary	Purpose: Structured questionnaires will be developed by RAI for the evaluation of the developed AI services/tools in the context of UC3 pilot trials. The questionnaires will include questions that cover issues such as usefulness, usability, visualisation & 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 for all evaluation phases. Data collected through the questionnaires is used exclusively for analytical and statistical purposes. The evaluation results are presented in relevant WP8 deliverables. Type/format: Text documents containing questions and user responses. Re-use of existing data: No. Data origin: Questionnaires filled by end-users in the context of use case 3 during the different pilot phases. Expected size: A few KBs per questionnaire. A few MBs in total. Data utility: This data will be used in the context of WP8 to evaluate different versions of AI4Media services/tools for use case 3. The evaluation results of each pilot phase will be used by the technical partners to improve and extend the functionalities of the developed services/tools during the next development phases. The final evaluation results will help partners to improve these services/tools as part of further commercial exploitation.
Making data findable, incl. provisions for metadata	Is data discoverable: The evaluation questionnaires (raw data) will be stored on RAI servers. A summary of the evaluation results will be presented in relevant WP8 deliverables. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The raw data will not be openly accessible since it contains personal information. It is protected and shared only among project partners. However, a summary of the evaluation results will be presented in relevant WP8 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. How it will be accessible: The dataset will be stored on RAI servers; the questionnaires are only accessible by project partners.

	Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: Data quality has been assured by asking end users to fill out the evaluation questionnaire in their own languages. Feedback collected has been translated to English, in order to ensure accurate data collection and analysis.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data will be stored on local repositories inside RAI's premises, subject to the same security measures already used for IT infrastructure in RAI. These include network isolation from external internet accesses, firewalling, account-based access control management to the storage where the data copies are located. RAI fully complies with the applicable national, European data security frameworks, and the GDPR.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The datset will not be shared since it may contain personal information of end-users. Is informed consent for data sharing and long term preservation given: An Informed Consent Form will be prepared for the participation to the requirements collection activity in case treatment of personal data will be needed. In that case, the 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 analytical and statistical purposes.
Other Issues	No

4.4.4 Data from user research activities in Use Case 4

DMP component	**AI4Media_Data_12_WP8_USER-RESEARCH_UseCase4_NISV_v1** **Partner: NISV**
Data Summary	Purpose: In order to realise Use Case 4 in WP8, partner NISV will perform various qualitative research activities with researchers form the humanities and social sciences. This research is required to define the user requirements brought forward in the project, as well as evaluate the demonstrator built on the AI-based tooling delivered by the project. The following, mainly qualitative research methods are planned with small, selected target groups: tailored questionnaires, interviews, focus

	groups or demo/evaluation sessions with single users. By conducting this research with professional users, NISV will generate a dataset containing *Participant Personal Data* and *User Responses*. This data will be stored internally on NISV's systems/servers and it is not planned - at this point in time - to share this data with external organisations or AI4Media project partners. Participant Personal Data is likely to include information such as Name, Company Email, Company, Position, and – if necessary – Area of Work and Additional Contact Details. The data will be used for internal analysis purposes by NISV and the production of aggregated results summaries in Deliverables D8.1 (M12), D8.3 (M24) and D8.6 (M48). Type/format: Documents containing Participant Personal Data and documents containing questions/user responses from user research activities. Re-use of existing data: No. Data origin: Internal research to identify potential user research participants, questionnaires or other research methods (focus groups, demo sessions or interviews). Expected size: A few KBs per document - a few MBs in total. Data utility: This data will be used in the context of WP8 to define the final set of user requirements and to evaluate the upgraded demonstrators (see above) for Use Case 4. Analysed, aggregated results will be used by use case and technical partners to develop/improve functionalities, the demonstrators and the plans for commercial exploitation.
Making data findable, incl. provisions for metadata	Is data discoverable: No, because the dataset described will be stored on NISV's internal corporate systems and there are no plans for data sharing. The deliverables containing aggregated, analysed results are classified as confidential (D8.1, D8.3 and D8.6) and therefore only available to the consortium. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data will not be openly accessible since it contains personal information. How it will be accessible: It is planned that the data will only be accessible by project partner NISV. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: N/A

	Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset described will be stored on internal corporate systems/servers of NISV. Data handling and protection follows standard NISV operations and national/EU guidelines. IT security measures mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The data will not be shared since it contains personal information of end users. Is informed consent for data sharing and long-term preservation given: An *Informed Consent Form* or similar methods to obtain consent will be prepared for potential participants in the user research activities described above. This will include information about the purpose of the research, the level of anonymisation of personal information provided, how responses are used/reported and data treatment/compliance.
Other Issues	No

4.4.5 Data from user research activities in Use Case 7

DMP component	**AI4Media_Data_13_WP8_QUESTIONNAIRE_VIDEO_UseCase7-UserFeedbackData_v1** **Partner: IMG**
Data Summary	Purpose: IMG shall perform interviews with users of the 2 demonstrators in Use Case 7. We shall also be filming test users in the Living Lab trials we conduct (content managers and team members) using our technology for demonstration and improvement of such technology. A structured questionnaire will be developed by IMG to collect user opinions about the two demonstrators in use case 7. Along with the questionary images, videos and heatmaps of how users interact with the two demonstrators will be captured. The collected feedback will then be analysed, and the results of this analysis will be used to drive use case 7 demonstrators' further improvement. Type/format: Text documents containing questions and user responses, along with image and video, as well as heatmap, of the user's interaction with the two demonstrators. Re-use of existing data: No Data origin: Questionnaires filled by end-users in the context of use case 7. Images, videos and heatmaps of users' interaction with the two demonstrators. Expected size: Several MBs per user. Several GBs in total. Data utility: The data from the tests on IMG's two technology demonstrators with real users from media companies, in Living Lab trials, will be used to validate their performance and help tailor them further to meet user requirements.
Making data findable, incl.	Is data discoverable: The user feedback data for use case 7 will be securely stored on

provisions for metadata	IMG servers. A summary of the feedback data results will be available in D8.1. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The raw data will not be openly accessible since it contains personal information. A summary of this data will only be presented in D8.1. In case of a report or paper submitted for publication, all findings will be integrated into the report or paper. Datasets will not be added to the publication. How it will be accessible: The dataset will be securely stored on IMG servers; the data will be only accessible by IMG. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: N/A Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: The data collection process will be supervised by IMG.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data will be stored on local repositories inside IMG's premises, subject to the same security measures already used for IT infrastructure in IMG. These include network isolation from external internet accesses, firewalling, and access control management to the storage where the data copies are located. IMG fully complies with the applicable national, European data security frameworks, and the GDPR.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The dataset will not be shared since it may contain personal information of end-users. Is informed consent for data sharing and long term preservation given: We shall provide consent forms for people to take part in the trials and to share this data with IMG for the purpose of improving our technology and software demonstrators. We will include a Privacy Notice that specifies that the treatment of the data is confidential, complies with GDPR and is carried out exclusively for analytical and statistical purposes.
Other Issues	No

4.4.6 TruthNest dataset

DMP component	AI4Media_Data_14_WP8_Text_ThruthNestDataset-UseCase1-ATC_v1 Partner: ATC
Data Summary	Purpose: TruthNest is a tool which extracts and analyzes data related to a Twitter account and produces useful information and insights based on rule-based algorithms. In the context of WP8, TruthNest will be enriched with AI tools coming from WP3-WP6. The collected dataset will contain metadata-enriched content aggregated from Twitter regarding a specific Twitter account. The aggregated content is presented in JSON format. The dataset will contain also tags and alerts which are extracted by a rule-based algorithm. Type/format: JSON files Re-use of existing data: No Data origin: Twitter data and TruthNest analytics Expected size: Several MBs Data utility: The plan is to produce such analyses for a significant number of twitter accounts and provide access to this data to project partners in order to train their AI modules.
Making data findable, incl. provisions for metadata	Is data discoverable: No. The data will be accessible only by AI4Media partners developing tools for Use Case 1. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The dataset will not be shared as it contains personal data. The dataset will only be used internally by consortium partners that will develop AI tools based on TruthNest for training purposes. How it will be accessible: Through custom API available to AI4Media partners only. Methods/software tools to access data: N/A Repository: The data will not be shared. The data will be stored on ATC's internal servers and will be only accessible to ATC personnel working in AI4Media and to selected project partners for training purposes. Restrictions on access: N/A
Making data interoperable	Interoperability: No Data and metadata vocabularies: No Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: No Usable by third parties after end of project: No Re-use timeframe: N/A

	Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data are stored on ATC's servers. Security measures implemented for data protection include controlled access, user authentication, firewalls, VPNs, encryption, and back-ups. Only ATC employees working with TruthNest have access to the data.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The dataset will not be openly accessible. Redistribution of Twitter data is against the terms of service of Twitter APIs. Moreover, TruthNest data are proprietary and will not be shared with anyone outside the AI4Media consortium. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.4.7 Truly Media dataset

DMP component	**AI4Media_Data_15_WP8_Text_TrulyMediaDataset-UseCase1-ATC_v1** **Partner: ATC**
Data Summary	Purpose: Truly Media will be used in the context of Use Case 1 as the base demonstartor. The dataset will be created for testing purposes. It will contain a selection of social media posts (from Twitter, Facebook, VKontakte, Reddit,YouTube) and other web content (news articles, blog posts) depending on the interests and searches made by Truly Media users. The dataset will contain a complete social media posts and their content, along with all the accompanying data, such as user name, date and time of post, profile picture, likes/retweets etc. The extent of the accompanying data depends on the API specifications of each social media platform (e.g. Twitter provides number of likes for a post, while Facebook does not). No private social media data will be included in the dataset. All social media data will be collected through APIs. The dataset will also contain notes and annotations made to the social media posts and other news items by Truly Media users, as well as manually inserted information to Truly Media by its users, such as the location of an event or a post, other related URLs, similar images found online, etc. Ai4media AI tools will use the relevant data in order to extract useful information and produce insights to be used by use case end-users for detecting misinformation attempts. Type/format: JSON files Re-use of existing data: No Data origin: Social media (Twitter, Facebook, VKontakte, Reddit,YouTube) and web Expected size: A few MBs. Data utility: The dataset will be used by AI4Media partners that develop tools for Use Case 1 for training purposes.
Making data findable, incl.	Is data discoverable: That data will be discoverable only by AI4Media partners that will develop tools for Use Case 1. The dataset will be accessible only through Truly Media's

provisions for metadata	environment and will be stored on Truly Media's servers. Search keywords: N/A Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: The dataset will not be shared outside the AI4Media consortium. How it will be accessible: N/A Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: No Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: No Usable by third parties after end of project: No Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be stored by ATC on third-party cloud servers. Appropriate and detailed security policies, rules, and technical measures are implemented to protect data that are used by the Truly Media platform and stored on the platform from improper or unauthorized access, including use of firewalls where appropriate. Security measures also include 2FA (2 Factor Authentication) with OTP (One Time Password) for extra security during login, as well as Auth2.0 and JWT for authentication and authorisation. End-to-end encryption protects from man-in-the-middle attacks and data theft. All ATC employees and data processors, who have access to and are associated with the processing of personal data, are obliged to respect the confidentiality of the stored personal data. Moreover, ATC's development team has received training from external auditors for security awareness and security best practices to avoid vulnerabilities in source code. The external auditors have performed black-box penetration testing to ensure that the platform is fully secure. ATC's Data Protection Officer ensures that all processes followed are fully compliant with the GDPR provisions.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The dataset will not be shared outside AI4Media since it contains personal data. Only AI4Media partners that develop tools for UC1 and ATC employees working on the project will have access to this dataset.

	Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.4.8 UI pose dataset for Use Case 5

DMP component	AI4Media_Data_16_WP8_Text_ UI_pose -UC5-IDF_v1 Partner: IDF
Data Summary	Purpose: This dataset will contain short video clips of IDF employees performing a given action (e.g. standing still, shifting to the left, etc.) and the associated label. The UI pose dataset will be used in the context of Use Case 5 to create a demonstrator. Type/format: movie files Re-use of existing data: Partially. Data origin: Videos of IDF employees created internally in the company. Expected size: A few GBs. Data utility: Demonstration and improvement of the technology in the context of use case 5.
Making data findable, incl. provisions for metadata	Is data discoverable: No Search keywords: N/A Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: The dataset will not be shared as the consent form signed by the subjects does not allow it. The data will be stored internally in IDF's servers. How it will be accessible: N/A Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: No Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally by IDF. Availability for re-use: No Usable by third parties after end of project: No Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be stored on IDF servers for the duration

	mentioned in the consent form. IDF fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The dataset will not be shared outside IDF since it contains personal data and the consent form signed by the subjects does not allow sharing outside IDF. Is informed consent for data sharing and long term preservation given: No

4.4.9 Game glitches dataset for Use Case 5

DMP component	AI4Media_Data_17_WP8_Video_GameGlitches-UC5-MODL_v1 Partner: MODL
Data Summary	Purpose: This dataset will contain several short videos of games' glitches and corresponding glitch type annotation recorded and collected from modl.ai employees. This dataset will be used in the context of Use Case 5 (feature 5A) to train an AI capable of recognizing automatically different kinds of game glitches. Type/format: movie data (.mp4) Re-use of existing data: No. Data origin: Videos of game glitches created internally in the company (MODL). Expected size: Several GBs. Data utility: Training data that will be used to improve the technology in the context of Use Case 5A.
Making data findable, incl. provisions for metadata	Is data discoverable: No Search keywords: N/A Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: The dataset will not be shared as it contains personal data and games developed internally by modl.ai for the Use Case 5A's purposes only. The data will be stored exclusively in modl.ai's server. How it will be accessible: N/A Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: No Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally by modl.ai. Availability for re-use: No

	Usable by third parties after end of project: No Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be stored in modl.ai's servers only. Modl.ai complies with the requirements of the European directive for personal data protection and the data is secured by state-of-the-art IT security measures that mitigate most of the possible illegitimate access risks.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The dataset will not be shared outside modl.ai since it contains personal data and games developed internally by modl.ai. Is informed consent for data sharing and long term preservation given: No

4.4.10 Musical production for AI co-creation dataset

DMP component	**AI4Media_Data_18_WP8_Audio_RAWcompositions_v1** **Partner: BSC**
Data Summary	Purpose: New AI4Media dataset generated in audio format from the Demonstrator of AI co-creation in WP8. The Demonstrator for use case 6 aims to produce a set of tools to assist artists -especially music compositors- to produce in an easy and accessible manner novel creations with the help of an AI engine. This approach requires the set up of a platform to load, train and experiment with generative models. These models are trained from an initial dataset, in our case principally formed by a collection of RAW audio files. These audio files may be decorated with a set of labels to extend the desciption of each track. Once trained, these models are used by the artist to generate new collections of audio material that can be used for further training of new models, to be used as an inspiration material for novel human compositions, or directly as a audio used in novel content. The input dataset used for training may be selected from an existing collection of audio material, but the output material has to be presented in a consisted way for exploration and analysis by the artist and developers. Along with the platform for audio geneation and model selection, this audio dataset is the main validation of the demonstrator. Type/format: Raw audio data, uncompressed music Re-use of existing data: No. Data origin: Generated audio from a ML application. Expected size: Around 2GB. Data utility: Useful to creators using audio to improve the composition of new music. Outside the project participants, general audience of music compositions.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is not uniquely identifiable, but can be labeled according to some label set. Search keywords: No.

	Versioning: Yes, version numbers are assigned according to the training conditions. Metadata creation: Labels referred to the generation of creation, and conditions of the generative process.
Making data openly accessible	Data openly accessible: Data will be openly accessible when required. Access will be granted through BSC data storage facilities, through an open repository like Zenodo, or through a dedicated channel for audio file sharing, like soundcloud. Special cases for sharing content include if an author/creator uses his own data with the generative model; then, data may be partially restricted. How it will be accessible: Online repository. Methods/software tools to access data: Raw audio is made accessible online. Repository: Online repository – TBD. Restrictions on access: Access control for restricted.
Making data interoperable	Interoperability: No. Data and metadata vocabularies: NA Use of standard vocabularies: NA Mappings to commonly used vocabularies: NA
Increase data re-use	Licence: Data will be shared under a Creative Commons CC Zero License. Availability for re-use: Data available without embargo. Usable by third parties after end of project: Data available for further analysis. Re-use timeframe: Data will be in audio format, with no restriction on use. Data quality assurance process: Data is in audio format, but quality of these tracks cannot objectively be assessed.
Allocation of resources	Costs for making data FAIR: Audio data in our case is not subject to FAIR criteria. Costs for long-term preservation: Server hosting, without additional maintenance.
Data security	Security measures: Controlled-access for restricted parts of the dataset, with periodic backup copies.
Ethical aspects	Possible ethical and legal aspects preventing sharing: We don't foresee any ethical implications for sharing our dataset. Is informed consent for data sharing and long term preservation given: Not needed.
Other Issues	No

4.4.11 User survey data for AI industrial needs (for T8.4)

DMP component	AI4Media_Data_19_WP8_ QUESTIONNAIRE_AI-IndustrialNeeds-T8.4_v1 Partner: ATC
Data Summary	Purpose: Structured questionnaires will be developed by WP8 partners in the framework of T8.4 "Harmonising AI research with industrial needs" for the collection of input by external stakeholders. The questionnaires will include questions that aim to gather input and insights from industry stakeholders regarding the AI needs of media organisations. This dataset will include questionnaires filled by the industry stakeholders. Data collected through the questionnaires will be used exclusively for

	analytical and statistical purposes to guide T8.4 partners in harmonising AI research with the needs of media organisations. Type/format: Text documents containing questions and responses. Re-use of existing data: No Data origin: Survey participants Expected size: A few KBs per questionnaire. A few MBs in total Data utility: This data will be used in the context of WP8, and more specifically T8.4 "Harmonising AI research with industrial needs" for the collection of input by external stakeholders. Data collected through the questionnaires will be used exclusively for analytical and statistical purposes to guide T8.4 partners in harmonising AI research with the needs of media organisations.
Making data findable, incl. provisions for metadata	Is data discoverable: The questionnaires (raw data) will be stored on the servers of the partners that participate in the relevant T8.4 activities. A summary of the results will be included in D8.3 and D8.6, which are confidential and will be accessible only to consortium partners and the EC. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The raw data will not be openly accessible since it contains personal information. It is protected and shared only among selected project partners. A summary of the results will be available through confidential deliverables D8.3 and D8.6. 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. How it will be accessible: Stored in the servers of T8.4 partners that participate in the related activities, the questionnaires are only accessible by WP8 project partners that participate in the related T8.4 activities. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: Data quality will be assured by asking participants to fill out the questionnaire in their own languages, and in the case this is not possible by ensuring that survey participants clearly understand all the questions. Feedback

	collected has been translated to English, in order to ensure accurate data collection and analysis.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The questionnaires will be stored on the partners’ servers. Appropriate security mechanisms enforced by each partner include use of firewalls and restricted access to the folders only to partner employess that work on AI4Media.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The questionnaires may contain personal information of end-users. Is informed consent for data sharing and long term preservation given: An Informed Consent Form will be prepared for the participation to T8.4 validation activities. The 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 analytical and statistical purposes (see D12.1).
Other Issues	No

5. Data management plan for third-party research datasets used in AI4Media

This section presents existing third-party research datasets that will be used by AI4Media partners to develop and test new AI algorithms, methodologies, and tools in the context of WP3,4,5 and 6. Most of these datasets are already publicly shared by their third party owners while some are private datasets that have been provided to AI4Media partners for research purposes.

In total, 51 third-party research datasets have been identified until now. The list is not exhaustive and represents the current status; more datasets may be used in the future, depending on the needs that may be identified during the lifetime of the project. In the following, we present the DMP plan for these datasets using the same template as in Section 4.

The following Table briefly summarizes the 51 datasets presented in this section and offers a glance at the structure of the section and its subsections.

Table 3: Summary of third-party research datasets used in AI4Media

DMP component	WP	Short summary	Relevant sub-section
Data used in WP3 (New Learning Paradigms & Distributed AI)			5.1
AI4Media_Data_20_WP3_EMAIL_Enron_v1	WP3	Enron email dataset	5.1.1
AI4Media_Data_21_WP3_SOCIALMEDIA_FacebookWall_v1	WP3	Facebook Wall dataset	5.1.2
AI4Media_Data_22_WP3_IMAGE_AGAIN _v1	WP3	Affect Game Annotation (AGAIN) dataset	5.1.3
AI4Media_Data_23_WP3_TEXT_IMDBreviews_v1	WP3	IMDB movie reviews dataset	5.1.4
AI4Media_Data_24_WP3_TEXT_MHAD2DPose_v1	WP3	MHAD 2D pose dataset	5.1.5
AI4Media_Data_25_WP3-5_TEXT_20Newsgroups_v1	WP3,5	20Newsgroups dataset	5.1.6
AI4Media_Data_26_WP3-5_TEXT_HPAmazonReviews_v1	WP3,5	HP Amazon reviews dataset	5.1.7
AI4Media_Data_27_WP3-5_TEXT_JRCAcquis_v1	WP3,5	JRCAcquis legislative text dataset	5.1.8
AI4Media_Data_28_WP3-5_TEXT_Kindle_v1	WP3,5	Kindle document dataset	5.1.9
AI4Media_Data_29_WP3-5_TEXT_OHSUMED_v1	WP3,5	OHSUMED MEDLINE document dataset	5.1.10
AI4Media_Data_30_WP3-5_TEXT_RCV1-Reuters_v1	WP3,5	RCV1 Reuters stories dataset	5.1.11
AI4Media_Data_31_WP3-5_TEXT_RCV1RCV2-Reuters_v1	WP3,5	RCV1RCV2 Reuters stories dataset	5.1.12
AI4Media_Data_32_WP3-5_TEXT_Reuters-21578_v1	WP3,5	Reuters-21578 dataset	5.1.13
AI4Media_Data_33_WP3-5_TEXT_11TweetSentiment_v1	WP3,5	11 Tweet Sentiment Datasets	5.1.14
AI4Media_Data_34_WP3-	WP3	WipoGamma patent document	5.1.15

5_TEXT_WipoGamma_v1		dataset	
Data used in WP4 (Explainability, Robustness and Privacy in AI)			5.2
AI4Media_Data_35_WP4_IMAGE_Imagenet_01	WP4	ImageNet-ILSVRC2012 image classification dataset	5.2.1
AI4Media_Data_36_WP4_IMAGE_FFHQ_v1	WP4	FFHQ dataset for GAN training	5.2.2
AI4Media_Data_37_WP4_IMAGE_MNIST_v1	WP4	MNIST image dataset	5.2.3
AI4Media_Data_38_WP4_IMAGE_VIDEO_Interestingness10k	WP4	Interestingness10k image +video dataset	5.2.4
AI4Media_Data_39_WP4_VIDEO_Memorability2020	WP4	MediaEval Memorability 2020 dataset	5.2.5
AI4Media_Data_40_WP4_IMAGE_drawnUI2021	WP4	ImageCLEF DrawnUI 2021 dataset	5.2.6
Data used in WP5 (Content-centered AI)			5.3
AI4Media_Data_41_WP5_VIDEO_SumMeGycli14_v1	WP5	SumMe video summarization dataset	5.3.1
AI4Media_Data_42_WP5_VIDEO_TVSumSong15_v1	WP5	TVSum video summarization dataset	5.3.2
AI4Media_Data_43_WP5_VIDEO_MonumentsOfItaly_v1	WP5	RAI Monuments of Italy dataset	5.3.3
AI4Media_Data_44_WP5_IMAGE_LVIS_v1	WP5	LVIS image dataset	5.3.4
AI4Media_Data_45_WP5_IMAGE_CIFAR_v1	WP5	CIFAR10/100 image dataset	5.3.5
AI4Media_Data_46_WP5_IMAGE_STL10_v1	WP5	STL-10 image dataset	5.3.6
AI4Media_Data_47_WP5_TEXT_CCNET_v1	WP5	CCNet text dataset for multilingual representation learning	5.3.7
AI4Media_Data_48_WP5_VIDEO_360VideoViewingHeadMountedVR_v1	WP5	360 Video Viewing Dataset in Head Mounted Virtual Reality	5.3.8
AI4Media_Data_49_WP5_VIDEO_HeadMovementinPanoramicVideo_v1	WP5	Predicting Head Movement in Panoramic Video Dataset	5.3.9
AI4Media_Data_50_WP5_VIDEO_GazePredictionDynamic360ImmersiveVideos_v1	WP5	Gaze prediction in Dynamic 360° Immersive Videos Dataset	5.3.10
AI4Media_Data_51_WP5_VIDEO_YourAttentionIsUnique_v1	WP5	Your Attention is Unique Dataset	5.3.11
AI4Media_Data_52_WP5_VIDEO_HeadEyeMovements360Videos_v1	WP5	Dataset of Head and Eye Movements for 360° Videos	5.3.12
AI4Media_Data_53_WP5_Audio_dim-sim music_v1	WP5	Dim-sim dataset for music similarity search	5.3.13
AI4Media_Data_54_WP5_Audio_spam_music_v1	WP5	SPAM dataset for music segmentation	5.3.14
AI4Media_Data_55_WP5_Audio_salami_music_v1	WP5	SALAMI dataset for music segmentation	5.3.15
AI4Media_Data_56_WP5_Audio_harmonix_music_v1	WP5	Harmonix dataset for music segmentation	5.3.16
AI4Media_Data_57_WP5_Audio_FMA_v1	WP5	Free Music Archive dataset	5.3.17
AI4Media_Data_58_WP5_Audio_LAK	WP5	LAKH MIDI music dataset	5.3.18

H-MIDI_v1			
AI4Media_Data_59_WP5_Audio_MIDI_Piano_v1	WP5	Piano Audio and MIDI music datasets	5.3.19
AI4Media_Data_60_WP5_Audio_GiantSteps _v1	WP5	GiantSteps music datasets	5.3.20
Data used in WP6 (Human- and Society-centred AI)			5.4
AI4Media_Data_61_WP6_VIDEO_Deepfake-Detection-Challenge-Dataset_v1	WP6	Deepfake Detection Challenge video dataset	5.4.1
AI4Media_Data_62_WP6_VIDEO_ FaceForensics++_v1	WP6	FaceForensics++ video dataset	5.4.2
AI4Media_Data_63_WP4_IMAGE_ImageCLEFaware-dataset_v1	WP6	Visual profile impact rating and ranking – ImageCLEFaware dataset	5.4.3
AI4Media_Data_64_WP6_EEG_DEAP _v1	WP6	DEAP EEG dataset	5.4.4
AI4Media_Data_65_WP6_EEG_SEED _v1	WP6	SEED EEG dataset	5.4.5
AI4Media_Data_66_WP6_EEG_SEED -IV_v1	WP6	SEED-IV EEG dataset	5.4.6
AI4Media_Data_67_WP6_Audio_ Clotho_v1	WP6	Clotho audio captioning dataset	5.4.7
AI4Media_Data_68_WP6_Audio_ ASVspoof_v1	WP6	ASVspoof2019 dataset DMP component	5.4.8
AI4Media_Data_69_WP6_Audio_MOBIPHONE_v1	WP6	MOBIPHONE audio dataset	5.4.9
AI4Media_Data_70_WP6_Audio_ Fake-or-Real_v1	WP6	Fake-or-Real (FoR) audio dataset	5.4.10

5.1 Datasets used in the context of WP3

5.1.1 Enron email dataset

DMP component	AI4Media_Data_20_WP3_EMAIL_Enron_v1 Partner: CERTH
Data Summary	Purpose: The Enron email dataset contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse. It will be used by CERTH in T3.5 to simulate decentralized settings, where each employee is considered to own a separate device. Results involving this dataset will be reported in D3.2 and D3.4. Type/format: Csv file containing timestamp, an anonymized email sender and anonymized email receiver Re-use of existing data: Yes. Data origin: Emails of Enron employees obtained by FERC and available at http://networkrepository.com/ia-enron-email-dynamic.php Expected size: 4.2 MB Data utility: It is useful to WP3 partners to benchmark graph mining algorithms, such as node ranking and graph neural networks on link prediction tasks.
Making data	Is data discoverable: Data is discoverable. The dataset is hosted in the Network

findable, incl. provisions for metadata	Repository. Search keywords: Enron email dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible at http://networkrepository.com/ia-enron-email-dynamic.php How it will be accessible: Shared through a third-party repository link Methods/software tools to access data: Web-browser to download the data as zip file Repository: Network Repository (http://networkrepository.com) Restrictions on access: None
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under an attribution licence (http://networkrepository.com/policy.php) Availability for re-use: The loading and pre-processing mechanism developed for using the dataset in experiments will be made publicly available to ensure reproducibility of research. Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: This dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the dataset should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.2 Facebook wall dataset

DMP component	AI4Media_Data_21_WP3_SOCIALMEDIA_FacebookWall_v1 Partner: CERTH
Data Summary	Purpose: This dataset captures a Facebook friendship graph where nodes are users and edges between the users represent wall post events. It will be used in T3.5 to simulate decentralized settings, where each user is considered to own a separate device. Results involving this dataset will be reported in D3.2 and D3.4. Type/format: Csv file containing timestamp, an anonymized message sender and anonymized message receiver Re-use of existing data: Yes. Data origin: Facebook - http://networkrepository.com/ia-facebook-wall-wosn-dir.php Expected size: 6.7 MB Data utility: It is useful to WP3 partners to benchmark graph mining algorithms, such as node ranking and graph neural networks on link prediction tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable. The dataset is hosted in the Network Repository. Search keywords: facebook-wall-wosn Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible at http://networkrepository.com/ia-facebook-wall-wosn-dir.php How it will be accessible: Shared through a third-party repository link Methods/software tools to access data: Web-browser to download the data as zip file Repository: Network Repository (http://networkrepository.com) Restrictions on access: None
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is already publicly available under an attribution licence (http://networkrepository.com/policy.php) Availability for re-use: The loading and pre-processing mechanism developed for using the dataset in experiments will be made publicly available to ensure reproducibility of research. Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A

	Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the dataset should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.3 Affect Game Annotation (AGAIN) dataset

DMP component	**AI4Media_Data_22_WP3_IMAGE_AGAIN _v1** **Partner: UM**
Data Summary	Purpose: AGAIN is a large-scale affective corpus that features over $1,100$ in-game videos (with corresponding gameplay data) from nine different games, which are annotated for arousal from 124 participants in a first-person continuous fashion. Even though AGAIN is created for the purpose of investigating the generality of affective computing across dissimilar tasks, affect modelling can be studied within each of its 9 specific interactive games. AGAIN will likely be used in WP3 to evaluate affect-driven quality diversity algorithms. Type/format: Annotated in-game video footage images and accompanying telemetry-and metadata. Re-use of existing data: Yes, the dataset is created within the TAMED Marie Curie project. Data origin: Annotated by MTurk workers, based on original artefacts created for the TAMED Marie Cure project: https://again.institutedigitalgames.com/ Expected size: 42.4 GB Data utility: It will be useful to AI4Media partners that investigate affect detection in relation to media.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is made available on the UM IDG Google Drive storage and the following site: https://again.institutedigitalgames.com Search keywords: dataset, videogame, affect, arousal, video repository, affective computing, affective dataset Versioning: Google Drive supports versioning. Metadata creation: Metadata written in a readable and searchable JSON file.
Making data openly accessible	Data openly accessible: The data is openly accessible via https://again.institutedigitalgames.com How it will be accessible: The data can be downloaded from an online archive after completing a form. Methods/software tools to access data: N/A Repository: https://drive.google.com/drive/u/1/folders/1f4eO0A0JH6FE8n5v7ozjJztOb

	PcdLzoJ Restrictions on access: The user should accept the terms of use.
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: datapackage.json includes schemas and field descriptions of the dataset. Use of standard vocabularies: The dataset metadata follows Data Package specification. Mappings to commonly used vocabularies: The Kaggle API implements the same data specifications for datasets.
Increase data re-use	Licence: The data is released under MIT License. Availability for re-use: The data is available openly for reuse for research purposes. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: The data has been sorted and cleaned for further use. Unusable data is removed from the dataset, and both raw unprocessed and preprocessed ready-to-use packages are available from the remaining data.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The full dataset (including images and non-anonymized metadata) will be hosted on UM's servers. UM fully complies with the applicable national, European and International framework, and the GDPR. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: Yes
Other Issues	N/A

5.1.4 IMDB movie reviews dataset

DMP component	AI4Media_Data_23_WP3_TEXT_IMDBreviews_v1 Partner: IDF
Data Summary	Purpose: This dataset includes movie reviews extracted from the IMDB website. It consists in 25,000 training and 25,000 test reviews annotated as positive or negative. This data will be used as benchmark for T3.5 by IDF. Type/format: Raw text and already processed bag of words formats Re-use of existing data: Yes, we use an existing dataset Data origin: https://ai.stanford.edu/~amaas/data/sentiment/ Expected size: 250MB Data utility: As benchmark for T3.5 for IDF.
Making data findable, incl. provisions for	Is data discoverable: Data is available online (https://ai.stanford.edu/~amaas/data/sentiment/) and indexed in Google. It is also

metadata	directly available in some frameworks such as tensorflow. Search keywords: imdb, sentiment analysis, stanford Versioning N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Data has already been made publicly available by its owners. We will not re-share it. How it will be accessible: Accessible for original source: https://ai.stanford.edu/~amaas/data/sentiment/ Methods/software tools to access data: Download the dataset. It is also available directly from some framework such as tensorflow Repository: https://ai.stanford.edu/~amaas/data/sentiment/ Restrictions on access: No
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: Data will not be altered as it is easy to read. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: When using this dataset, please cite *Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Availability for re-use: Data is already available online. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on IDF servers provided that this is not forbidden by the owner of the dataset. IDF fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.5 MHAD 2D pose dataset

DMP component	AI4Media_Data_24_WP3_TEXT_MHAD2DPose_v1 Partner: IDF
Data Summary	Purpose: This is a human action recognition dataset consisting of 2D pose estimations extracted from videos. The 2D pose estimation was performed on a subset of the

	Berkeley Multimodal Human Action Database (MHAD) dataset. Six actions (Jumping in place, Jumping jacks, Punching (boxing), Waving two hands, Waving one hand, Clapping hands) are considered for a total of 1438 videos recorded at 22Hz. This data will be used as benchmark for T3.5 by IDF. Type/format: Text files Re-use of existing data: Yes Data origin: Pose estimation performed on a subset of the MHAD dataset, available at https://github.com/stuarteiffert/RNN-for-Human-Activity-Recognition-using-2D-Pose-Input Expected size: 250MB Data utility: As benchmark for T3.5 for IDF.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available online at Github https://github.com/stuarteiffert/RNN-for-Human-Activity-Recognition-using-2D-Pose-Input and indexed in Google. Search keywords: N/A Versioning N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Data is already available online, shared by its owners. We will not re-share it. How it will be accessible: Can be downloaded from GitHub at https://github.com/stuarteiffert/RNN-for-Human-Activity-Recognition-using-2D-Pose-Input Methods/software tools to access data: Download the dataset. Repository: GitHub Restrictions on access: Access based on license on https://github.com/stuarteiffert/RNN-for-Human-Activity-Recognition-using-2D-Pose-Input
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: Data will not be altered as it is easy to read. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: The data is already openly shared under a BSD-2 license. Availability for re-use: Data is already available. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will

	be stored on IDF servers provided that this is not forbidden by the owner of the dataset. IDF fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A Is informed consent for data sharing and long term preservation given: Yes.
Other Issues	N/A

5.1.6 20Newsgroups dataset

This dataset will also be used in the context of WP5.

DMP component	AI4Media_Data_25_WP3-5_TEXT_20Newsgroups_v1 Partner: CNR
Data Summary	Purpose: The 20Newsgroups dataset consists of about 20,000 messages posted in the early '90s on 20 different Usenet discussion groups. The dataset is used in AI4media (and has been used elsewhere since the '90s) as a benchmark (training set + test set) for testing text classification systems, e.g., in the context of T3.7 and T5.4. Type/format: Raw text Re-use of existing data: Yes, this dataset has been in the public domain since the '90s. Data origin: The data consist of messages posted in the early '90s on Usenet discussion groups. Expected size: Approximately 20,000 documents. Data utility: It is, and it has been for decades, useful to text classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their text classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by its creator on http://qwone.com/~jason/20Newsgroups/. In every paper we write, we indicate the URL from where the dataset can be downloaded. This URL can be located by just typing the dataset name into any web search engine. Search keywords: No search keywords provided. Versioning: There is only one version which has been made available by its creator. Metadata creation: No metadata were attached to this dataset by its creator, aside from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The creator of the dataset makes it openly accessible at http://qwone.com/~jason/20Newsgroups/. We will not reshare the data. How it will be accessible: The dataset has been accessible from its creator's home page ever since the '90s. Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: http://qwone.com/~jason/20Newsgroups/. Restrictions on access: There are no restrictions on the use of this dataset.
Making data	Interoperability: The dataset consists of interoperable data, for the simple fact that it

interoperable	consists of raw text. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Already publicly available without license at http://qwone.com/~jason/20Newsgroups/. Availability for re-use: Data is already available for re-use. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The creator of the dataset did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.7 HP Amazon reviews dataset

This dataset will also be used in the context of WP5.

DMP component	AI4Media_Data_26_WP3-5_TEXT_HPAmazonReviews_v1 Partner: CNR
Data Summary	Purpose: The HP dataset consists of 27,932 documents. The dataset is used in AI4media as a benchmark (training set of 9,533 documents + test set of 18,399) for testing text quantification methods, e.g., in the context of T3.7 and T5.4. Every document in the dataset is a plain-text tokenized review, with associated a sentiment label: “positive” or “negative”. Type/format: Raw text, with a binary label associated to every document. Re-use of existing data: Yes. The dataset was built by CNR prior to the start of the project. Data origin: The dataset consists of product reviews for the Amazon Kindle e-book reader, collected from public reviews published on the Amazon.com website. Expected size: 25,421 documents. Data utility: Quantification is an emerging research topic in the field of aggregated data analysis. Sentiment quantification on text is of special interest for its usefulness in

	text mining application on social data.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by publishing it on Zenodo (https://doi.org/10.5281/zenodo.4117827). In every paper we write, we cite the resource and the URL from where the dataset can be downloaded. Search keywords: No search keywords provided. Versioning: There is only one version, which has been made available by its creator. Metadata creation: The dataset has no metadata attached.
Making data openly accessible	Data openly accessible: The data is already publicly shared at https://doi.org/10.5281/zenodo.4117827. We will not reshare the data. How it will be accessible: The data are downloadable from the Zenodo website. Methods/software tools to access data: The dataset consists of two text files. A Web browser is required to download them. Repository: The dataset is published on Zenodo with DOI 10.5281/zenodo.4117827. Restrictions on access: There are no restrictions on the access and use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text. Data and metadata vocabularies: No data or metadata vocabularies are used, since content is plain text and no metadata are available. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: The data is already shared on Zenodo under a Creative Commons Attribution 4.0 International license. Availability for re-use: Data are already available for re-use. Usable by third parties after end of project: Yes. Re-use timeframe: Unlimited. Data quality assurance process: The data has not been subjected to a data quality assurance process.
Allocation of resources	Costs for making data FAIR: No Costs for long-term preservation: No
Data security	Security measures: Data is released with CC licence, no security measures for data protection have been implemented.
Ethical aspects	Possible ethical and legal aspects preventing sharing: As the creators of this dataset, CNR did not foresee any ethical or legal issues that can have an impact on data sharing. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.8 JRCAcquis legislative text dataset

This dataset will also be used in the context of WP5.

DMP component	AI4Media_Data_27_WP3-5_TEXT_JRCAcquis_v1 Partner: CNR
Data Summary	Purpose: JRC-Acquis is a collection of legislative texts of European Union law written between the 1950s and 2006, and classified according to a set of 6,000 classes which describe what the text is about; the data are multilingual, i.e., each news story is written in one of 22 official European languages. The dataset is used in AI4media (and has been used elsewhere since the mid 2000's) as a benchmark (training set + test set) for testing multilingual text classification systems, e.g., in the context of T3.7 and T5.4. Type/format: Raw text Re-use of existing data: Yes, this dataset has been in the public domain since the mid 2000's. Data origin: The stories are legislative texts of European Union law. Expected size: About 111,700 documents. Data utility: It is, and it has been for many years, useful to multilingual text classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their multilingual text classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by its creators from the EU website at https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis. In every paper we write, we indicate the URL from where the dataset can be downloaded. This URL can be located by just typing the dataset name into any web search engine. Search keywords: JRC Acquis Versioning: There is only one version (3.0) which has been made available by its creators. Metadata creation: No metadata were attached to this dataset by its creators, aside from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The creators of the dataset make it openly accessible at https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis. We will not reshare the data. How it will be accessible: The dataset has been accessible from https://ec.europa.eu/ from the beginning Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: The data is deposited in the UCI ML repository. Restrictions on access: There are no restrictions on the use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A

Increase data re-use	Licence: Already publicly available dataset. Licesne and terms of use defined at https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis. Availability for re-use: Data is already available for re-use. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The creator of the dataset did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.9 Kindle document dataset

This dataset will also be used in the context of WP5.

DMP component	AI4Media_Data_28_WP3-5_TEXT_Kindle_v1 Partner: CNR
Data Summary	Purpose: The Kindle dataset consists of 25,421 documents. The dataset is used in AI4media as a benchmark (training set of 21,591 documents + test set of 3,821) for testing text quantification methods, e.g., in the context of T3.7 and T5.4. Every document in the dataset is a plain-text tokenized review, with associated a sentiment label: “positive” or “negative”. Type/format: Raw text, with a binary label associated to every document. Re-use of existing data: The dataset was built by CNR prior to the start of the project. Data origin: The dataset consists of product reviews for the Amazon Kindle e-book reader, collected from public reviews published on the Amazon.com website. Expected size: 25,421 documents. Data utility: Quantification is an emerging research topic in the field of aggregated data analysis. Sentiment quantification on text is of special interest for its usefulness in text mining application on social data.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by publishing it on Zenodo (https://doi.org/10.5281/zenodo.4117827). In every paper we write, we cite the resource and the URL from where the dataset can be downloaded. Search keywords: No search keywords provided.

	Versioning: There is only one version, which has been made available by its creator. Metadata creation: The dataset has no metadata attached.
Making data openly accessible	Data openly accessible: Yes, the data is already openly accesible at https://doi.org/10.5281/zenodo.4117827. We will not reshare the data. How it will be accessible: The data are downloadable from the Zenodo website. Methods/software tools to access data: The dataset consists of two text files. A Web browser is required to download them. Repository: The dataset is published on Zenodo with DOI 10.5281/zenodo.4117827. Restrictions on access: There are no restrictions on the access and use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text. Data and metadata vocabularies: No data or metadata vocabularies are used, since content is plain text and no metadata are available. Use of standard vocabularies: No. Mappings to commonly used vocabularies: No.
Increase data re-use	Licence: The data is already openly shared in Zenodo under a Creative Commons Attribution 4.0 International license. Availability for re-use: Data are already available for re-use. Usable by third parties after end of project: Yes. Re-use timeframe: Unlimited. Data quality assurance process: The data has not been subjected to a data quality assurance process.
Allocation of resources	Costs for making data FAIR: No. Costs for long-term preservation: No.
Data security	Security measures: Data is released with CC licence, no security measures for data protection have been implemented.
Ethical aspects	Possible ethical and legal aspects preventing sharing: As the creators of this dataset, CNR did not foresee any ethical or legal issues that can have an impact on data sharing. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.10 OHSUMED MEDLINE document dataset

This dataset will also be used in the context of WP5.

DMP component	**AI4Media_Data_29_WP3-5_TEXT_OHSUMED_v1** **Partner: CNR**
Data Summary	Purpose: The OHSUMED dataset consists of oa set of about 348,000 MEDLINE documents spanning the years from 1987 to 1991, and classified according to a a set of classes representing disease, which describe what the document is about. The dataset

	is used in AI4media (and has been used elsewhere since the mid '90s) as a benchmark (training set + test set) for testing text classification systems, e.g., in the context of T3.7 and T5.4. Type/format: Raw text Re-use of existing data: Yes, this dataset has been in the public domain since the mid '90s. Data origin: The documents consist of title+abstract+metadata from scientific articles available from the MEDLINE service. Ohio State University (OHSU) released this dataset to the public in the mid '90s. Expected size: About 348,000 documents. Data utility: It is, and it has been for decades, useful to text classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their text classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by their creator at https://dmice.ohsu.edu/hersh/ohsumed/index1.html. In every paper we write, we indicate the URL from where the dataset can be downloaded. This URL can be located by just typing the dataset name into any web search engine. Search keywords: No search keywords provided. Versioning: There is only one version which has been made available by its creator. Metadata creation: No metadata were attached to this dataset by its creator, aside from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The creator of the dataset makes it openly accessible at https://dmice.ohsu.edu/hersh/ohsumed/index1.html. We will not reshare the data. How it will be accessible: The dataset has been accessible from its creator's home page ever since the mid '90s. Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: N/A Restrictions on access: There are no restrictions on the use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is already openly shared by the creator under terms of use expalined at https://dmice.ohsu.edu/hersh/ohsumed/index1.html. Availability for re-use: Data is already available for re-use. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A

Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The creator of the dataset did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.11 RCV1 Reuters stories dataset

This dataset will also be used in the context of WP5.

DMP component	**AI4Media_Data_30_WP3-5_TEXT_RCV1-Reuters_v1** **Partner: CNR**
Data Summary	Purpose: The RCV1-v2 dataset consists of about 800,000 news stories written by Reuters journalists, and classified according to a set of 101 classes related to economics, which describe what the news story is about. The dataset is used in AI4media (and has been used elsewhere since the mid 2000's) as a benchmark (training set + test set) for testing text classification systems, e.g., in the context of T3.7 and T5.4. Type/format: Raw text Re-use of existing data: Yes, this dataset has been in the public domain since the mid 2000's. Data origin: The stories were written by Reuters journalists. Reuters released this dataset to the public in the mid 2000's. Expected size: about 800,000 documents. Data utility: It is, and it has been for decades, useful to text classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their text classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by its creator on his home page at http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm, from where it can be obtained by signing an agreement; however, the data in preprocessed (matrix) form can also be simply downloaded, without signing any agreement. In every paper we write, we indicate the URL from where the dataset can be downloaded. This URL can be located by just typing the dataset name into any web search engine. Search keywords: No search keywords provided. Versioning: There is only one version which has been made available by its creator. Metadata creation: No metadata were attached to this dataset by its creator, aside

	from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The creator of the dataset makes it openly accessible at http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm. We will not reshare the data. How it will be accessible: The dataset has been accessible from its creator's home page ever since the mid 2000's. Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: Data available at http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm Restrictions on access: There are no restrictions on the use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is already shared by their creator without a license. However, those who want to aqcuire the data are encouraged to sign an agreement: http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm Availability for re-use: Data is already available for re-use. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The creator of the dataset did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing. The data are made available in preprocessed (matrix) form, a form from which the original text cannot be reconstructed. Reuters personnel have stated that distributing term/document matrices is not a violation of the Agreement. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.12 RCV1RCV2 Reuters stories dataset

This dataset will also be used in the context of WP5.

DMP component	AI4Media_Data_31_WP3-5_TEXT_RCV1RCV2-Reuters_v1 Partner: CNR
Data Summary	Purpose: The RCV1 RCV2 dataset consists of about 111,700 news stories written by Reuters journalists, and classified according to a set of 101 classes related to economics, which describe what the news story is about; the data are multilingual, i.e., each news story is written in one of 5 different languages. The dataset is used in AI4media (and has been used elsewhere since year 2000) as a benchmark (training set + test set) for testing multilingual text classification systems, e.g., in the context of T3.7 and T5.4. Type/format: Raw text Re-use of existing data: Yes, this dataset has been in the public domain since the mid 2000's. Data origin: The stories were written by Reuters journalists. Reuters released this dataset to the public in the mid 2000's. Expected size: about 111,700 documents. Data utility: It is, and it has been for decades, useful to multilingual text classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their multilingual text classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by its creators from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual%2C+Multiview+Text+Categorization+Test+collection. In every paper we write, we indicate the URL from where the dataset can be downloaded. This URL can be located by just typing the dataset name into any web search engine. Search keywords: No search keywords provided. Versioning: There is only one version which has been made available by its creators. Metadata creation: No metadata were attached to this dataset by its creators, aside from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The dataset is already openly accessible at https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual%2C+Multiview+Text+Categorization+Test+collection by its creator. We will not reshare the data. How it will be accessible: The dataset has been accessible from the UCI ML repository ever since 2013. Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: The data is deposited in the UCI ML repository. Restrictions on access: There are no restrictions on the use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text.

	Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is already shared by their creator without a license. Users of the dataset should acknowledge its use, by referring to: *M.-R. Amini, N. Usunier, C. Goutte. Learning from Multiple Partially Observed Views - an Application to Multilingual Text Categorization. Advances in Neural Information Processing Systems 22, p. 28-36, 2009* Availability for re-use: Data is already available for re-use. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The creator of the dataset did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing. The data are made available in preprocessed (matrix) form, a form from which the original text cannot be reconstructed. Reuters personnel have stated that distributing term/document matrices is not a violation of the Agreement. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.13 Reuters-21578 dataset

This dataset will also be used in the context of WP5.

DMP component	AI4Media_Data_32_WP3-5_TEXT_Reuters-21578_v1 Partner: CNR
Data Summary	Purpose: The Reuters-21578 dataset consists of 12,904 news stories written by Reuters journalists in the late ‘90s, and classified according to a a set of 115 classes related to economics (e.g., “acquisitions”, “interest rates”), which describe what the news story is about. The dataset is used in AI4media (and has been used elsewhere since the ‘90s) as a benchmark (training set + test set) for testing text classification systems, e.g., in the context of T3.7 and T5.4. Type/format: Raw text Re-use of existing data: Yes, this dataset has been in the public domain since the ‘90s. Data origin: The stories were written by Reuters journalists. Reuters released this dataset to the public in the mid ‘90s.

	Expected size: 12,904 documents. Data utility: It is, and it has been for decades, useful to text classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their text classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by its creator on his home page and on the UCI machine learning repository at https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection. In every paper we write, we indicate the URL from where the dataset can be downloaded. This URL can be located by just typing the dataset name into any web search engine. Search keywords: No search keywords provided. Versioning: There is only one version which has been made available by its creator. Metadata creation: No metadata were attached to this dataset by its creator, aside from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The dataset is already openly accessible at https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection by its creator. We will not reshare the data. How it will be accessible: The dataset has been accessible from its creator's home page ever since the '90s. Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: The data is deposited in the UCI machine learning repository at https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection Restrictions on access: There are no restrictions on the use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The copyright for the text of newswire articles and Reuters annotations in the Reuters-21578 collection resides with Reuters Ltd. Reuters Ltd. and Carnegie Group, Inc. have agreed to allow the free distribution of this data *for research purposes only*. Users of the dataset should acknowledge its use, refer to the data set by the name "Reuters-21578, Distribution 1.0", and inform of the current location of the dataset. Availability for re-use: Data is already available for re-use. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A

	Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The creator of the dataset did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.14 11 Tweet Sentiment Datasets

This dataset will also be used in the context of WP5.

DMP component	**AI4Media_Data_33_WP3-5_TEXT_11TweetSentiment_v1** **Partner: CNR**
Data Summary	Purpose: The 11 Tweet Sentiment Datasets is a set of 11 datasets (called GASP, HCS, OMD, Sanders, SemEval2013, SemEval2014,, SemEval2015, SemEval2016, SST, WA, WB, respectively), all of a similar nature, often used all together for experimentation purposes, consisting of tweets classified according to the Positive, Neutral, Negative, sentiment-based classes. The datasets are used in AI4media (and have been used elsewhere in the last ten years) as benchmarks (training sets + test sets) for testing sentiment classification systems or sentiment quantification systems, e.g., in the context of T3.7 and T5.4. Type/format: Vectors of features extracted from text Re-use of existing data: Yes, these datasets have been in the public domain for 5 years or more. Data origin: The data consist of posts crawled from Twitter by several authors; they are in the form of feature vectors so as to comply with the Twitter terms of use. Expected size: Approximately 20,000 tweets altogether. Data utility: These datasets are useful to sentiment classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their sentiment classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The datasets are available, among other places, on Zenodo at https://zenodo.org/record/4255764 . In every paper we write, we indicate the URL from where the datasets can be downloaded. This URL can be located by just typing "tweet sentiment classification datasets Zenodo" into any web search engine. Search keywords: No search keywords provided. Versioning: N/A Metadata creation: No metadata are attached to these datasets, aside from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The datasets are openly accessible on Zenodo at https://zenodo.org/record/4255764. We will not reshare the datasets.

	How it will be accessible: The datasets are openly accessible on Zenodo, an open-access repository. Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: The data are deposited in the Zenodo repository at https://zenodo.org/record/4255764 Restrictions on access: There are no restrictions on the use of these datasets.
Making data interoperable	Interoperability: The datasets consist of interoperable data, for the simple fact that they consist of raw text. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is alreday shared on Zenodo under a Creative Commons Attribution 4.0 International license. Availability for re-use: Data are already available for re-use. Usable by third parties after end of project: Without limits. Re-use timeframe: Perpetual Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The authors who made the datasets available did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing, since the tweets that the datasets consist of are made available in the form of feature vectors only, which means that the tweets in their original form cannot be recovered. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.1.15 WipoGamma patent document dataset

This dataset will also be used in the context of WP5.

DMP component	**AI4Media_Data_34_WP3-5_TEXT_WipoGamma_v1** **Partner: CNR**
Data Summary	Purpose: The WipoGamma dataset consists of about 1,100,000 patent documents made available by the World Intellectual Property Organization (WIPO), and classified according to classes representing sectors and subsectors of technology, which describe what the patent document is about. The dataset is used in AI4media (and has been

	used elsewhere) as a benchmark (training set + test set) for testing text classification systems, e.g., in the context of T3.7 and T5.4. Type/format: Raw text Re-use of existing data: Yes, this dataset has been in the public domain for years. Data origin: The dataset consists of patent documents made available by the World Intellectual Property Organization. Expected size: About 1,100,000 documents. Data utility: It is, and it has been for years, useful to text classification researchers (see e.g., T3.7 and T5.4) wishing to benchmark their text classification systems.
Making data findable, incl. provisions for metadata	Is data discoverable: The data have been made available by WIPO on their website, at https://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/. In every paper we write, we indicate the URL from where the dataset can be downloaded. This URL can be located by just typing the dataset name into any web search engine. Search keywords: No search keywords provided. Versioning: There is only one version which has been made available by its creators. Metadata creation: No metadata were attached to this dataset by its creators, aside from the set of classes that is to be used for labelling the documents.
Making data openly accessible	Data openly accessible: The dataset is already openly accessible at https://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/ by WIPO. We will not reshare the data. How it will be accessible: The dataset has been accessible from the WIPO website after filling in a registration form: https://www.wipo.int/classifications/ipc/en/forms/index.html Methods/software tools to access data: The only software tool needed to access the data is a web browser. Repository: https://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/ Restrictions on access: There are no restrictions on the use of this dataset.
Making data interoperable	Interoperability: The dataset consists of interoperable data, for the simple fact that it consists of raw text. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is already openly shared by WIPO under the terms of use defined at https://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/ Availability for re-use: Data is already available for re-use. Usable by third parties after end of project: N/A Re-use timeframe: N/A

	Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CNR servers. CNR fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The creator of the dataset did not, in all evidence, foresee any ethical or legal issues that can have an impact on data sharing. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.2 Datasets used in the context of WP4

5.2.1 ImageNet-ILSVRC2012 image classification dataset

DMP component	**AI4Media_Data_35_WP4_IMAGE_Imagenet_01** **Partner: CERTH**
Data Summary	Purpose: ImageNet is an image dataset organized according to the WordNet hierarchy: images annotated with concept labels. ImageNet is among the most popular large-scale image dataset for image semantic concept classification tasks. In this version of the dataset, we use a subset of ImageNet, the so called Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) dataset. ILSVRC2012 contains approximately 1.3 million images, associated object bounding boxes, and 1,000 semantic concept categories. It will be used by CERTH in T4.3 for evaluating the explainable AI methods developed in this task. Type/format: jpeg Re-use of existing data: Yes Data origin: ImageNet project site: http://www.image-net.org/index Expected size: Train partition: 137 GB; Validation partition: 6.28 GB; Test partition: 12.7 GB. Data utility: It is useful in the context of T4.3 for evaluating the XAI methods developed in this task. In general, this dataset is also useful for any researcher that wants to train deep learning models for image classification/localization using a large-scale image dataset.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted on the ImageNet project site (http://image-net.org/download). It is discoverable by googling “ImageNet Dataset”. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data	Data openly accessible: No. The data is accessible through registration on the

openly accessible	ImageNet project site (http://image-net.org/download) under certain restrictions. The data will not be re-shared by AI4Media partners. How it will be accessible: The data is hosted on ImageNet project site and requires registration to download. The images are provided for non-commercial research and/or educational purposes under certain conditions and terms. Details are provided on: http://image-net.org/download Methods/software tools to access data: Creation of an ImageNet account as described on: http://image-net.org/signup.php?next=download-images Repository: The data repository (data, metadata, documentation, processing code) is hosted on the ImageNet project site. Restrictions on access: ImageNet does not own the copyright of the images. ImageNet provides the images for non-commercial research and/or educational purposes under certain conditions and terms. The users have to sign up for an ImageNet Account.
Making data interoperable	Interoperability: The data is already interoperable and widely used in the research community. Data and metadata vocabularies: Images are in jpeg format. All synsets are assigned to an integer ID between 1 and 1000 (ILSVRC2012_ID). Moreover, its synset has a WordNet ID (WNID) used to uniquely identify a synset in ImageNet or WordNet Use of standard vocabularies: N/A Mappings to commonly used vocabularies: A mapping to WordNet is already provided.
Increase data re-use	Licence: A registration is required to download the dataset. ImageNet does not own the copyright of the images. However, it provides the images for non-commercial research and/or educational purposes under certain conditions and terms. Details are provided on: http://image-net.org/download Availability for re-use: N/A Usable by third parties after end of project: Data already shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate most of the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: The images of the dataset are provided for non-commercial research and/or educational purposes under certain conditions and terms. Details are provided on: http://image-net.org/download Is informed consent for data sharing and long term preservation given: N/A (Note that ImageNet does not own the copyright of the images)

Other Issues	N/A

5.2.2 FFHQ dataset for GAN training

DMP component	AI4Media_Data_36_WP4_IMAGE_FFHQ_v1 Partner: CEA
Data Summary	Purpose: Flickr-Faces-HQ (FFHQ) is a high-quality image dataset of human faces, originally created as a benchmark for generative adversarial networks (GAN). We will use the dataset stored as multi-resolution TF records. It will be used in T4.3 to evaluate the generative models developed in the task. Results involving this dataset will be reported in D4.1, D4.4 and D4.6. Type/format: JSON file containing metadata and 70k images stored as multi-resolution TF records. Re-use of existing data: Yes, were reusing an existing dataset Data origin: https://reposhub.com/python/deep-learning/NVlabs-ffhq-dataset.html Expected size: 575 MB Data utility: It is useful to WP4 partners to benchmark generative models.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable. The dataset is hosted on the NVIDIA website (actually Google drive of NVIDIA). Search keywords: FFHQ, Flickr-Faces-HQ Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible at https://reposhub.com/python/deep-learning/NVlabs-ffhq-dataset.html. We will not reshare the data. How it will be accessible: Already shared through a third-party repository link Methods/software tools to access data: python scripts to download data are provided Repository: Network Repository (http://networkrepository.com) Restrictions on access: None
Making data interoperable	Interoperability: The data is interoperable. Data and metadata vocabularies: The metadata schema can be found at https://reposhub.com/python/deep-learning/NVlabs-ffhq-dataset.html#articleHeader4 Use of standard vocabularies N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under an attribution licence (https://reposhub.com/python/deep-learning/NVlabs-ffhq-dataset.html) Availability for re-use: The loading and pre-processing mechanism developed for using the dataset in experiments will be made publicly available to ensure reproducibility of research.

	Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CEA's servers. CEA fully complies with the applicable national, European and International framework, and the GDPR. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The dataset was made available from FlickR images with appropriate licence. It will be kept only during the length of the T4.3 and will not be re-shared since it is available on NVIDIA website. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.2.3 MNIST image dataset

DMP component	AI4Media_Data_37_WP4_IMAGE_MNIST_v1 Partner: IBM
Data Summary	Purpose: The MNIST database of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. Type/format: Images Re-use of existing data: Yes, were reusing an existing dataset Data origin: https://tensorflow.google.cn/datasets/catalog/mnist?hl=en Expected size: 21 MB Data utility: It is useful to WP4 partners to benchmark generative AI models.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable. The dataset is hosted on the Google Tensorflow website at https://tensorflow.google.cn/datasets/catalog/mnist?hl=en Search keywords: MNIST Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible at https://tensorflow.google.cn/datasets/catalog/mnist?hl=en , http://yann.lecun.com/exdb/mnist/ and https://keras.io/api/datasets/mnist/ How it will be accessible: Already shared through a third-party repository link Methods/software tools to access data: python scripts to download data are provided Repository: N/A Restrictions on access: None

Making data interoperable	Interoperability: The data simply consists of images so it is interoperable with any program that can read images Data and metadata vocabularies: N/A Use of standard vocabularies N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is already publicly available under an attribution licence Creative Commons Attribution-Share Alike 3.0 license. Availability for re-use: The loading and pre-processing mechanism developed for using the dataset is available to the public through the tensorflow library. Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: N/A (The dataset is already publicly available)
Ethical aspects	Possible ethical and legal aspects preventing sharing: No. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.2.4 Interestingness10k image +video dataset

DMP component	AI4Media_Data_38_WP4_IMAGE_VIDEO_Interestingness10k Partner: UPB
Data Summary	Purpose: Interestingness10k is the most comprehensive collection of image and video information annotated for training and evaluating algorithms for visual interestingness prediction. It is used for T4.6 'Benchmarking of AI Systems' and for T6.6 'Measuring and Predicting User Perception of Social Media' and for T6.3 'Hybrid, privacy-enhanced recommendation'. Results involving this data are to be reported to the task corresponding deliverables. Type/format: 9,831 images, 4 hours of video, interestigness scores determined based on more than 1M pair-wise annotations of 800 trusted annotators, some pre-computed multi-modal descriptors, and 192 system output results as baselines. Re-use of existing data: No, it was newly generated. Part of the data usage and algorithms analysis was carried out within the project. The data was generated outside the project. Data origin: IDF dataset: https://www.interdigital.com/data_sets/interestingness-dataset Expected size: 20 GB Data utility: It is useful to WP4 and WP6 partners.
Making data findable, incl.	Is data discoverable: Data is discoverable. The dataset is hosted on the Interdigital

provisions for metadata	website. Search keywords: Interestingness10k, Predicting Media Interestingness. Versioning: N/A Metadata creation: The data comes with metadata.
Making data openly accessible	Data openly accessible: The data is already openly accessible at https://www.interdigital.com/data_sets/interestingness-dataset. We will not reshare the data. How it will be accessible: Already shared through a third-party repository link. Methods/software tools to access data: Confirmation of data usage agreement via email is required. Repository: Interdigital. Restrictions on access: Access is made via email request.
Making data interoperable	Interoperability: The data is interoperable. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under a Creative Commons license that allows redistribution. Availability for re-use: All the details of the data are available on the website. A detailed article describing the data, usage and many baseline systems is also available *M.G. Constantin, L.-D. Ștefan, B. Ionescu, Q.-K.-N. Duong, C.-H. Demarty, M. Sjoberg "Visual Interestingness Prediction: A Benchmark Framework and Literature Review", International Journal of Computer Vision* Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UPB servers. UPB fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: We didn’t identify any as materials are already publicly available Creative Commons content. The annotations are not recording any personal information of the user. Is informed consent for data sharing and long term preservation given: N/A.
Other Issues	N/A

5.2.5 MediaEval Memorability 2020 dataset

DMP component	**AI4Media_Data_39_WP4_VIDEO_Memorability2020** **Partner: UPB**
Data Summary	Purpose: The Predicting Media Memorability 2020 dataset is a collection of videos annotated for their long- and short-term memorability impact on users. All materials are under Creative Commons licenses that allow redistribution. It is used for T4.6 'Benchmarking of AI Systems' and for T6.6 'Measuring and Predicting User Perception of Social Media' and for T6.3 'Hybrid, privacy-enhanced recommendation'. Results involving this data are to be reported to the task corresponding deliverables. Type/format: 1,500 short videos, long- (72 hours) and short-(24 hours) term memorability scores, some pre-computed multi-modal descriptors. Re-use of existing data: The video data is recovered from the TRECVid 2019 Video-to-Text dataset. The annotations were newly created. Part of the data usage and algorithms analysis was carried out within the project. The data was generated outside the project. Data origin: Videos recovered from the TRECVid 2019 Video-to-Text dataset. Expected size: 1.8 GB Data utility: It is useful to WP4 and WP6 partners.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable. Search keywords: Predicting Media Memorability. Versioning: N/A Metadata creation: The data comes with metadata.
Making data openly accessible	Data openly accessible: The data is open for distribution. It has not yet been published publicly. How it will be accessible: The data can be obtained via request to owners. Methods/software tools to access data: N/A. Repository: N/A Restrictions on access: The data is provided by the authors.
Making data interoperable	Interoperability: The data is interoperable. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under a Creative Commons license that allows redistribution. Availability for re-use: All the details of the data are available on the https://multimediaeval.github.io/editions/2020/tasks/memorability/ website. A detailed article describing the data, usage and baseline systems is also available *Alba García Seco De Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-Hélène Demarty, Faiyaz Doctor, Bogdan Ionescu and Alan F. Smeaton, Overview of MediaEval 2020 Predicting Media Memorability task: What*

	Makes a Video Memorable? MediaEval Workshop 2021. Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UPB servers. UPB fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: We didn't identify any as materials are already publicly available Creative Commons content. The annotations are not recording any user information. Is informed consent for data sharing and long term preservation given: N/A.
Other Issues	N/A

5.2.6 ImageCLEF DrawnUI 2021 dataset

DMP component	**AI4Media_Data_40_WP4_IMAGE_drawnUI2021** **Partner: UPB**
Data Summary	Purpose: The ImageCLEF drawnUI 2021 dataset contains hand drawn images of website user interface units and real website screenshots which are annotated for their user interface components. It serves for training systems capable of automatically identifying a website template from a drawing or a image of it. It is used for T4.6 'Benchmarking of AI Systems'. Results involving this data are to be reported to the task corresponding deliverables. Type/format: 4,291 hand drawn images and 9,630 screenshot images, manual labelling of the positions of UI bounding boxes. Re-use of existing data: The data and annotations were newly created. Part of the data usage and algorithms analysis was carried out within the project. The data was generated outside the project. Data origin: Images from ImageCLEF drawnUI 2021 dataset. Expected size: 9 GB Data utility: It is useful to WP4 partners.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable. Search keywords: ImageCLEFdrawnUI. Versioning: N/A Metadata creation: The data comes with metadata.
Making data openly accessible	Data openly accessible: The data is not open yet. The data set is so far owned by teleportHQ. We plan to release it publicly.

	How it will be accessible: The data can be obtained via request to authors. Methods/software tools to access data: N/A. Repository: N/A Restrictions on access: The data is provided by the authors.
Making data interoperable	Interoperability: The data is interoperable. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is owned by company teleportHQ, Romania. Availability for re-use: All the details of the data are available on the https://www.imageclef.org/2021/drawnui website. Usable by third parties after end of project: The data set is so far owned by teleportHQ. We plan to release it publicly. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UPB servers. UPB fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: We didn't identify any as materials are either generated by the authors or recovered from public sources. The annotations are not recording any user information. Is informed consent for data sharing and long term preservation given: N/A.
Other Issues	N/A

5.3 Datasets used in the context of WP5

5.3.1 SumMe video summarization dataset

DMP component	**AI4Media_Data_41_WP5_VIDEO_SumMeGycli14_v1** **Partner: CERTH**
Data Summary	Purpose: This dataset is composed of a set of videos from various genres (e.g. holidays, sports) and the associated ground-truth data that indicate the preferences of multiple human annotators with respect to the optimal visual summary for each video. It will be used in T5.1 for training and evaluating purposes, assisting the development of deep-learning-based architectures for video summarization. Type/format: Video files in MP4 and WEBM format; MAT files with the ground-truth annotations; TXT and XLS files with information about the statistics of each video category; Matlab and Python scripts for evaluating the performance of a video

	summarization algorithm. Re-use of existing data: Yes. Data origin: This dataset was introduced in an ECCV 2014 paper titled “Creating Summaries from User Videos”, and was made publicly available through the following link: https://gyglim.github.io/me/vsum/index.html#benchmark Expected size: ~2.5GB Data utility: It will be useful to WP5 partners working on the development of video summarization methods, for training and evaluation purposes.
Making data findable, incl. provisions for metadata	Is data discoverable: Data are already publicly available at: https://gyglim.github.io/me/vsum/index.html#benchmark Search keywords: N/A Versioning: N/A Metadata creation: Any metadata generated to assist training and evaluation of video summarization methods (e.g. deep feature vectors representing the visual content of video frames, or data about the shot-level structure of the videos) will be stored in HDF5 files; a documentation of these metadata will be also created to facilitate their re-use.
Making data openly accessible	Data openly accessible: The data are already openly accessible at: https://gyglim.github.io/me/vsum/index.html#benchmark How it will be accessible: The original data are available through a third-party repository link. Any created metadata and their associated documentation will be made publicly available through a GitHub repository. Methods/software tools to access data: Web-browser to download the data as zip file. Repository: https://gyglim.github.io/me/vsum/index.html#benchmark Any generated metadata will be deposited at a publicly accessible GitHub repository. Restrictions on access: None
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: Any metadata generated to assist training and evaluation of video summarization methods (e.g. deep feature vectors representing the visual content of video frames, or data about the shot-level structure of the videos) will be stored in HDF5 files; a documentation of these metadata will be also created to facilitate their re-use. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under an attribution non-commercial licence (https://gyglim.github.io/me/vsum/index.html#benchmark) Availability for re-use: The data are already publicly-available for re-use. Any generated metadata will be made permanently publicly-available for re-use as soon as they are complete and appropriately documented. Usable by third parties after end of project: The dataset is already available for use by third parties.

	Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate most of the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the original dataset and the associated ECCV2014 paper should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.2 TVSum video summarization dataset

DMP component	AI4Media_Data_42_WP5_VIDEO_TVSumSong15_v1 Partner: CERTH
Data Summary	Purpose: This dataset is composed of a set of videos from 10 categories of the TRECVid MED dataset - including news, how-to's, user-generated-content and documentaries - and the associated ground-truth data that indicate the opinion of multiple human annotators with respect to the importance of video frames and fragments. It will be used in T5.1 for training and evaluating purposes, assisting the development of deep-learning-based architectures for video summarization. Type/format: Video files in MP4 format; video thumbnails in JPG format; MAT and TSV files with the ground-truth annotations and video-level metadata; Matlab scripts for evaluating the performance of a video summarization algorithm. Re-use of existing data: Yes Data origin: This dataset was introduced in a CVPR 2015 paper titled “TVSum: Summarizing web videos using titles”, and was made publicly available through the following links: https://github.com/yalesong/tvsum and http://people.csail.mit.edu/yalesong/tvsum/ Expected size: ~650MB Data utility: It will be useful to WP5 partners working on the development of video summarization methods, for training and evaluation purposes.
Making data findable, incl. provisions for metadata	Is data discoverable: Data are already publicly-available at: http://people.csail.mit.edu/yalesong/tvsum/ Search keywords: N/A Versioning: N/A Metadata creation: Any metadata generated to assist training and evaluation of video summarization methods (e.g. deep feature vectors representing the visual content of

	video frames, or data about the shot-level structure of the videos) will be stored in HDF5 files; a documentation of these metadata will be also created to facilitate their re-use.
Making data openly accessible	Data openly accessible: The data are already openly accessible at: http://people.csail.mit.edu/yalesong/tvsum/ How it will be accessible: The original data are shared through a third-party repository link. Any created metadata and their associated documentation will be made publicly-available through a GitHub repo. Methods/software tools to access data: Web-browser to download the data as tgz file. Repository: Original data are already deposited at: http://people.csail.mit.edu/yalesong/tvsum/. Any generated metadata will be deposited at a publicly-accessible GitHub repo. Restrictions on access: None
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: Any metadata generated to assist training and evaluation of video summarization methods (e.g. deep feature vectors representing the visual content of video frames, or data about the shot-level structure of the videos) will be stored in HDF5 files; a documentation of these metadata will be also created to facilitate their re-use. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The videos of this dataset, collected from YouTube, come with a Creative Commons CC-BY (v3.0) license. (https://github.com/yalesong/tvsum#overview) Availability for re-use: The data are already publicly-available for re-use. Any generated metadata will be made permanently publicly-available for re-use as soon as they are complete and appropriately documented. Usable by third parties after end of project: The dataset is already available for use by third parties. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate most of the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the original dataset and the associated CVPR2015 paper should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A

Other Issues	N/A

5.3.3 RAI Monuments of Italy dataset

DMP component	**AI4Media_Data_43_WP5_VIDEO_MonumentsOfItaly_v1** **Partner: RAI**
Data Summary	Purpose: A collection of videos depicting various Italian monuments. This dataset will be useful as a reference for developments in WP5 about landmark recognition. The typical processing will include indexing of images and storage of resulting features in a database for search/match. Type/format: Videos in MP4 or WMV format Re-use of existing data: All videos are coming from RAI Archives. Data origin: RAI Archives, mainly news production. Expected size: 4 GB Data utility: The dataset can be used to test landmark recognition algorithms.
Making data findable, incl. provisions for metadata	Is data discoverable: No. Search keywords: N/A Versioning: N/A Metadata creation: Videos are organised in folders with self-descriptive names.
Making data openly accessible	Data openly accessible: No, the dataset will not be made openly accessible. It is stored in private repository and will only be accessed by partners based on bilateral agreements with RAI. How it will be accessible: Bilateral agreement with RAI. Methods/software tools to access data: File transfer. Repository: Private RAI repository. Restrictions on access: Direct access is not allowed.
Making data interoperable	Interoperability: Video encoding schemes used in the datasets are widely accepted by all common software tools. Data and metadata vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: RAI specific license. Availability for re-use: N/A Usable by third parties after end of project: To be defined. Re-use timeframe: To be defined. Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A

Data security	<u>Security measures</u>: Data will be stored on local repositories inside RAI's premises, subject to the same security measures already used for IT infrastructure in RAI. These include network isolation from external internet accesses, firewalling, account-based access control management to the storage where the data copies are located. RAI fully complies with the applicable national, European data security frameworks, and the GDPR.
Ethical aspects	<u>Possible ethical and legal aspects preventing sharing</u>: Possible rights issues related to material included in news items for which license of usage has expired. <u>Is informed consent for data sharing and long term preservation given</u>: No
Other Issues	No

5.3.4 LVIS image dataset

DMP component	**AI4Media_Data_44_WP5_IMAGE_LVIS_v1** **Partner: JRC**
Data Summary	<u>Purpose</u>: LVIS is a dataset for large vocabulary instance segmentation, with > 1,200 categories, a large number of rare categories (long-tail), and high-quality instance segmentation mask. The dataset will be used by JR within Task 5.3 (Learning with Scarce Data), specifically for training the few-shot object detection algorithms, which will be researched and developed within this task. <u>Type/format</u>: Images, segmentation mask + json annotations <u>Re-use of existing data</u>: Yes, we are reusing an existing dataset <u>Data origin</u>: https://www.lvisdataset.org/dataset <u>Expected size</u>: 30 GB <u>Date utility</u>: It is useful for various tasks, but especially for T5.3 (learning from scarce data) for the few-shot instance segmentation.
Making data findable, incl. provisions for metadata	<u>Is data discoverable</u>: Data is discoverable. The dataset is hosted in the Network Repository https://www.lvisdataset.org/dataset <u>Search keywords</u>: N/A <u>Versioning</u>: N/A <u>Metadata creation</u>: N/A
Making data openly accessible	<u>Data openly accessible</u>: The data is already openly accessible at https://www.lvisdataset.org/dataset by its owners thus, we will not re-share it. <u>How it will be accessible</u>: Shared through a third-party repository link <u>Methods/software tools to access data</u>: Web-browser to download the data as zip file <u>Repository</u>: Network Repository (https://www.lvisdataset.org/dataset) <u>Restrictions on access</u>: None
Making data interoperable	<u>Interoperability</u>: N/A <u>Data and metadata vocabularies</u>: LVIS has annotations for instance segmentations in a commonly used format similar to MS COCO. The annotations are stored using JSON. An API is provided to access and manipulate annotations.

	Use of standard vocabularies N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The LVIS annotations along with this website are licensed under a Creative Commons Attribution 4.0 License. All LVIS dataset images come from the COCO dataset; please see https://cocodataset.org/#termsofuse for their terms of use. Availability for re-use: Yes. Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on JR's servers. JR fully complies with the applicable national and European data protection frameworks. State-of-the-art IT security measures and company policies (firewalls, right-based file system, etc.) mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the dataset should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.5 CIFAR10/100 image dataset

DMP component	**AI4Media_Data_45_WP5_IMAGE_CIFAR_v1** **Partner: QMUL**
Data Summary	Purpose: The CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. CIFAR-10 consists of 60,000 labeled images, split between 10 including objects and animals. CIFAR-100 consists of 60,000 labeled images split between 100 “fine” classes and 20 “coarse” superclasses, including classes related to people. The CIFAR-10 and CIFAR-100 datasets are among the most popular image datasets for classification tasks. The datasets will be used to by QMUL toevaluate novel representation learning techniques developed for T5.3. Type/format: Numpy array Re-use of existing data: Yes Data origin: https://www.cs.toronto.edu/~kriz/cifar.html Expected size: ~160MB for each of the two datasets Data utility: The datasets will be used for WP5 T5.3 to evaluate the developed methods and contrast their performance with existing works.
Making data findable, incl. provisions for metadata	Is data discoverable: The dataset is discoverable from its website: https://www.cs.toronto.edu/~kriz/cifar.html

	Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly available at https://www.cs.toronto.edu/~kriz/cifar.html. We will not re-share it. How it will be accessible: From original source. Methods/software tools to access data: No specialized software is required to access the data. They can be downloaded directly from the source Repository: https://www.cs.toronto.edu/~kriz/cifar.html Restrictions on access: None
Making data interoperable	Interoperability: The data are interoperable. Data and metadata vocabularies: The data are provided in numpy format. The labels are also provided as numpy files, where each sample's label is represented by an integer (0-9 for CIFAR-10, 0-99 for CIFAR-100). The dataset also includes a list that maps semantic labels to their numerical representation (e.g. label 0 in CIFAR-10 corresponds to the class "airplane". Use of standard vocabularies: The classes of the dataset relate to real world objects and entities and use common words as labels to identify them. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Already publicly shared. Availability for re-use: The data are publicly available with no restrictions as to who can acquire it. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the data are stored in the servers of Queen Mary University of London. Access requires username/password authentication. The security measures taken prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.6 STL-10 image dataset

DMP component	AI4Media_Data_46_WP5_IMAGE_STL10_v1 Partner: QMUL
Data	Purpose: STL-10 is a CIFAR-10 inspired dataset whose samples are drawn from the

Summary	ImageNet dataset. It consists of 13,000 labeled samples that are equally split between 10 labels, the same as those of ImageNet. 5,000 of those samples belong to the training set and 8,000 to the test set. Furthermore, the dataset includes 100,000 unlabeled samples. Notably The datasets will be used by QMULto evaluate novel representation learning techniques developed for T5.3. Type/format: Binary files Re-use of existing data: Yes Data origin: https://cs.stanford.edu/~acoates/stl10/ Expected size: 2.5 GB Data utility: The datasets will be used for WP5 T5.3 to evaluate the developed methods and contrast their performance with existing works.
Making data findable, incl. provisions for metadata	Is data discoverable: The dataset is discoverable from its website: https://www.cs.toronto.edu/~kriz/cifar.html. No registration is required and no restrictions are present as to who can access the data Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly available at https://cs.stanford.edu/~acoates/stl10/. We will not re-share it. How it will be accessible: From original source. Methods/software tools to access data: No specialized software is required to access the data. They can be downloaded directly from the source. They can be read via python code that the source provides. Repository: https://cs.stanford.edu/~acoates/stl10/ Restrictions on access: None
Making data interoperable	Interoperability: The data are interoperable. Data and metadata vocabularies: The data are provided in binary format. The labels are also provided as numpy files, where each sample's label is represented by an integer (range 1-10). The dataset also includes a txt file that maps semantic labels to their numerical representation. Finally, a separate txt file proposes specific data splits to be used for multi-fold validation purposes Use of standard vocabularies: The classes of the dataset relate to real world objects and entities and use common words as labels to identify them. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Already publicly shared. Availability for re-use: The data are publicly available with no restrictions as to who can acquire it. Usable by third parties after end of project: N/A Re-use timeframe: N/A

	Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the data are stored in the servers of Queen Mary University of London. Access requires username/password authentication. The security measures taken prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.7 CCNet text dataset for multilingual representation learning

DMP component	**AI4Media_Data_47_WP5_TEXT_CCNET_v1** **Partner: CEA**
Data Summary	Purpose: CCNet is a large text dataset composed of high-quality monolingual datasets from Common Crawl for a variety of languages. CEA will use this dataset to create languages models in different languages. These models will be used to train information extraction models (e.g. opinion mining or named entity recognition). Results involving this dataset will be reported in D5.4. Type/format: compressed JSON files (one per line) Re-use of existing data: Yes, we are reusing an existing dataset Data origin: https://github.com/facebookresearch/cc_net Expected size: Unknown Data utility: Useful for training language models. Will be used in WP5 for developing information extraction models in different languages.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable on the orginal source website. Search keywords: CCNET text dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible at https://github.com/facebookresearch/cc_net. The data will not be re-shared. How it will be accessible: Shared through a third-party repository link Methods/software tools to access data: python scripts to download data are provided. Repository: Network Repository (http://networkrepository.com) Restrictions on access: None
Making data interoperable	Interoperability: The data is shared as JSON files. Data and metadata vocabularies: The JSON files follow the following format (see

	https://github.com/facebookresearch/cc_net): • url: webpage URL (part of CC) • date_download: date of download (part of CC) • digest: sha1 digest of the webpage (part of CC) • length: number of chars • nlines: number of lines • source_domain: web domain of the webpage • title: page title (part of CC) • raw_content: webpage content after deduplication • original_nlines: number of lines before deduplication • original_length: number of chars before deduplication • language: language detected by FastText LID • language_score: language score • perplexity: perplexity of a LM trained on Wikipedia Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is shared through an MIT license. Availability for re-use: The loading and pre-processing mechanism developed for using the dataset in experiments will be made publicly available to ensure reproducibility of research. Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on CEA's servers. CEA fully complies with the applicable national, European and International framework, and the GDPR. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A The dataset is alredy open. We will not reshare it. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.8 360 Video Viewing Dataset in Head Mounted Virtual Reality

DMP component	AI4Media_Data_48_WP5_VIDEO_360VideoViewingHeadMountedVR_v1 Partner: UCA
Data Summary	Purpose: The 360° Video Viewing Dataset in Head-Mounted Virtual Reality consists of the content (10 equirectangular videos) and sensory data (50 subjects) of 360-degree videos to Head Mounted Display (HMD). The dataset contains both content data (such as image saliency maps and motion maps derived from 360° videos) and sensor data (such as viewer head positions and orientations derived from HMD sensors). The content and sensor data are aligned using timestamps in the log files. The dataset was used by their creators to optimize 360° video streaming applications by the prediction of fixation points in 360° Videos. The dataset will be used to develop and test new algorithms for the prediction of head motion in 360° videos in WP5. Type/format: videos are compressed using H-264 in MP4 container, while sensor (text) data is stored as comma separated values files in ASCII Re-use of existing data: Yes, we are re-using an existing dataset from Lo, W. C., Fan, C. L., Lee, J., Huang, C. Y., Chen, K. T., & Hsu, C. H. (2017, June). 360 video viewing dataset in head-mounted virtual reality. In Proceedings of the 8th ACM on Multimedia Systems Conference (pp. 211-216). Data origin: Lo, W. C., Fan, C. L., Lee, J., Huang, C. Y., Chen, K. T., & Hsu, C. H. (2017, June). 360 video viewing dataset in head-mounted virtual reality. In Proceedings of the 8th ACM on Multimedia Systems Conference (pp. 211-216). https://nmsl.cs.nthu.edu.tw/360-video-project/ Expected size: ~1GB Data utility: This dataset is useful in our context to train deep neural networks to predict which parts of 360° videos attract viewers to watch the most. This dataset however can be leveraged in various novel applications in a much broader scope. For example, it can be used by researchers, engineers, and hobbyists to either optimize existing 360° video streaming applications (rate-distortion optimization, saliency prediction) and novel applications (crowd-driven camera movements, head motion prediction).
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is under the Networking and Multimedia Systems Lab of the Department of Computer Science at National Tsing Hua University (NTHU). Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on the website of the National Tsing Hua University https://nmsl.cs.nthu.edu.tw/360-video-project/. How it will be accessible: The data is accessible directly by following a Google Drive link: https://drive.google.com/file/d/1s_EEUjUTa_N5u94Nuwir9gl3pwDEY_7_/view Methods/software tools to access data: The data can be directly downloaded.

	Repository: Web server of the National Tsing Hua University. Restrictions on access: No
Making data interoperable	Interoperability: The data is already interoperable. Data and metadata vocabularies: Videos for saliency maps and motion maps are in mp4 format and information about sensory data (head orientation or viewed tiles) are text files in csv format. There is also a readme file in txt format explaining the different folders in the dataset. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: To use the dataset, you should cite the work of *Lo, W. C., Fan, C. L., Lee, J., Huang, C. Y., Chen, K. T., & Hsu, C. H. (2017, June). 360 video viewing dataset in head-mounted virtual reality. In Proceedings of the 8th ACM on Multimedia Systems Conference (pp. 211-216).* Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UCA servers. UCA fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A. Is informed consent for data sharing and long term preservation given: N/A (Note that actors in the dataset provided their consent for the creation of these traces and the users cannot be identified, and therefore no personal data is used).
Other Issues	N/A

5.3.9 Predicting Head Movement in Panoramic Video Dataset

DMP component	**AI4Media_Data_49_WP5_VIDEO_HeadMovementinPanoramicVideo_v1** **Partner: UCA**
Data Summary	Purpose: The dataset for the Head Movement Prediction in Panoramic Video contains both head movement and eye movement data of 56 subjects on 76 panoramic videos. Of 360-degree videos to Head Mounted Display (HMD). The dataset contains both content data (the 360° videos) and sensor data (the head positions of subjects exploring each video). The sensor data is stored in a Matlab file, this file includes 76 cells, corresponding to the head motion data of all 76 videos. Each cell records (longitude, latitude) of head motion pairs for 58 subjects in the colums of the Matlab file.

	The dataset will be used in WP5 to develop and test new algorithms for the prediction of head motion in 360° videos. Type/format: videos are compressed using H-264 in MP4 container, while sensor data is stored as a MATLAB file. Re-use of existing data: Yes, we are re-using an existing dataset from Xu, M., Song, Y., Wang, J., Qiao, M., Huo, L., & Wang, Z. (2018). Predicting head movement in panoramic video: A deep reinforcement learning approach. IEEE transactions on pattern analysis and machine intelligence, 41(11), 2693-2708. Data origin: M., Song, Y., Wang, J., Qiao, M., Huo, L., & Wang, Z. (2018). Predicting head movement in panoramic video: A deep reinforcement learning approach. IEEE transactions on pattern analysis and machine intelligence, 41(11), 2693-2708. Expected size: ~4GB Data utility: This dataset is useful in our context to train deep neural networks to predict which parts of 360° videos attract viewers to watch the most. This dataset however can be leveraged in various novel applications in a much broader scope. For example, it can be used by researchers, engineers, and hobbyists to either optimize existing 360° video streaming applications (rate-distortion optimization, saliency prediction) and novel applications (crowd-driven camera movements, head-eye relationship in video exploration).
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted in the github repository of the authors https://github.com/YuhangSong/DHP. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on the repository of the authors https://github.com/YuhangSong/DHP. How it will be accessible: The data is accessible directly by following a Dropbox link: https://www.dropbox.com/s/swenk8b33vs6151/PVS-HM.tar.gz?dl=0 Methods/software tools to access data: The data can be directly downloaded. Repository: Github Restrictions on access: There is no restriction to access the dataset, it is hosted in the repository of the authors at https://github.com/YuhangSong/DHP, with a direct link to download in Dropbox https://www.dropbox.com/s/swenk8b33vs6151/PVS-HM.tar.gz?dl=0. However, the authors recommend contacting them so that they can grant permission to the file.
Making data interoperable	Interoperability: The data is already interoperable. Data and metadata vocabularies: Equirectangular videos are in mp4 format and information about sensory data (head orientation) is stored as a MATLAB file with an entry per video, each with a matrix with a column per user and alternating longitude and latitudes in the rows. Use of standard vocabularies: N/A

	Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: To use the dataset, you should cite the work of M., Song, Y., Wang, J., Qiao, M., Huo, L., & Wang, Z. (2018). Predicting head movement in panoramic video: A deep reinforcement learning approach. IEEE transactions on pattern analysis and machine intelligence, 41(11), 2693-2708. Availability for re-use: N/A Usable by third parties after end of project: The authors recommend contacting them by mail to ask for permission to access the file, before a password was needed to download the dataset, but now the data is publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UCA servers. UCA fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A. Is informed consent for data sharing and long term preservation given: N/A (Note that actors in the dataset provided their consent for the creation of these traces and the users cannot be identified, and therefore no personal data is used.)
Other Issues	N/A

5.3.10 Gaze prediction in Dynamic 360° Immersive Videos Dataset

DMP component	**AI4Media_Data_50_WP5_VIDEO_GazePredictionDynamic360ImmersiveVideos_v1** **Partner: UCA**
Data Summary	Purpose: The dataset for Gaze Prediction in VR Videos contains 208 videos captured in dynamic scenes, and the traces of head and gaze position of 45 subjects, the traces of at least 31 subjects are recorded per video. The sensor data is stored in a csv files with a folder per user and a text file per trace. The dataset will be used in WP5 to develop and test new algorithms for the prediction of head motion in 360° videos. Type/format: videos are compressed using H-264 in MP4 container, while sensor data is stored as text files. Re-use of existing data: Yes, we are re-using an existing dataset from Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., & Gao, S. (2018). Gaze prediction in dynamic 360 immersive videos. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5333-5342). Data origin: Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., & Gao, S. (2018). Gaze prediction in dynamic 360 immersive videos. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5333-5342).

	https://github.com/xuyanyu-shh/VR-EyeTracking. Expected size: ~4GB Data utility: This dataset is useful in our context to train deep neural networks to predict which parts of 360° videos attract viewers to watch the most. This dataset however can be leveraged in various novel applications in a much broader scope. For example, it can be used by researchers, engineers, and hobbyists to either optimize existing 360° video streaming applications (rate-distortion optimization, saliency prediction) and novel applications (crowd-driven camera movements, head-eye relationship in video exploration).
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted in the github repository of the authors https://github.com/xuyanyu-shh/VR-EyeTracking. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on the repository of the authors https://github.com/xuyanyu-shh/VR-EyeTracking. How it will be accessible: The data is accessible by following a Baidu link: https://pan.baidu.com/s/1RKTZoeiLjKidW3S1E0l_yw with the code “olxt” given in the github repository. Methods/software tools to access data: The data can be downloaded from the baidu website using the baidu-pan-downloader https://github.com/dotennin/baidu-pan-downloader. Repository: Github Restrictions on access: There is a password to access the Baidu link, the password “olxt” can be found in the github repository.
Making data interoperable	Interoperability: The data is already interoperable. Data and metadata vocabularies: Equirectangular videos are in mp4 format and information about sensory data (head and gaze position) is stored as text files with a folder per user and a text file per video. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: To use the dataset, you should cite the work of Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., & Gao, S. (2018). Gaze prediction in dynamic 360 immersive videos. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5333-5342). Availability for re-use: N/A Usable by third parties after end of project: Data is already publicly available. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A

	Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UCA servers. UCA fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A. Is informed consent for data sharing and long term preservation given: N/A (Note that actors in the dataset provided their consent for the creation of these traces and the users cannot be identified, and therefore no personal data is used.)
Other Issues	N/A

5.3.11 Your Attention is Unique Dataset

DMP component	**AI4Media_Data_51_WP5_VIDEO_YourAttentionIsUnique_v1** **Partner: UCA**
Data Summary	Purpose: The dataset for the PanoSalnet model consists on the post-processing of a publicly available dataset. The dataset includes 18 videos viewed by 48 users, from which 9 videos are selected. The dataset also includes the model weights of a Python Caffe implementation. The sensor data and the saliency data are stored in a Python dictionary with an entry per video, each with a list with an entry per user. This dataset is useful in our context to train deep neural networks to predict which parts of 360° videos attract viewers to watch the most (WP5). Type/format: All the data is contained within a python dictionary, numpy arrays are used to store the values of the saliency maps and the head motion traces. Re-use of existing data: Yes, we are re-using an existing dataset from Nguyen, A., Yan, Z., & Nahrstedt, K. (2018, October). Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In Proceedings of the 26th ACM international conference on Multimedia (pp. 1190-1198). Data origin: Nguyen, A., Yan, Z., & Nahrstedt, K. (2018, October). Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In Proceedings of the 26th ACM international conference on Multimedia (pp. 1190-1198). https://github.com/phananh1010/PanoSalNet. Expected size: ~1GB Data utility: This dataset is useful in our context to train deep neural networks to predict which parts of 360° videos attract viewers to watch the most. This dataset however can be leveraged in various novel applications in a much broader scope. For example, it can be used by researchers, engineers, and hobbyists to either optimize existing 360° video streaming applications (rate-distortion optimization, saliency prediction) and novel applications (crowd-driven camera movements).
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted in the github repository of the authors https://github.com/phananh1010/PanoSalNet. Search keywords: N/A

	Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on the repository of the authors https://github.com/phananh1010/PanoSalNet. How it will be accessible: The data is accessible by following a Dropbox link: https://www.dropbox.com/s/smiplkpqlv0npsm/panosalnet_iter_800.caffemodel?dl=0 given in the github repository. Methods/software tools to access data: N/A. Repository: Github. Restrictions on access: N/A.
Making data interoperable	Interoperability: The data is already interoperable. Data and metadata vocabularies: Equirectangular saliency maps and sensory data (head position) are in numpy array format inside a Python dictionary the weights of the model are given in a caffemodel format. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: To use the dataset, you should cite the work of Nguyen, A., Yan, Z., & Nahrstedt, K. (2018, October). Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In Proceedings of the 26th ACM international conference on Multimedia (pp. 1190-1198). Availability for re-use: N/A Usable by third parties after end of project: Data is already publicly available. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UCA servers. UCA fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A. Is informed consent for data sharing and long term preservation given: N/A (Note that actors in the dataset provided their consent for the creation of these traces and the users cannot be identified, and therefore no personal data is used.)
Other Issues	N/A

5.3.12 Dataset of Head and Eye Movements for 360° Videos

DMP component	AI4Media_Data_52_WP5_VIDEO_HeadEyeMovements360Videos_v1 Partner: UCA
Data	Purpose: The dataset of Head and Eye Movements for 360° Videos is composed of 19

Summary	videos in 4K resolution in equirectangular format. It contains the exploration of 57 observers on all 19 videos for a duration of 20 seconds. Visual attention data is organized in folders according to their data type: whether if they come from Head-only movements or head and eye movements. For both cases, saliency maps are stored in a separate folder. It contains one saliency map per stimulus as a compressed binary file. The scanpaths are stored in a directory as CSV text files for each video. These CSV files contain all identified fixations for one video, ordered temporally for each observer one after the other in the file. The first data column reports fixation indexes for each participant, this value is incremented with each new fixation until reaching the end of an observer's trial, after which indexing starts over at 0 for the next observer. Next two columns are gaze positions in longitudes and latitudes normalized between 0 and 1 and then two columns contain the head positions in the same format. This dataset is useful in the context of WP5 to train deep neural networks to predict which parts of 360° videos attract viewers to watch the most. Type/format: videos in mp4, scanpaths in csv files and saliency maps in binary files. Re-use of existing data: Yes, we are re-using an existing dataset from David, E. J., Gutiérrez, J., Coutrot, A., Da Silva, M. P., & Callet, P. L. (2018, June). A dataset of head and eye movements for 360 videos. In Proceedings of the 9th ACM Multimedia Systems Conference (pp. 432-437) this dataset was used in the Salient360°! ICME'18 Grand Challenge. Data origin: David, E. J., Gutiérrez, J., Coutrot, A., Da Silva, M. P., & Callet, P. L. (2018, June). A dataset of head and eye movements for 360 videos. In Proceedings of the 9th ACM Multimedia Systems Conference (pp. 432-437). https://salient360.ls2n.fr/datasets/training-dataset/. Expected size: ~1GB Data utility: This dataset is useful in our context to train deep neural networks to predict which parts of 360° videos attract viewers to watch the most. This dataset however can be leveraged in various novel applications in a much broader scope. For example, it can be used by researchers, engineers, and hobbyists to either optimize existing 360° video streaming applications (rate-distortion optimization, saliency prediction) and novel applications for coding, transmitting, and rendering 360° content.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted in the website of the salient360! challenge https://salient360.ls2n.fr/datasets/training-dataset/. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on the repository of the Salient360! Challenge https://salient360.ls2n.fr/datasets/training-dataset/. How it will be accessible: The data is accessible by a ftp link: ftp://gdchallenge18:1piN4nte5@ftp.ivc.polytech.univ-nantes.fr given in the repository of the University of Nantes. Methods/software tools to access data: The data could be downloaded using any

	software to download ftp files. Repository: https://salient360.ls2n.fr/datasets/training-dataset/. Restrictions on access: N/A.
Making data interoperable	Interoperability: The data is already interoperable. Data and metadata vocabularies: Equirectangular videos are in mp4 format, saliency maps extracted from the videos are in binary format and sensory data (head orientation and eye motion) is stored in text files in csv format. There is also a readme file in txt format explaining the different folders in the dataset and its usage. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: To use the dataset, you should cite the work of David, E. J., Gutiérrez, J., Coutrot, A., Da Silva, M. P., & Callet, P. L. (2018, June). A dataset of head and eye movements for 360 videos. In Proceedings of the 9th ACM Multimedia Systems Conference (pp. 432-437). Availability for re-use: N/A Usable by third parties after end of project: Data is already publicly available. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on UCA servers. UCA fully complies with the applicable national and European data protection framework. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A. Is informed consent for data sharing and long term preservation given: N/A (Note that actors in the dataset provided their consent for the creation of these traces and they received monetary compensation for their time, the users cannot be identified in the dataset, and therefore no personal data is used).
Other Issues	N/A

5.3.13 Dim-sim dataset for music similarity search

DMP component	**AI4Media_Data_53_WP5_Audio_dim-sim_v1** **Partner: FhG-IDMT**
Data Summary	Purpose: The dim-sim dataset is a collection of user-annotated triplet ratings for music similarity. The dataset can be used to train and to evaluate algorithms for music similarity search and music recommendation. All annotations relate to audio files from the Million Song Dataset (MSD). 4,000 3-secong triplets were randomly sampled from the MSD and were each annotated by 5-12 annotators w.r.t. to which song being more similar to the anchor song. In total, the dataset includes 39,400 human annotations. Furthermore, a subset

	of cleaned annotations with higher agreement is additionally provided. Type/format: Triplet annotations (csv / json) with audio file names (MSD) and similarity ratings[9] Re-use of existing data: Yes Data origin: Million Song Dataset (MSD) http://millionsongdataset.com/ Expected size: <1MB Data utility: The dataset is useful in the context of T5.6 for the evaluation of algorithms for disentangled music similarity search.
	Is data discoverable: Data is discoverable. The dataset is hosted on Zenodo: https://zenodo.org/record/3889149#.XuovcxMzbyV Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The dataset is openly accessible. How it will be accessible: Shared through a third-party repository link: https://zenodo.org/record/3889149#.XuovcxMzbyV Methods/software tools to access data: Web-browser to download the data as zip file. Repository: Zenodo Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under a non-commercial attribution licence (Creative Commons Attribution Non Commercial 4.0 International) Availability for re-use: Yes Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on FhG-IDMT's servers. FhG-IDMT fully complies with the applicable national and European data protection frameworks. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.

[9] As documented here: https://jongpillee.github.io/multi-dim-music-sim/

Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the dataset should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.14 SPAM dataset for music segmentation

DMP component	AI4Media_Data_54_WP5_Audio_spam_music_v1 Partner: FhG-IDMT
Data Summary	Purpose: The SPAM dataset includes structural annotations for 50 tracks from 5 human annotators each[10]. These annotations include the segment boundary times as well as segment labels indicating similar segments such as chorus or vers. As the original audio files are copyright-protected, the dataset instead includes 5 different types of (pre-computed) audio features relating to timbre, tonality/harmony, and rhythm. Type/format: Annotations (json), Pre-computed audio features (hosted separately at CCRMA Stanford)[11], audio file metadata (tsv), python scripts for dataset parsing Re-use of existing data: Yes Data origin: 50 copyright-protected audio files of various origins Expected size: ~1.2 GB Data utility: The dataset is useful in the context of T5.6 for the evaluation of algorithms for music structure analysis.
	Is data discoverable: Data is discoverable. The dataset is hosted on Github and (linked from there) on a website hosted by CCRMA: https://github.com/urinieto/msaf-data/tree/master/SPAM Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The dataset is openly accessible. How it will be accessible: Shared through a third-party repository link Methods/software tools to access data: Web-browser to download the dataset as zip file. Repository: Github Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A

[10] https://github.com/urinieto/msaf-data/tree/master/SPAM

[11] https://ccrma.stanford.edu/~urinieto/SPAM/SPAM-features.tgz

	Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under a not-specified licence. Availability for re-use: Yes Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on FhG-IDMT's servers. FhG-IDMT fully complies with the applicable national and European data protection frameworks. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the dataset should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.15 SALAMI dataset for music segmentation

DMP component	**AI4Media_Data_55_WP5_Audio_salami_music_v1** **Partner: FhG-IDMT**
Data Summary	Purpose: The SALAMI (Structural Analysis of Large Amounts of Music Information) dataset includes structural annotations for around 1300 licence free music recordings from one or multiple annotators. Type/format: Annotations (txt) + Metadata Re-use of existing data: Yes Data origin: 50 copyright-protected audio files of various origins: https://github.com/DDMAL/salami-data-public Expected size: ~1.2 GB Data utility: The dataset is useful in the context of T5.6 for the evaluation of algorithms for music structure analysis.
	Is data discoverable: Data is discoverable. The dataset is hosted on Github (https://github.com/DDMAL/salami-data-public) (a later version was hosted on the facility of McGill University: https://ddmal.music.mcgill.ca/research/SALAMI/annotation/) Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly	Data openly accessible: The dataset is openly accessible.

accessible	How it will be accessible: Shared through a third-party repository link: https://github.com/DDMAL/salami-data-public Methods/software tools to access data: Web-browser to download the dataset, audio files can be partly accessed via matching Youtube-links[12]. Repository: Github Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under a Creative Commons 0 licence. Availability for re-use: Yes Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on FhG-IDMT's servers. FhG-IDMT fully complies with the applicable national and European data protection frameworks. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the dataset should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.16 Harmonix dataset for music segmentation

DMP component	**AI4Media_Data_56_WP5_Audio_harmonix_music_v1** **Partner: FhG-IDMT**
Data Summary	Purpose: The Harmonix dataset includes metrical (beats & downbeats) and structural annotations for 912 pop tracks. Type/format: Annotations (json), Pre-computed audio features (mel-spectrograms) hosted on a dropbox account Re-use of existing data: Yes Data origin: 912 copyright-protected Western pop music recordings at https://github.com/urinieto/harmonixset

[12] https://github.com/jblsmith/matching-salami

	Expected size: ~1.3 GB Data utility: The dataset is useful in the context of T5.6 for the evaluation of algorithms for music structure analysis.
	Is data discoverable: Data is discoverable. The dataset is hosted on Github (https://github.com/urinieto/harmonixset) and (linked from there) on an Dropbox account (https://www.dropbox.com/s/zxnqlx0hxz0lsyc/Harmonix_melspecs.tgz?dl=0). Audio files can be retrieved via matching Youtube-URLs. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The dataset is openly accessible. How it will be accessible: Shared through a third-party repository link https://github.com/urinieto/harmonixset Methods/software tools to access data: Web-browser to download the dataset as zip file. Repository: Github Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is publicly available under the MIT Licence. Availability for re-use: Yes Usable by third parties after end of project: This is an open dataset. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on FhG-IDMT's servers. FhG-IDMT fully complies with the applicable national and European data protection frameworks. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Links to the dataset should be shared instead of raw data. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.17 Free Music Archive dataset

DMP component	AI4Media_Data_57_WP5_Audio_FMA_v1 Partner: IRCAM
Data Summary	Purpose: The Free Music Archive (FMA) song dataset is a collection of 106,574 musical recordings, from 16,341 artists. It comes with musical genre annotations, artist and album names, and other metadata. This dataset has mainly an interest for non-supervised learning, and training of GANs for the sound synthesis of musical mixes. Type/format: MP3 files for the audio recordings, and CSV files for the metadata. Re-use of existing data: Yes, we reuse a public dataset from EPFL. Data origin: http://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis Expected size: 879 GB Data utility: This dataset will be used by IRCAM for the sound synthesis and audio analysis (T5.2, T5.6, and Use Case 5).
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable in the original source: http://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis Search keywords: FMA music Versioning: N/A Metadata creation: The metadata are written in the standard CSV files.
Making data openly accessible	Data openly accessible: The data is already publicly shared at http://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis. We will not re-share the data. How it will be accessible: Can be downloaded from http://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis. Methods/software tools to access data: Web-browser Repository: http://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis. Restrictions on access: No
Making data interoperable	Interoperability: The data and metadata formats are standard, which makes the use of the dataset easy. Data and metadata vocabularies: A documentation and a publication are accessible on the repository website for an explanation of the content. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: The data is already shared under a Creative Commons license. Availability for re-use: Already publicly shared Usable by third parties after end of project: N/A Re-use timeframe: N/A

	Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the dataset is stored in IRCAM's servers, for which access requires username/password authentication. Security measures prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.18 LAKH MIDI music dataset

DMP component	**AI4Media_Data_58_WP5_Audio_LAKH-MIDI_v1** **Partner: IRCAM**
Data Summary	Purpose: The Lakh MIDI dataset is a collection of digital scores of 176,581 songs, with the MIDI format and other annotations. It will be used in WP5 for the sound synthesis of full musical mixes, and the learning of score augmentation. Type/format: MIDI files. Re-use of existing data: Yes, we re-use a public dataset from Columbia University. Data origin: https://colinraffel.com/projects/lmd/ Expected size: 7.6 GB Data utility: This dataset will be used by IRCAM for sound synthesis, MIDI score augmentation, and possibly for automatic transcription and time alignment (T5.2, T5.6, and Use Case 5).
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable in the original source: https://colinraffel.com/projects/lmd/ Search keywords: LAKH MIDI music Versioning: N/A Metadata creation: The Lakh dataset is made of MIDI files without external metadata.
Making data openly accessible	Data openly accessible: The data is already publicly shared at https://colinraffel.com/projects/lmd/. We will not re-share the data. How it will be accessible: Can be downloaded from https://colinraffel.com/projects/lmd/ Methods/software tools to access data: Web-browser Repository: https://colinraffel.com/projects/lmd/ Restrictions on access: No
Making data interoperable	Interoperability: The MIDI format is standard in computer music. It is well documented and many libraries and software are available to read it.

	Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is already shared under a CC-BY 4.0 license. Availability for re-use: Already publicly shared Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the dataset is stored in IRCAM's servers, for which access requires username/password authentication. Security measures prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: No. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.3.19 Piano Audio and MIDI music datasets

DMP component	**AI4Media_Data_59_WP5_Audio_MIDI_Piano_v1** **Partner: IRCAM**
Data Summary	Purpose: The ENST-MAPS and the MAESTRO datasets are two collections of real piano recordings: (1) individual notes and (2) full musical piano pieces, coming with the corresponding digital scores. The recordings and the MIDI scores have been produced using a mechanic piano (YAMAHA disklavier). These data are useful for piano sound synthesis and analyses, such as automatic transcription. Type/format: WAV files for the audio recordings, and MIDI files for the scores. Re-use of existing data: Yes, we re-use public datasets from Telecom-Paris and Google AI. Data origin: ENST-MAPS available on the Telecom-Paris website: http://www.tsi.telecom-paristech.fr/aao/en/2010/07/08/maps-database-a-piano-database-for-multipitch-estimation-and-automatic-transcription-of-music/ , and MAESTRO available with the Magenta Framework of Google AI: https://magenta.tensorflow.org/datasets/maestro Expected size: ENST-MAPS: 32 GB, and MAESTRO: 120GB. Data utility: These datasets will be used by IRCAM for sound synthesis, MIDI score augmentation, and for piano synthesis and analysis (T5.2, T5.6, and Use Case 5).
Making data findable, incl. provisions for metadata	Is data discoverable: Data are discoverable in the original sources: http://www.tsi.telecom-paristech.fr/aao/en/2010/07/08/maps-database-a-piano-database-for-multipitch-estimation-and-automatic-transcription-of-music/ and

	https://magenta.tensorflow.org/datasets/maestro. Search keywords: ENST MAPS piano, MAESTRO Versioning: N/A Metadata creation: The metadata are in the standard MIDI format for computer music.
Making data openly accessible	Data openly accessible: The data are already publicly shared. We will not re-share the data. How it will be accessible: Can be downloaded from http://www.tsi.telecom-paristech.fr/aao/en/2010/07/08/maps-database-a-piano-database-for-multipitch-estimation-and-automatic-transcription-of-music/ and https://magenta.tensorflow.org/datasets/maestro. Methods/software tools to access data: Web-browser Repository: France Telecom repository, Magenta Framework of Google AI Restrictions on access: No
Making data interoperable	Interoperability: The data and metadata formats are standard, which makes the use of the dataset easy. Data and metadata vocabularies: Documentations and publications are accessible on the repository websites for an explanation of the content. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: For ENST-MAPS, a subset is under Creative Commons License, and the three other subsets have no given License. MAESTRO is fully under a CC BY-NC-SA 4.0 (Non-Commercial Creative Commons). Availability for re-use: Already publicly shared Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the dataset is stored in IRCAM's servers, for which access requires username/password authentication. Security measures prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: No
Other Issues	No

5.3.20 GiantSteps music datasets

DMP component	AI4Media_Data_60_WP5_Audio_ GiantSteps _v1 Partner: IRCAM
Data Summary	Purpose: The GiantSteps datasets are two collections of musical recordings annotated in tempo (664 songs) and in harmonic key (604 songs) for research purposes. These datasets are useful for training and testing tempo and key recognition methods in WP5. Type/format: MP3 files for the audio recordings, and JAMS files for the annotations. Re-use of existing data: Yes, we re-use a public dataset from Universitat Pompeu Fabra. Data origin: Key dataset: https://github.com/GiantSteps/giantsteps-key-dataset, and Tempo dataset: https://github.com/GiantSteps/giantsteps-tempo-datase Expected size: 1.8 GB Data utility: This dataset will be used by IRCAM evaluate the recognition of tempo and harmonic key (T5.6, and Use Case 5).
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable in the original source: https://github.com/GiantSteps/ Search keywords: GiantSteps music Versioning: N/A Metadata creation: The metadata are in JAMS format (JSON Annotated Music Specification), a format for reproducible MIR research. It is documented on the following webpage: https://github.com/marl/jams
Making data openly accessible	Data openly accessible: The data is already publicly shared at https://github.com/GiantSteps/. We will not re-share the data. How it will be accessible: Can be downloaded from https://github.com/GiantSteps/ Methods/software tools to access data: Web-browser Repository: GitHub Restrictions on access: No
Making data interoperable	Interoperability: The data and metadata formats are standard, which makes the use of the dataset easy. Data and metadata vocabularies: Documentations and publications are accessible on the repository websites for an explanation of the content. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The datasets are already publicly shared in GitHub. Availability for re-use: Already publicly shared Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A

	Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the dataset is stored in IRCAM's servers, for which access requires username/password authentication. Security measures prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.4 Datasets used in the context of WP6

5.4.1 Deepfake Detection Challenge dataset

DMP component	AI4Media_Data_61_WP6_VIDEO_Deepfake-Detection-Challenge-Dataset_v1 Partner: CERTH
Data Summary	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. This full dataset was used by participants during a Kaggle competition to create new and better models to detect manipulated media. The dataset was created by Facebook with paid actors who entered into an agreement to the use and manipulation of their likenesses in the creation of the dataset. The dataset will be used by CERTH is the context of T6.2 to develop and test new algorithms for the detection of deep fake videos in the web, focusing on facial manipulation. Type/format: Videos (mp4) Re-use of existing data: Yes, we are re-using an existing dataset from Kaggle.com Data origin: Kaggle.com/ AWS Expected size: ~500GB Data utility: It is useful in the context of T6.2 for the detection of synthetic content in videos including faces. In general, this dataset is useful for any researcher that wants to train deep learning models for facial manipulation detection using a large-scale dataset.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted on Kaggle or AWS platforms and is discoverable by googling "Deepfake Detection Challenge dataset". See https://ai.facebook.com/datasets/dfdc/ Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on Kaggle or AWS platforms. The data will not be re-shared by AI4Media partners. How it will be accessible: The data is hosted on Kaggle or AWS platforms. Details on:

	https://ai.facebook.com/datasets/dfdc/ Methods/software tools to access data: Creation of Kaggle account or AWS account with an IAM user and Access Keys. Repository: Kaggle or AWS platforms Restrictions on access: The user should accept licence agreement first.
Making data interoperable	Interoperability: The data is already interoperable. Data and metadata vocabularies: Videos are in mp3 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 the manipulated video. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: To download the dataset from Kaggle the user had to agree to the challenge rules https://www.kaggle.com/c/deepfake-detection-challenge/rules. The guidelines and licence for the DFDC dataset are listed in section 7. COMPETITION DATA. Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate most of the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: Licence prevents the user from re-sharing the dataset. Is informed consent for data sharing and long term preservation given: N/A (Note that actors in the dataset provided their consent for the creation of these videos)
Other Issues	N/A

5.4.2 FaceForensics++ dataset

DMP component	**AI4Media_Data_62_WP6_VIDEO_ FaceForensics++_v1** **Partner: CERTH**
Data Summary	Purpose: FaceForensics++ is a forensics dataset consisting of 1,000 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. The dataset will be used by CERTH is the context of T6.2 to develop and test new algorithms for the detection of deep fake videos in the web, focusing on facial manipulation. Type/format: Videos (mp4) Re-use of existing data: Yes, we are re-using existing data. Original videos are taken from YouTube. Data origin: youtube.com Expected size: ~400GB (all video qualities) Data utility: It is useful in the context of T6.2 for the detection of synthetic content in videos including faces. In general, this dataset is useful for any researcher that wants to train deep learning models for facial manipulation detection using a large-scale dataset.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable in the authors' GitHub repo https://github.com/ondyari/FaceForensics Search keywords: N/A Versioning: GitHub supports versioning Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on GitHub at https://github.com/ondyari/FaceForensics We will not re-share the data. How it will be accessible: The data can be downloaded from the original source after filling an online form: https://docs.google.com/forms/d/e/1FAIpQLSdRRR3L5zAv6tQ_CKxmK4W96tAab_pfBu2EKAgQbeDVhmXagg/viewform Methods/software tools to access data: Data owner provides a download script Repository: N/A Restrictions on access: The user should accept the terms of use: http://kaldir.vc.in.tum.de/faceforensics_tos.pdf
Making data interoperable	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. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is released under the FaceForensics Terms of Use, and the code is released under the MIT license. Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A

Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate most of the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: The user may provide research associates and colleagues with access to the database if they first agree to be bound by the terms and conditions. Is informed consent for data sharing and long term preservation given: N/A (Original data was mined from youtube and there is no consent from the subject appearing in the videos)
Other Issues	N/A

5.4.3 Visual profile impact rating and ranking – ImageCLEFaware dataset

DMP component	AI4Media_Data_63_WP4_IMAGE_ImageCLEFaware-dataset_v1 Partner: CEA
Data Summary	Purpose: This dataset is designed to evaluate the ability of machine learning methods to compute automatic ratings of visual user profiles made of photos in impactful real-life situations. Photos which compose the profiles have been sampled from the YFCC100M dataset. With regard to object detections, visual concept scores were crowdsourced in an experiment carried out by partner CEA. Object detectors were trained using a model trained with a combination of the publicly available MS-COCO, ImageNet and OpenImages datasets. The dataset will be used by partners CEA and UPB in the context of T6.7 to develop and test new algorithms for visual profile rating. A minimized version of the dataset (excluding photographs and with anonymized visual concepts) will be released publicly as part of the ImageCLEF 2021 evaluation campaign. Type/format: Images (JPEG), metadata (JSON) Re-use of existing data: Yes, we are re-using existing data. Original photos are from the YFCC100M dataset, which is itself collected from Flickr. Data origin: flickr.com Expected size: ~6MB Data utility: It is useful in the context of T6.2 for providing users with feedback about potentially serious real-life effects of personal data sharing.
Making data findable, incl. provisions for metadata	Is data discoverable: Data will be made available as a subproject of the CEA's github account: https://github.com/cea-list-lasti Search keywords: N/A Versioning: GitHub supports versioning Metadata creation: N/A

Making data openly accessible	Data openly accessible: The data will be openly accessible via GitHub at https://github.com/cea-list-lasti How it will be accessible: The data can be downloaded from an online archive after completing a form. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: The user should accept the terms of use.
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Anonymized image detections are provided for train/val/test users in separate files. Anonymized visual concept ratings are provided per situation. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is released under the ImageCLEFaware Terms of Use, and the code is released under the CC license. Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The full dataset (including images and non-anonymized metadata) will be hosted on CEA's servers. CEA fully complies with the applicable national, European and International framework, and the GDPR. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The version of the dataset which is shared publicly includes data minimization, in compliance with art. 9 of GDPR. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

5.4.4 DEAP EEG dataset

DMP component	**AI4Media_Data_64_WP6_EEG_DEAP_v1** **Partner: QMUL**
Data Summary	Purpose: DEAP is a dataset of human physiological signal recordings (EEG) and facial video recordings, originally created for emotion recognition purposes. QMUL will use the dataset stored in Python (.npy) and video (.avi) file format. It will be used by QMUL in T6.6 to evaluate the models developed in the task.

	Type/format: CSV file containing metadata, 880 videos stored in .avi format. Re-use of existing data: Yes, we will reuse an existing dataset Data origin: https://www.eecs.qmul.ac.uk/mmv/datasets/deap/index.html Expected size: 12 GB Data utility: It is useful to WP6 partners to evaluate EEG-based emotion recognition models.
Making data findable, incl. provisions for metadata	Is data discoverable: The dataset is discoverable from its website: https://www.eecs.qmul.ac.uk/mmv/datasets/deap/index.html. The dataset is stored on the servers of Queen Mary University of London. Search keywords: DEAP Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The dataset is not openly accessible. Access is allowed only to users that have been given credentials after signing an End User License Agreement (EULA) form. How it will be accessible: The EULA form has been sent and access has been granted, together with credentials to download it. Methods/software tools to access data: The dataset is downloaded from an internet browser, without use of any other tool Repository: Network Repository (http://networkrepository.com) Restrictions on access: None
Making data interoperable	Interoperability: The data are interoperable. Data and metadata vocabularies: Vocabularies used in the dataset are clearly defined in the dataset description: https://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html Use of standard vocabularies: As defined above. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is already publicly available under an End User License Agreement: https://www.eecs.qmul.ac.uk/mmv/datasets/deap/doc/eula.pdf Availability for re-use: The dataset is available for re-use, only under the terms of the EULA form. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the data are stored in the servers of Queen Mary University of London. Access requires username/password authentication. The security measures taken prevent illegitimate access (firewalls and rights-based-file

	system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: The EULA terms clearly prevent sharing the dataset. Is informed consent for data sharing and long term preservation given: N/A (Participants provided their consent for the creation of the dataset)
Other Issues	N/A

5.4.5 SEED EEG dataset

DMP component	**AI4Media_Data_65_WP6_EEG_SEED_v1** **Partner: QMUL**
Data Summary	Purpose: SEED is a dataset of human physiological signal recordings (EEG), originally created for emotion recognition purposes. We will use the dataset stored in Matlab (.mat) file format. It will be used in T6.6 to evaluate the models developed in the task. Type/format: Excel (.xls) files containing metadata, EEG signals and EEG features stored in .mat format. Re-use of existing data: Yes, we will reuse an existing dataset Data origin: http://bcmi.sjtu.edu.cn/~seed/seed.html Expected size: 10 GB Data utility: It is useful to WP6 partners to evaluate EEG-based emotion recognition models.
Making data findable, incl. provisions for metadata	Is data discoverable: The dataset is discoverable from its website: http://bcmi.sjtu.edu.cn/~seed/seed.html. Access to the dataset is handled by Shanghai Jiao Tong University. Search keywords: SEED Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The dataset is not openly accessible. Access is allowed only to users that have been given credentials after signing an End User License Agreement (EULA) form. How it will be accessible: The EULA form has been sent and access has been granted, together with credentials to download it. Methods/software tools to access data: The dataset is downloaded from an internet browser, without use of any other tool Repository: N/A Restrictions on access: None
Making data interoperable	Interoperability: The data are interoperable. Data and metadata vocabularies: Vocabularies used in the dataset are clearly defined in the dataset description: http://bcmi.sjtu.edu.cn/~seed/seed.html Use of standard vocabularies: As defined above.

	Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is already publicly available under an End User License Agreement: http://bcmi.sjtu.edu.cn/~seed/resource/license/license. Availability for re-use: The dataset is available for re-use, only under the terms of the EULA form. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the data are stored in the servers of Queen Mary University of London. Access requires username/password authentication. The security measures taken prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: The EULA terms clearly prevent sharing the dataset. Is informed consent for data sharing and long term preservation given: N/A (Participants provided their consent for the creation of the dataset)
Other Issues	N/A

5.4.6 SEED-IV EEG dataset

DMP component	**AI4Media_Data_66_WP6_EEG_SEED-IV_v1** **Partner: QMUL**
Data Summary	Purpose: SEED-IV is a dataset of human physiological signal recordings (EEG), originally created for emotion recognition purposes. We will use the dataset stored in Matlab (.mat) file format. It will be used in T6.6 by QMUL to evaluate the models developed in the task. Type/format: Excel (.xls) files containing metadata, EEG signals and EEG features stored in .mat format. Re-use of existing data: Yes, we will reuse an existing dataset Data origin: http://bcmi.sjtu.edu.cn/~seed/seed-iv.html Expected size: 7 GB Data utility: It is useful to WP6 partners to evaluate EEG-based emotion recognition models.
Making data findable, incl. provisions for metadata	Is data discoverable: http://bcmi.sjtu.edu.cn/~seed/seed-iv.html. Access to the dataset is handled by Shanghai Jiao Tong University. Search keywords: SEED-IV Versioning: N/A Metadata creation: N/A
Making data	Data openly accessible: The dataset is not openly accessible. Access is allowed only to

openly accessible	users that have been given credentials after signing an End User License Agreement (EULA) form. How it will be accessible: The EULA form has been sent and access has been granted, together with credentials to download it. Methods/software tools to access data: The dataset is downloaded from an internet browser, without use of any other tool. Repository: N/A Restrictions on access: None
Making data interoperable	Interoperability: The data are interoperable. Data and metadata vocabularies: Vocabularies used in the dataset are clearly defined in the dataset description: http://bcmi.sjtu.edu.cn/~seed/seed-iv.html Use of standard vocabularies: As defined above. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is already publicly available under an End User License Agreement: http://bcmi.sjtu.edu.cn/~seed/resource/license/license-SEED-IV.pdf Availability for re-use: The dataset is available for re-use, only under the terms of the EULA form. Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: After downloading, the data are stored in the servers of Queen Mary University of London. Access requires username/password authentication. The security measures taken prevent illegitimate access (firewalls and rights-based-file system).
Ethical aspects	Possible ethical and legal aspects preventing sharing: The EULA terms clearly prevent sharing the dataset. Is informed consent for data sharing and long term preservation given: N/A (Participants provided their consent for the creation of the dataset)
Other Issues	N/A

5.4.7 Clotho audio captioning dataset

DMP component	**AI4Media_Data_67_WP6_Audio_ Clotho_v1** **Partner: CERTH**
Data Summary	Purpose: The Clotho dataset consists of audio samples of 15 to 30 seconds duration, in WAVE format, and each audio sample has five captions of eight to 20 words length. There is a total of 4,981 audio samples in the dataset with 24,905 captions. The dataset will be used by CERTH is the context of T6.2 to develop and test new algorithms for automated audio captioning, where general audio content is described

	using free text. The final system will use an input audio signal and it will output the textual description (i.e., the caption) of that signal. Type/format: Audio (wav) Re-use of existing data: Yes, we are reusing an existing dataset. Data origin: https://zenodo.org/record/3490684 Expected size: 6 GB Data utility: It is useful in the context of T6.2 for automated audio captioning. In general, this dataset is useful for any researcher that wants to train deep learning models for audio signal processing by using a dataset captured in the wild.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable in the authors' GitHub repo https://github.com/audio-captioning/clotho-dataset. Search keywords: N/A Versioning: GitHub supports versioning Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on GitHub at https://github.com/audio-captioning/clotho-dataset. We will not re-share the data. How it will be accessible: It can be accessed via running the '.sh' scripts provided by the authors on the aforementioned GitHub link, or it can be downloaded directly from Zenodo. Methods/software tools to access data: Data owner provides a download script. Repository: Zenodo Restrictions on access: N/A
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. The authors have provided all the necessary files for training, validation and testing, separately, along with the ground truth for each file. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is released under the Tampere University, Finland Terms of Use. Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and

	institutional policies mitigate most of the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: The user may provide research associates and colleagues with access to the database if they first agree to be bound by the terms and conditions. The data will not be reshared by CERTH. Is informed consent for data sharing and long term preservation given: N/A (Original data was mined from Zenodo and there is no consent from the subject that is heard in the recordings).
Other Issues	No

5.4.8 ASVspoof2019 dataset

DMP component	**AI4Media_Data_68_WP6_Audio_ ASVspoof_v1** **Partner: CERTH**
Data Summary	Purpose:The ASVspoof 2019 dataset consists of audio samples from 107 speakers (46 males and 61 females). It was released by the University of Edinburgh in collaboration with Google. The database encompasses two partitions for the assessment of logical access and physical access scenarios. The original waveform format is PCM and compressed using FLAC. No telephone or mobile codec was used. The dataset will be used by CERTH is the context of T6.2 to develop and test new algorithms for deepfake detection, focusing on manipulated speech. Type/format: Audio (flac) Re-use of existing data: Yes, we use an existing dataset Data origin: Audio dataset created by Uni Edinburgh and Google https://datashare.ed.ac.uk/handle/10283/3336 Expected size: ∽ 25 GB Data utility: It is useful in the context of T6.2 for the detection of fake audios to spoof automatic speaker verification (ASV) systems. The dataset is suited not only to study of ASV replay spoofing and countermeasures, but also the study of fake audio detection in the case of e.g., smart home devices that facilitate the security of online banking and payment solutions.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable in the University of Edinburgh DataShare repo: https://datashare.ed.ac.uk/handle/10283/3336 Search keywords: N/A Versioning: The DataShare repository supports versioning (e.g., 2013, 2015, 2017 and 2019) Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on the DataShare repo of the University of Edinburgh. We will not re-share the data. How it will be accessible: It can be accessed via downloading the necessary parts from the DataShare repo

	Methods/software tools to access data: N/A Repository: DataShare repo of the University of Edinburgh Restrictions on access: N/A
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. The authors have provided the evaluation plan for the database in the following link: https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is released under the Open Data Commons Attribution License Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded and stored on CERTH's servers. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate most of the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: The user may provide research associates and colleagues with access to the database if they first agree to be bound by the terms and conditions Is informed consent for data sharing and long term preservation given: N/A(Original data was mined from the University of Edinburgh and there is no consent from the subject heard in the recordings).
Other Issues	No

5.4.9 MOBIPHONE audio dataset

DMP component	**AI4Media_Data_69_WP6_Audio_MOBIPHONE_v1** **Partner: FhG-IDMT**
Data Summary	Purpose: MOBIPHONE is a forensics dataset consisting of audio recordings acquired by 21 mobile phones of various models from 7 different brands, collected by recording 10 utterances, uttered by 12 male speakers and another 12 female speakers, randomly chosen from the TIMIT database. The dataset is the only publicly available dataset for microphone classification at present and has been used by several publications in this domain for benchmarking and comparison with other state-of-the-art algorithms.

	Type/format: Audio (wav) Re-use of existing data: Yes, re-use of utterances from the TIMIT speech corpus. Data origin: The data was created by the authors themselves, by replaying a subset of TIMIT utterances with a high-end loudspeaker, and recording the outcome with several devices at once Expected size: ~941MB Data utility: The dataset is useful in the context of T6.2 for developing and testing new algorithms for microphone classification and device identification, and for any research focused on manipulation detection on the basis of changes in the recording device.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable by reading the authors' paper on https://ieeexplore.ieee.org/document/6900732. A Google search for "microphone classification dataset" or "MOBIPHONE dataset" is not sufficient to discover the dataset. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible on Dropbox at https://www.dropbox.com/sh/9n7fy7moi825bgk/WFLBKxUitV. We will not re-share the data. How it will be accessible: Already accessible at the original dropbox location. Methods/software tools to access data: Web-browser to download the data as zip file. Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Audios are separated in folders, with each folder corresponding to a single microphone. Splitting the data for training, validation and testing is left to the users Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is not released under any license. Further communications with the authors are necessary to clear and clarify usage rights. Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A

Data security	Security measures: The dataset is stored on Dropbox servers, which have mechanisms in place for minimizing the risk of data loss.
Ethical aspects	Possible ethical and legal aspects preventing sharing: From a legal perspective, the lack of license prevents usage of all sorts, including sharing and re-hosting. No ethical aspects are present which may prevent re-distribution. We plan on contacting the original authors for clearing any license issues, before making use of the data. Is informed consent for data sharing and long term preservation given: All recordings in the dataset have been acquired in controlled conditions and with the informed consent of all the speakers involved, in compliance with the GDPR).
Other Issues	N/A

5.4.10 Fake-or-Real (FoR) audio dataset

DMP component	AI4Media_Data_70_WP6_Audio_Fake-or-Real_v1 Partner: FhG-IDMT
Data Summary	Purpose: The Fake-or-Real (FoR) dataset is a collection of utterances from real humans and computer generated speech. The dataset can be used to train classifiers for synthetic speech detection. The dataset aggregates data from the latest TTS solutions (such as Deep Voice 3 and Google Wavenet TTS) as well as a variety of real human speech, including pre-existing speech datasets, and the authors' own speech recordings. The data has been normalized in terms of speakers' genre, but is not clear how many of the natural voices have a synthetic counterpart for avoiding inherent biases. Nevertheless, it's the only database of this kind ever release to the research community. Type/format: Audio (wav) Re-use of existing data: The datasets is re-using recordings from the Arctic Dataset (http://festvox.org/cmu_arctic/), LJSpeech Dataset (https://keithito.com/LJ-Speech-Dataset/), and VoxForge Dataset (http://www.voxforge.org). Data origin: Previous datasets, own recordings from the authors Expected size: ~22GB (four different variants) Data utility: The dataset is useful in the context of T6.2 for the detection of synthetic speech content in audio files. The authors also included a variant of the dataset acquired by simulating a re-recording, to let the researcher community also address the problem of washed-up synthetic recordings
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable by reading the authors' paper on https://ieeexplore.ieee.org/document/8906599. A Google search for "synthetic speech dataset" is sufficient to discover the authors' paper, and thus the dataset itself. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data	Data openly accessible: The data is already openly accessible on the authors'

openly accessible	institutional page https://bil.eecs.yorku.ca/datasets/. We do not plan to re-share the data. How it will be accessible: The data can be downloaded from the original source. Methods/software tools to access data: Web-browser to download the data as zip file. Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. The files are split into disjoint training, validation and testing set. Moreover, for each split, original audios are in a separate folder from synthetic ones. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is released under the GNU General Public License (V3). Availability for re-use: Yes, if the license terms and conditions are fulfilled. Usable by third parties after end of project: Data already publicly shared. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset will be downloaded from the original source and will be stored on FHG-IDMT's servers. FHG-IDMT fully complies with the applicable national and European data protection frameworks. State-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A Is informed consent for data sharing and long term preservation given: All recordings in the dataset have been acquired in controlled conditions and with the informed consent of all the speakers involved, in compliance with the GDPR).
Other Issues	N/A

6. Data management plan for non-research datasets collected in AI4Media

This section presents non-research datasets collected by AI4Media partners to support the activities of WP1 (management), WP8 (use cases), WP9 (IAIDA), WP10 (open calls) and WP11 (dissemination). 12 non-research datasets have been identified. The list is not exhaustive and represents the current status.

In the following, we present the DMP plan for each of these datasets using the template presented in Section 4. The Table below briefly summarizes the 12 datasets presented in this section and offers a glance at the structure of the section and its subsections.

Table 4: Summary of non-research datasets collected in AI4Media

DMP component	WP	Short summary	Relevant sub-section
Data collected in WP1 (Project Management)			6.1
AI4Media_Data_71_WP1_Text_AI4MediaConsortiumContactInfo_v1	WP1	AI4Media consortium contact info dataset	6.1.1
Data collected in WP8 (Use cases & demonstrators in media, society and politics)			6.2
AI4Media_Data_72_WP8_UserData-TrulyMedia-UC1-ATC_v1	WP8	User data from Truly Media for Use case 1	6.2.1
Data collected in WP9 (Doctoral Academy and exchange programme)			6.3
AI4Media_Data_73_WP9_Text_IAIDA CourseOfferings_v1	WP9	IAIDA course offerings dataset	6.3.1
AI4Media_Data_74_WP9_Text_IAIDA Students_v1	WP9	IAIDA students dataset	6.3.2
AI4Media_Data_75_WP9_Text_IAIDA MailingList_v1	WP9	IAIDA mailing list dataset	6.3.3
AI4Media_Data_76_WP9_Text_Junior FellowsExchange_v1	WP9	AI4Media Junior Fellows exchange program dataset	6.3.4
Data collected in WP10 (Community Outreach and Growth)			6.4
AI4Media_Data_77_WP10_Competitive_call_application_datasets_v1	WP10	Competitive call application datasets	6.4.1
AI4Media_Data_78_WP10_sub-granted_projects_dataset_v1	WP10	Sub-granted projects datasets	6.4.2
AI4Media_Data_79_WP10_sub-granted_projects_dataset_v1	WP10	External experts and evaluators datasets	6.4.3
Data collected in WP11 (Communication, dissemination, exploitation and sustainability)			6.5
AI4Media_Data_80_WP1_Text_AI4MediaAssociateMembersContactInfo_v1	WP11	AI4Media associate members contact info dataset	6.5.1
AI4Media_Data_81_WP11_Text_NewsletterSubscribers _v1	WP11	AI4Media newsletter subscribers dataset	6.5.2
AI4Media_Data_82_WP11_Text_WebsiteMessages_v1	WP11	AI4Media website messages dataset	6.5.3

6.1 Datasets collected in the context of WP1

6.1.1 AI4Media consortium contact info dataset

DMP component	**AI4Media_Data_71_WP1_Text_AI4MediaConsortiumContactInfo_v1** **Partner: CERTH**
Data Summary	Purpose: This dataset contains business contact information from AI4Media consortium members, including names, affiliation, emails, office phone numbers, Skype Ids, office postal addresses, wiki user names, etc. The collected data are necessary for the communication among project partners and are collected in the context of WP1-Management. Type/format: text Re-use of existing data: No. Data origin: Data provided by AI4Media partners to CERTH in emails or excel files. Expected size: < 1MB Data utility: This data is necessary for facilitating communication among consortium members.
Making data findable, incl. provisions for metadata	Is data discoverable: No, data are not discoverable from outside. The data is stored on the project wiki and is only discoverable by consortium members with a wiki account. Search keywords: N/A Versioning: Yes, through scheduled website backups. Metadata creation: N/A.
Making data openly accessible	Data openly accessible: No. The data will only be shared internally in AI4Media since it contains personal information. How it will be accessible: Restricted access. The data is accessible only by wiki users. Information about wiki usernames and passwords can only be accessed by the wiki administrator. Methods/software tools to access data: Web browser. Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Data will not be shared. Availability for re-use: N/A Usable by third parties after end of project N/A

	Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data is stored on the project wiki, which is on a dedicated web server hosted in CERTH's premises. The wiki web site uses for its domain an SSL certificate enabling the SHA256RSA signature algorithm and forces all visits to use HTTPS to ensure the traffic is secure. The wiki is restricted only to registered users while registration is possible only by invitation. Access requires username/password authentication. CERTH fully complies with the applicable national, European and International framework, and the GDPR. The wiki uses a file-based RDBMS to enhance security. Web server and file-based DB are running on a Linux encrypted partition, which conforms to the data-at-rest GDPR guidelines. Moreover, state-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk. Regular rolling daily backups are scheduled to minimize the risk of data loss. The data will be preserved there for three years after the end of the project and will then be deleted.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Dataset includes personal data of consortium members and will thus not be shared. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

6.2 Datasets collected in the context of W8

6.2.1 User data from Truly Media for Use case 1

DMP component	**AI4Media_Data_72_WP8_UserData-TrulyMedia-UC1-ATC_v1** **Partner: ATC**
Data Summary	Purpose: In order to realise Use Case 1 in WP8, Truly Media, a web-based platform for collaborative verification co-owned by ATC and DW, will be used as the basis of the main demonstrator in UC1. In order for the test users, as well as for the project partners developing AI tools and components for UC1, to access Truly Media, ATC will collect and store personal data. The data will be used for registration and login purposes by ATC. ATC uses Twitter login that enables users to register and login to the platform with their Twitter credentials. Collected data include: Name; Email; Organization; Department; Role; Expertise; Office phone number; Twitter profile image; and Twitter handle. Type/format: Data stored in JSON format. Re-use of existing data: No. Data origin: Truly Media users registering and logging in the platform. Expected size: A few MBs.

	Data utility: This data will be used in the context of WP8 to allow test users and project partners developing AI tools and components for UC1 to access Truly Media.
Making data findable, incl. provisions for metadata	Is data discoverable: No, because the dataset described will be stored on Truly Media's databases and there are no plans for data sharing. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data will not be openly accessible since it contains personal information. How it will be accessible: It is planned that the data will only be accessible by project partner ATC. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data will not be licensed since it will only be used internally. Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset described will be stored by ATC on third-party cloud servers. Appropriate and detailed security policies, rules, and technical measures are implemented to protect data are used by the Truly Media platform and stored on the platform from improper or unauthorized access, including use of firewalls where appropriate. Security measures also include 2FA (2 Factor Authentication) with OTP (One Time Password) for extra security during login, as well as Auth2.0 and JWT for authentication and authorisation. End-to-end encryption protects from man-in-the-middle attacks and data theft. All ATC employees and data processors, who have access to and are associated with the processing of personal data, are obliged to respect the confidentiality of the stored personal data. Moreover, ATC's development team has received training from external auditors for security awareness and security best practices to avoid vulnerabilities in source code. External auditors have performed black-box penetration testing to ensure that the platform is fully secure. ATC's Data Protection Officer ensures that all processes followed are fully compliant with the GDPR provisions.

Ethical aspects	Possible ethical and legal aspects preventing sharing: The data will not be shared since it contains personal information of end users. Is informed consent for data sharing and long-term preservation given: Informed consent is given implicitly by users when completing the relevant information on the registration form and when authorising Truly Media to use their Twitter account for login purposes.
Other Issues	No

6.3 Datasets collected in the context of WP9

6.3.1 IAIDA course offerings dataset

DMP component	AI4Media_Data_73_WP9_Text_IAIDACourseOfferings_v1 Partner: AUTH
Data Summary	Purpose: This dataset will contain information about course offerings from collaborating IAIDA full and international members, for the purpose of advertising them to IAIDA students. Such information includes non-personal data (course title, date, course offer affiliation), minimal personal data about the lecturers (Full Name, lecturer affiliation) and also electronic business contact information (e-mail). The collected data are necessary for the execution of WP9 Tasks T9.1 and T9.2. Type/format: text Re-use of existing data: No. Data origin: Lecturers will enter the details in web applications or provide them in shared excel files and will be responsible for maintenance updates. All data will be moderated by AUTH to resolve inconsistencies. At any time, lectures may alter or remove their personal data and course offerings that appear on i-aida.org website, using the web applications or contacting the website moderators (AUTH). Lecturers will maintain the responsibility/option of adding, editing, and deleting course offerings, having full access to their own content at all times. Expected size: few kb Data utility: The data will be useful to I-AIDA registered students to apply for I-AIDA course offerings.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data will be discoverable in the web, though search engines. Search keywords: Through web metadata. Versioning: Yes, through scheduled website backups. Metadata creation: Course title, offer type (web/short/semester), course affiliation and lecturer full names and affiliation.
Making data openly accessible	Data openly accessible: The dataset is publicly available on www.i-aida.org website. How it will be accessible: Though web. Methods/software tools to access data: Web browser.

	Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/AT Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: Data quality assurance will be ensured though moderation from AUTH. AUTH will consistently check if the appearing information is in the correct form, and will request the respective lectures for updates, if necessary.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data will be stored at an internal AUTH server in an encrypted format. A Data Protection Impact Assessment will be performed before storage, according to GDPR provisions, in order to identify proper data security measures.
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A. Is informed consent for data sharing and long-term preservation given: Yes. A form will be requested to be signed for every lecturer involved in IAIDA program.
Other Issues	No

6.3.2 IAIDA students dataset

DMP component	**AI4Media_Data_74_WP9_Text_IAIDAStudents_v1** **Partner: AUTH**
Data Summary	Purpose: This dataset will contain minimal personal identity information about IAIDA PhD students (full name, e-mail, affiliation, supervisor id) and details about their progress in the form of courses attended, grades, ECTS collected within the IADA program. The collected data are necessary for the execution of WP9 Tasks T9.1 and T9.2, for book-keeping and administrative purposes (e.g., Provide IAIDA certificates). Type/format: text Re-use of existing data: No. Data origin: Students will enter their personal details through a secure registration login process. Their supervisors will be notified and validate this information. Details about

	the procedure will be analytically provided future i-aida.org website updates. Expected size: few kb Data utility: The data will be useful for offering IAIDA services to students (course attendance certificates etc.)
Making data findable, incl. provisions for metadata	Is data discoverable: No. Search keywords: N/A Versioning: Yes, through scheduled website backups. Metadata creation: No.
Making data openly accessible	Data openly accessible: No, because it includes personal information of students. How it will be accessible: It will be restricted Methods/software tools to access data: Web browser. Repository: N/A Restrictions on access: Personal registered student data will only be accessible by themselves, registered lecturers that offer the respective courses that the students attended, and website moderators (in encrypted/pseudo-anonymized form).
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: Data quality assurance will be ensured though moderation from AUTH.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data will be stored at an internal AUTH server in an encrypted form. A Data Protection Impact Assessment will be performed before storage, according to GDPR provisions, in order to identify proper data security measures.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The data will not be shred since it contains some personal information about IAIDA students. Is informed consent for data sharing and long-term preservation given: Yes. A form will be requested to be signed for every student involved in the IAIDA program.

Other Issues	No

6.3.3 IAIDA mailing list dataset

DMP component	**AI4Media_Data_75_WP9_Text_IAIDAMailingList_v1** **Partner: AUTH**
Data Summary	Purpose: This dataset will contain business contact information from students and lecturers and interested researches in IAIDA activities from the general AI community. The collected data are necessary for the execution of WP9 Task T9.3. Type/format: text Re-use of existing data: No. Data origin: Mailing list participants will register by following the following instructions: https://lists.auth.gr/sympa/info/aida Expected size: few kb Data utility: This mailing list will advertise IADA activities to the general public.
Making data findable, incl. provisions for metadata	Is data discoverable: No personal data is retrievable except by list moderators. Search keywords: N/A Versioning: Yes, though scheduled website backups. Metadata creation: N/A.
Making data openly accessible	Data openly accessible: No How it will be accessible: It is restricted. Methods/software tools to access data: Web browser. Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A

Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Personal data are stored in secure AUTH servers according to internal institutional procedures.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Dataset includes personal data (email addresses) and will thus not be shared. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

6.3.4 AI4Media Junior Fellows exchange program dataset

DMP component	AI4Media_Data_76_WP9_Text_JuniorFellowsExchange_v1 Partner: CERTH
Data Summary	Purpose: This dataset contains information of host and sender organizations involved in the AI4Media exchange program for Junior Fellows. This information includes: organization, organization type, country, exchange topics and interests, type of mobility, availability period, contact person name and email, logo/photo, short bios of researchers, organization profiles, etc. The collected data is necessary for the implementation of the Junior Fellows exchange program of AI4Media and it is collected in the context of WP9. Type/format: Text Re-use of existing data: No. Data origin: Data provided by Host and Sender institutions (AI4Media partners but also other organizations involved in AI research, e.g. partners in other ICT-48 projects) when submitting their profiles through an online form in the project website in order to participate in AI4Media's Junior Fellows exchange program. Expected size: <10 MBs Data utility: This data is necessary for the operation of AI4Media's Junior Fellows exchange program.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, data is discoverable from outside. The data will be displayed in a dedicated page in the project website. Search keywords: Visitors of the website will be able to search for preferred profiles using keywords and filters. Versioning: Yes, through scheduled website backups. Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, the data will be openly accessible via the project website. How it will be accessible: Displayed in a dedicated page in the project website. Methods/software tools to access data: Web browser.

	Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: N/A Usable by third parties after end of project: N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data is stored on CERTH servers. Access requires username/password authentication. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: Dataset includes personal data (name, email, short bio) from Host and Sender institutions that create their profiles to be publicly displayed in the website. Consent is provided for making this data openly accessible in the project website. Is informed consent for data sharing and long term preservation given: Yes.
Other Issues	No

6.4 Datasets collected in the context of W10

6.4.1 Competitive call application datasets

DMP component	**AI4Media_Data_77_WP10_Competitive_call_application_datasets_v1** **Partner: F6S**
Data Summary	Purpose: The competitive call applications dataset will include data collected during the open call application process. It will contain: • The applications to the competitive calls; • The evaluation results; • Communications with applicants.

	Type/format: Several types of data: documents, emails, texts etc. Re-use of existing data: No, this is an original dataset to be created in the context of WP10 as a result of the open call procedure. Data origin: The data is collected from applicants who submit a proposal via the F6S portal. Data will be collected and consolidated by the leader of the evaluation process. Communication channels with applicants will be defined in the "Guidelines for Applicants"; relevant messages must be collected by the intervening consortium members. All the information related to competitive calls will be exported to the project wiki repository (or similar platform) to enable the long-term storage and access to data by consortium members and auditors. As soon as the competitive call closes, the leader of the competitive calls process is responsible for exporting all applications from the F6S portal to the project repository. Expected size: Several GB Data utility: Data collected will be used by consortium members and external evaluators to perform the evaluation of the submitted applications and to decide whether they should be selected to participate in AI4Media. Information on applications selected to participate in the project will be used to create the sub-granted project dataset. The dataset will also be used to generate statistics and reports about the AI4Media project as requested by the Grant Agreement, which will be aggregated or anonymised data that will not compromise personal details of applicants nor any other confidential information about their projects.
Making data findable, incl. provisions for metadata	Is data discoverable: The data will be stored in the project wiki or similar platform. It will be discoverable only by registered wiki users with metadata, identifiable by participant name and/or organisation and in some cases, be indexable/findable using a persistent and unique actor key. Search keywords: N/A Versioning: Yes, the wiki supports versioning. Metadata creation: A part of the data, such as applicant and evaluation data, will be organised as structured data. Metadata used will include: • How data was created; • Time and date of creation or modification; • Source of data; • Who created data; • Expected quality of data.
Making data openly accessible	Data openly accessible: The raw data will not be openly accessible due to AI4Media's commitments to its applicants and sub-grantees in relation to personal information and business private information. However, anonymised or aggregated data from the open call participation will be made public. How it will be accessible: The collected raw data will be stored in the project wiki or a similar platform, only accessible to those partners with direct management of the open calls and respective information. A dataset including anonymised or aggregated data from the open call participation will be made public in a public repository (i.e., project website) in the form of a non-editable document). Methods/software tools to access data: Download files via web browser.

	Repository: The collected data will be stored in the project wiki or a similar platform. A dataset including anonymised or aggregated data from the open call participation will be made public in a public repository like the project website. Restrictions on access: Given the confidential nature of the information, access will be restricted to only those required for managing the data.
Making data interoperable	Interoperability: Yes, effort will be made so that the data is interoperable. Data and metadata vocabularies: Data will be available using standard spreadsheet formats (Open Office, Excel, Google Sheets, etc.), using standard text editors (Open Office, Word, Google format) and using standard encoding and formatting to reduce the likelihood of incompatibility. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset will not be made openly available due AI4Media's commitments to its applicants and sub-grantees in relation to personal information and business private information. Availability for re-use: No Usable by third parties after end of project: No Re-use timeframe: N/A Data quality assurance process: Completeness and conformity of the competitive calls exports is ensured by the leader of the evaluation process and the person responsible for performing the eligibility check as they will have to check all documents. Completeness and conformity of the evaluation data, including communications with applicants, is ensured by the evaluation committee as they will have to check all process documents to approve final ranking and selection.
Allocation of resources	Costs for making data FAIR: There are no additional costs to make data FAIR in the project, as the costs to operate each platform used in the project are already integrated into the project costs. Costs for long-term preservation: N/A
Data security	Security measures: AI4Media will maintain protection of personal data and compliance with Data Regulations as per national and European legislation regarding the protection of personal data. Procedures will be in place for applicable technical means to avoid the loss, misuse, alteration, access by unauthorised persons and/or theft of the data provided to this entity. Notwithstanding, security measures (particularly for Internet accessible data) are not impregnable. To mitigate risk of unauthorised access, access controls will be applied to data sources. As an example, applications to the AI4Media project and feedback forms will be made accessible to a limited number of team members using user-level security and permissions. AI4Media participants (data owners) will be able to exercise their right to be forgotten.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The raw data will not be made openly available due AI4Media's commitments to its applicants and sub-grantees in relation to personal information and business private information. Only anonymised or aggregated data from the open call participation will be made public. Is informed consent for data sharing and long term preservation given: The applicants

	will provide their informed consent for the collection and processing of their personal data in the context of AI4Media. The data will not be shared outside the consortium.
Other Issues	N/A

6.4.2 Sub-granted projects dataset

DMP component	AI4Media_Data_78_WP10_sub-granted_projects_dataset_v1 Partner: F6S
Data Summary	Purpose: In each AI4Media competitive open call, several projects will be selected to sign a sub-grant contract. The seed for this dataset is the information provided by applicants in the application process and by the selection committee in the evaluation report. The dataset will be extended with all information needed to run the AI4Media programme, such as deliverables submitted by sub-grantees, evaluation reports, payment requests as will be defined in the Guidelines for Applicants. To summarize, the dataset will include (but is not limited to): applications, application evaluation, contracts, deliverables submitted, deliverables' evaluations, payment requests, proof of payments, amendments and copies or summaries of messages in any form, whose content may have an impact in the programme outcome. Type/format: Several types of data: documents, emails, etc. Re-use of existing data: No, this is an original dataset to be created in the context of WP10 as a result of the open call procedure. Data origin: The data is collected from applicants, selection committee and project partners in the context of the WP10 open calls. When the evaluation process ends and the list of selected applications is available, the leader of AI4Media programme is responsible for creating the initial dataset with the applications and evaluations of the sub-grantees. Additional data will be collected as defined in the Guidelines for Applicants. Expected size: Several GB Data utility: The data is useful to WP10 partners to run the open call procedures and monitor the smooth execution of the funded projects as defined in the Guidelines for Applicants. The dataset will also be used to generate statistics and reports about the AI4Media project as requested by the Grant Agreement.
Making data findable, incl. provisions for metadata	Is data discoverable: The data will be stored in the project wiki or similar platform. It will be discoverable only by registered wiki users with metadata, identifiable by participant name and/or organisation and in some cases, be indexable/findable using a persistent and unique actor key. Search keywords: N/A Versioning: Yes, the wiki supports versioning. Metadata creation: A part of the data, such as applicant and evaluation data, will be organised as structured data. Metadata used will include: • How data was created; • Time and date of creation or modification; • Source of data; • Who created data;

	• Expected quality of data.
Making data openly accessible	Data openly accessible: The raw data will not be openly accessible due to AI4Media's commitments to its applicants and sub-grantees in relation to personal information and business private information. However, anonymised or aggregated data from the open call participation will be made public. How it will be accessible: The collected raw data will be stored in the project wiki or a similar platform and will only be accessible to partners responsible for managing the sub-grants and projects. A dataset including anonymised or aggregated data from the open call participation will be made public in a public repository (i.e., the project website) in the form of a non-editable document to ensure integrity of the information. Methods/software tools to access data: Download files via web browser. Repository: The collected data will be stored in the project wiki or a similar platform. A dataset including anonymised or aggregated data from the approved sub-projects will be made public in a public repository like the project website. Restrictions on access: Given the confidential nature of the information, access will be restricted to only those required for managing the data.
Making data interoperable	Interoperability: Yes, effort will be made so that the data is interoperable. Data and metadata vocabularies: Data will be available using standard spreadsheet formats (Open Office, Excel, Google Sheets, etc.), using standard text editors (Open Office, Word, Google format) and using standard encoding and formatting to reduce the likelihood of incompatibility. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset will not be made openly available due AI4Media's commitments to its applicants and sub-grantees in relation to personal information and business private information. Availability for re-use: No Usable by third parties after end of project: No Re-use timeframe: N/A Data quality assurance process: The Guidelines for Applicants will define the process to collect and verify the data with multiple checkpoints guaranteeing the quality of data.
Allocation of resources	Costs for making data FAIR: There are no additional costs to make data FAIR in the project, as the costs to operate each platform used in the project are already integrated into the project costs. Costs for long-term preservation: N/A
Data security	Security measures: AI4Media will maintain protection of personal data and compliance with Data Regulations as per national and European legislation regarding the protection of personal data. Procedures will be in place for applicable technical means to avoid the loss, misuse, alteration, access by unauthorised persons and/or theft of the data provided to this entity. Notwithstanding, security measures (particularly for Internet accessible data) are not impregnable. To mitigate risk of unauthorised access, access controls will be applied to data sources. As an example, applications to the AI4Media

	project and feedback forms will be made accessible to a limited number of team members using user-level security and permissions. AI4Media participants (data owners) will be able to exercise their right to be forgotten.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The raw data will not be made openly available due AI4Media's commitments to its applicants and sub-grantees in relation to personal information and business private information. Only anonymised or aggregated data from the open call participation will be made public. Is informed consent for data sharing and long term preservation given: The applicants will provide their informed consent for the collection and processing of their personal data in the context of AI4Media. The data will not be shared outside the consortium.
Other Issues	N/A

6.4.3 External experts and evaluators datasets

DMP component	**AI4Media_Data_79_WP10_external_experts_evaluators_datasets_v1** **Partner: F6S**
Data Summary	Purpose: As will be defined in the Guidelines for Applicants for the WP10 open calls, external evaluators will be involved in multiple evaluation tasks during the AI4Media programme. The purpose of this dataset is to manage the contractual relationship between experts and AI4Media. All data related with evaluation tasks is stored in the datasets “Competitive call applications” and “Sub-granted projects”. The dataset will include (but is not limited to): contracts, deliverables submitted, deliverables evaluations, payments request, proof of payments, amendments and copies or summaries of messages in any form, whose content may have an impact in the in the management of the contractual relationship with the evaluators. Type/format: Several types of data: documents, emails, texts etc. Re-use of existing data: No, this is an original dataset to be created in the context of WP10 as a result of the open call procedure. Data origin: External experts’ availability will be collected through a form on the F6S portal by submitting their expression of interest. Contracts and related documents for selected evaluators, such as declarations of honour, receipts and other expenses information will be submitted either in a form in the F6S portal or by email. Expected size: Several GB Data utility: Data collected will be used to manage the AI4Media programme as will be defined in the Guidelines for Applicants, including the evaluation of the submitted deliverables by external evaluators. The dataset will also be used to generate statistics and reports about the AI4Media project as requested by the Grant Agreement.
Making data findable, incl. provisions for metadata	Is data discoverable: The data will be stored in the project wiki or similar platform. It will be discoverable only by registered wiki users with metadata, identifiable by participant name and/or organisation and in some cases, be indexable/findable using a persistent and unique actor key. Search keywords: N/A Versioning: Yes, the wiki supports versioning. Metadata creation: A part of the data, such as applicant and evaluation data, will be

	organised as structured data. Metadata used will include: - How data was created; - Time and date of creation or modification; - Source of data; - Who created data; - Expected quality of data.
Making data openly accessible	Data openly accessible: The raw data will not be openly accessible due to AI4Media's commitments to its applicants, sub-grantees and evaluators in relation to personal information and business private information. Anonymised or aggregated data of evaluators' involvement may be made public. (In specific cases to be defined in the Guidelines for Applicants, the name and affiliation of evaluators selected to evaluate deliverables may be shared with sub-grantees to assess eventual conflicts of interest.) How it will be accessible: Data regarding evaluators will not be made accessible outside of those responsible for recruiting/ managing the evaluators Methods/software tools to access data: N/A Repository: N/A Restrictions on access: Access will be limited to those required for managing evaluators' data.
Making data interoperable	Interoperability: Yes, effort will be made so that the data is interoperable. Data and metadata vocabularies: Data will be available using standard spreadsheet formats (Open Office, Excel, Google Sheets, etc.), using standard text editors (Open Office, Word, Google format) and using standard encoding and formatting to reduce the likelihood of incompatibility. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset will not be made openly available due AI4Media's commitments to its applicants, sub-grantees and evaluators in relation to personal information and business private information. Availability for re-use: No Usable by third parties after end of project: No Re-use timeframe: N/A Data quality assurance process: The persons acting as project coordinator and project treasurer, as defined in the Grant Agreement, ensure that all documentation complies with legal requirements. The leader of the evaluation task ensures that the assigned tasks have been completed as contractually agreed and evidence is stored in relevant datasets.
Allocation of resources	Costs for making data FAIR: There are no additional costs to make data FAIR in the project, as the costs to operate each platform used in the project are already integrated into the project costs. Costs for long-term preservation: N/A
Data security	Security measures: AI4Media will maintain protection of personal data and compliance

	with Data Regulations as per national and European legislation regarding the protection of personal data. Procedures will be in place for applicable technical means to avoid the loss, misuse, alteration, access by unauthorised persons and/or theft of the data provided to this entity. Notwithstanding, security measures (particularly for Internet accessible data) are not impregnable. To mitigate risk of unauthorised access, access controls will be applied to data sources. As an example, applications to the AI4Media project and feedback forms will be made accessible to a limited number of team members using user-level security and permissions. AI4Media participants (data owners) will be able to exercise their right to be forgotten.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The raw data will not be made openly available due AI4Media's commitments to its applicants, sub-grantees and evaluators in relation to personal information and business private information. Only anonymised or aggregated data will be made public. Is informed consent for data sharing and long term preservation given: The applicants will provide their informed consent for the collection and processing of their personal data in the context of AI4Media. The data will not be shared outside the consortium.
Other Issues	N/A

6.5 Datasets collected in the context of W11

6.5.1 AI4Media associate members contact info dataset

DMP component	AI4Media_Data_80_WP1_Text_AI4MediaAssociateMembersContactInfo_v1 Partner: CERTH
Data Summary	Purpose: This dataset contains business contact information from AI4Media associate members, including name, affiliation and email of contact person. The collected data are necessary for the communication of AI4Media partners with the Associate members and are collected in the context of WP11. Type/format: text Re-use of existing data: No. Data origin: Data provided by AI4Media Associate members to CERTH as part of an application form template (https://ai4media.eu/associate-members/). Expected size: < 1MB Data utility: This data is necessary for facilitating communication with the Associate members.
Making data findable, incl. provisions for metadata	Is data discoverable: No, data are not discoverable from outside. The data is stored on the project wiki and is only discoverable by consortium members with a wiki account. Search keywords: N/A Versioning: Yes, through scheduled website backups. Metadata creation: N/A.
Making data openly	Data openly accessible: No. The data will only be shared internally in AI4Media since it

accessible	contains personal information. How it will be accessible: Restricted access. The data is accessible only by wiki users. Methods/software tools to access data: Web browser. Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data is stored on the project wiki, which is on a dedicated web server hosted in CERTH's premises. The wiki web site uses for its domain an SSL certificate enabling the SHA256RSA signature algorithm and forces all visits to use HTTPS to ensure the traffic is secure. The wiki is restricted only to registered users while registration is possible only by invitation. Access requires username/password authentication. CERTH fully complies with the applicable national, European and International framework, and the GDPR. The wiki uses a file-based RDBMS to enhance security. Web server and file-based DB are running on a Linux encrypted partition, which conforms to the data-at-rest GDPR guidelines. Moreover, state-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk. Regular rolling daily backups are scheduled to minimize the risk of data loss. The data will be preserved there for three years after the end of the project and will then be deleted.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Dataset includes personal data (name and email) of associate members and will thus not be shared. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

6.5.2 AI4Media newsletter subscribers dataset

DMP component	**AI4Media_Data_81_WP11_Text_NewsletterSubscribers _v1** **Partner: LOBA**
Data Summary	Purpose: This dataset contains contact information including name and email, from people subscribing to the AI4Media newsletter. The collected data is necessary for distributing the newsletters to the subscribers and disseminating the project in the context of WP11. Type/format: text Re-use of existing data: No. Data origin: Data provided from people interested in subscribing to the projects' newsletter, through the subscription form available in the website (https://ai4media.eu/newsletters/) Expected size: < 1MB Data utility: This data is necessary for facilitating communication with the subscribers.
Making data findable, incl. provisions for metadata	Is data discoverable: No, data are not discoverable from outside. The data is stored on Zoho Campaigns, the email marketing software used by LOBA for managing and distribution of newsletters. The data is only accessed by LOBA. Search keywords: N/A Versioning: Yes, through scheduled website backups. Metadata creation: N/A.
Making data openly accessible	Data openly accessible: No. The data will only be shared internally in AI4Media since it contains personal information. How it will be accessible: Restricted access. The data is accessible only by LOBA, who can share this information with the coordinator if requested. Methods/software tools to access data: Website's back office (word press). Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A. The data will not be shared. Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A

	Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data is stored on Zoho Campaigns, an email marketing software used for distribution of newsletters and email marketing. Zoho Campaigns fully complies with GDPR, from data collection and processing to managing data subject rights. The software handles and processes data, to ensure the additional level of security that GDPR encourages. Data at rest is encrypted using industry-standard AES-256. All customer data is encrypted in transit over public networks using Transport Layer Security (TLS) 1.2/1.3 with Perfect Forward Secrecy (PFS) to protect it from unauthorized disclosure or modification.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Dataset includes personal data (name and email) of subscribers and will thus not be shared. Is informed consent for data sharing and long-term preservation given: N/A (Consent is given by the subscribers to only use their personal data for sending the newsletters)
Other Issues	No

6.5.3 AI4Media website messages dataset

DMP component	**AI4Media_Data_82_WP11_Text_WebsiteMessages_v1** **Partner: CERTH, LOBA**
Data Summary	Purpose: This dataset contains contact information (including name and email and messages) from people sending a message to AI4Media through the website's contact form. The collected data is necessary for replying to messages received through the website in the context of WP11. Type/format: text Re-use of existing data: No. Data origin: Data provided from people sending messages through the contact form available in the website (https://ai4media.eu/contact/) Expected size: < 1MB Data utility: This data is necessary for facilitating communication with the stakeholders.
Making data findable, incl. provisions for metadata	Is data discoverable: No, data are not discoverable from the outside. The data is forwarded to the project's email list (info@ai4media.eu) which only includes the emails from CERTH (project coordinator) and from LOBA (dissemination leader). The data is stored on CERTH's servers. Search keywords: N/A Versioning: Yes, through scheduled website backups. Metadata creation: N/A.
Making data	Data openly accessible: No. The data will only be shared internally in AI4Media since it

openly accessible	contains personal information. How it will be accessible: Restricted access. The data is accessible only by CERTH and LOBA. Methods/software tools to access data: Website back office (word press) Repository: N/A Restrictions on access: N/A
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: N/A Usable by third parties after end of project N/A Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The data is stored on CERTH's servers. Access requires username/password authentication. CERTH fully complies with the applicable national and European data protection frameworks & guidelines, including the GDPR. State-of-the-art IT security measures and institutional policies mitigate the risk of illegitimate access, including firewalls (to prevent illegitimate access from outside) and a rights-based-file access system (to prevent illegitimate access from inside).
Ethical aspects	Possible ethical and legal aspects preventing sharing: Dataset includes personal data (name and email) of people sending messages to the project and will thus not be shared. Is informed consent for data sharing and long-term preservation given: N/A
Other Issues	No

7. Conclusions

This DMP identifies the datasets managed by the AI4Media consortium, organized by work package. In the present document, we discuss 70 different research datasets, both pre-existing ones and newly-created within the project. Most of the datasets are or will be made openly available, to the benefit of the broader scientific community. In addition, we describe how we will manage and protect non-research data (12 datasets) collected in the context of WP1, WP8, WP9, WP10 and WP11.

The datasets collected and used in the project include: **media-related datasets** (including videos, audio files, social media posts, user profiles, etc.) that will be employed for the design and development of AI methodologies, algorithms, and tools in the context of WP3-WP6, aiming to support tha seven use cases; **questionnaire data** including questionnaires collected from project partners, associate members and end-users aiming to identify user requirements, provide guidance to technical partners for the development of the AI tools but also assess the effectiveness and impact of the developed tools; **user activity data and software analytics** automatically collected by the AI tools during the use case trial with the purpose of evaluating and improving these tools; **personal data of members of the research, academic, student and business communities** participating in AI4Media educational, outreach and dissemination activities or applying for funding through the AI4Media open calls.

With regard to making the research data FAIR, our approach may be summarized as follows:

- **Findable**: The datasets that will be made publicly available will be uploaded in open repositories like Zenodo but also platforms like AI4EU, thus making this data both easily discoverable and identifiable from the outside. Datasets that will be only used internally by project partners will be stored either on the project wiki or in the partner servers. In both cases, the datasets are internally discoverable and identifiable using simple queries with keywords or filters.
- **Accessible**: In the context of the project, we will try to make publicly available the research datasets created by consortium members in the context of WP3,4, 5, 6, and 8. The data will be shared in open repositories like Zenodo or institutional (open) repositories of partners.

  Many of the data used in the project is already **open data**, made openly available by research organizations or media companies. Since it is already open, as a general policy, we will not re-share them. Sharing some of this data will be handled on a case-by-case basis, and will only be pursued in cases where the data license allows it and re-sharing of the data (in some new form or after some processing) provides some additional benefit to the research community. In any case, we will try to provide open software tools that will allow other researchers to easily crawl and collect data from all open data sources.

  In addition to open data, there is **privately owned data**, usually collected by project partners in the context of other projects or internal processes. Such data has been provided to the project for research purposes and will be only used internally by project partners. Effort will be made to make some of this data available in cooperation with the data owners, as long as there are no legal or ethical issues for their sharing.

Data from surveys addressed to project partners, associate members, and end users of AI tools in the context of the seven use cases will not be made openly accessible (at least most of them) since they may contain sensitive or personal information. Where possible and in case there is added value from their sharing, data will be fully anonymized before being shared. The collected data (whether public, private, or personal) will be analysed and analysis results will be made open as part of public project deliverables or publications that will be available in open repositories like Zenodo.

- **Interoperable**: Effort will be dedicated to making the data interoperable, mainly in the context of WP7 and in order to share it through the AI4EU platform. The data and metadata vocabularies adopted for each data source have been also discussed.
- **Reusable**: Effort will be made to increase the re-use of the data that we plan to make open, through clarifying licenses. Licensing will be examined on a case-by-case basis depending on the dataset. In case of data coming from external sources or in cases where the data comes with a license of its own, the data will be re-shared (where necessary) under the same licence. For other datasets, a CC-BY 4.0 (Creative Commons Attribution 4.0 International License) license will be selected.

Datasets generated within AI4Media will either be openly shared (by uploading them in open repositories) or shared internally among specific partners (stored on the project wiki or the servers of AI4Media partners). Datasets to be openly shared, will be deposited in certified repositories like Zenodo and platforms like AI4EU that have in place strong mechanisms and protocols for **data security** and long-term data preservation. Similar mechanisms are in place in both the project wiki and the partners' servers to ensure data protection.

Finally, addressing **legal and ethics challenges** is an important part of the AI4Media work plan. A legal partner (KUL) forms part of the consortium and dedicated tasks (T1.3 - *Ethical issue management* and Task 4.1 - *Legal and ethical frameworks for trusted AI*) and work packages (WP12 – *Ethics Management*) deal specifically with such issues. Moreover, an Ethical Advisory Board will provide guidance on such issues. The aim is to identify the relevant EU legal and ethical frameworks and relevant requirements and provide guidance to partners on issues of data collection and data privacy and protection. To conform to the GDPR, consent forms and information sheets will be provided to users whose personal data will be collected and processed in the context of the AI4Media use cases.

The AI4Media datasets are evolving. Therefore, the DMP is a **living document** that will keep being updated through the lifetime of the project. This is the initial version of the DMP. The final version, including new datasets, will be provided at the end of the project.

AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu