AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

D1.6

Final 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

info@ai4media.eu www.ai4media.eu

Deliverable title	Final Data Management Plan
Deliverable number	D1.6
Deliverable version	1.0
Previous version(s)	-
Contractual date of delivery	31 August 2024
Actual date of delivery	26 August 2024
Deliverable filename	AI4Media_D1.6_Final_Data_Management_Plan_final.docx
Nature of deliverable	ORDP: Open Research Data Pilot
Dissemination level	Public
Number of pages	289
Work Package	WP1
Task(s)	T1.1
Partner responsible	CERTH
Author(s)	Filareti Tsalakanidou (CERTH), Yiannis Kompatsiaris(CERTH), Vasileios Mezaris (CERTH), Symeon Papadopoulos (CERTH)
Editor	Filareti Tsalakanidou (CERTH)
EC Project Officer	Evangelia Markidou

Abstract	This deliverable presents the final Data Management Plan (DMP) of AI4Media and offers an update of the initial DMP (D1.2) that was delivered in M6. The final DMP summarises the strategy for data management within the AI4Media project and provides a detailed description of the datasets that have been collected, processed or generated within the project. It describes the handling of data during and after the project lifetime and discusses how they are 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. The report discusses 166 datasets: 61 research datasets created by project partners within the project, 81 research datasets of third-parties used in the project's research activities, and 24 non-research datasets collected within the project.

Keywords	Artificial intelligence, data management, data management plan, 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 Papadopoulos	CERTH
Evlampios Apostolidis	CERTH
Spiros Baxevanakis	CERTH
Noémie Krack	KUL
Lidia Dutkiewicz	KUL
Nicu Sebe	UNITN
Marco Formentini	UNITN
Lorenzo Seidenari	UNIFI
Ioannis Pitas	AUTH
Ioannis Patras	QMUL
Ioannis Maniadis Metaxas	QMUL
Anna Hansen	UvA
Matthew Barthet	UM
Henning Müller	HES-SO
Riccardo Fratti	HES-SO
Adrian Popescu	CEA
Sven Becker	FhG-IAIS
Milica Gerhardt	FhG-IDMT
Lucile Sassatelli	UCA
Lucia Vadicamo	CNR
Giuseppe Amato	CNR
Fabrizio Sebastiani	CNR
Adrian Tormos	BSC
Enrique Lopez	BSC
Dario Garcia Gasulla	BSC
Victor Bros	IDIAP
David Alonso del Barrio	IDIAP
Remi Mignot	IRCAM
Werner Bailer	JR
Danae Tsabouraki	ATC
Rasa Bocyte	NISV
Birgit Gray	DW
Chaja Libot	VRT
Alberto Messina	RAI
Fulvio Negro	RAI
Maurizio Montagnuolo	RAI
Roberto Iacoviello	RAI
Angelo Bruccoleri	RAI
Stefano Scotta	RAI
Lorenzo Canale	RAI

Samuel Almeida	F6S
Ellie Shtereva	F6S
Candela Bravo	LOBA
Christoffer Holmgard	MODL
Angel Spasov	IMG
Carmen Mac Williams	GAR

Peer Reviews

NAME	ORGANISATION
Henning Müller	HES-SO
Danae Tsabouraki	ATC

Revision History

VERSION	DATE	REVIEWER	MODIFICATIONS
0.1	29/04/2024	Filareti Tsalakanidou, Yiannis Kompatsiaris	First draft sent to partners for contributions
0.2	11/07/2024	Filareti Tsalakanidou	Updated version including inputs from all partners (new datasets added in sections 4, 5, 6)
0.3	15/07/2024	Filareti Tsalakanidou	Updated version including additional datasets
0.4	15/07/2024	Filareti Tsalakanidou	Ready for internal review
0.5	23/08/2024	Henning Müller, Danae Tsabouraki	Internal review
0.6	26/08/2024	Filareti Tsalakanidou	Updated version based on internal review comments
1.0	26/08/2024	Filareti Tsalakanidou, Yiannis Kompatsiaris	Final version

Table of Acronyms and Abbreviations

Acronym	Meaning
1D, 2D, 3D, 4D	One-, two-, three-, four-dimensional
2FA	2 Factor Authentication
AI	Artificial Intelligence
AIDA	International AI Doctoral Academy
AIoD	AI-on-Demand Platform
API	Application Programming Interface
ASV	Automatic Speaker Verification
AWS	Amazon Web Services
BSD-2	Berkeley Software Distribution 2
CC	Creative Commons
CLEF	Conference and Labs of the Evaluation Forum
CoMA	Convolutional Mesh Autoencoders
CSV	Comma-Separated Values file format
DB	Database
DFDC	DeepFake Detection Challenge
DFDM	DeepFakes from Different Models dataset
DFEW	Dynamic Facial Expression in-the-Wild dataset
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
FaVCI2D	Face Verification with Challenging Imposters and Diversified Demographics dataset
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
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

Acronym	Meaning
	Electrotechnical Commission
IT	Information Technology
JAMS	JSON Annotated Music Specification
JSON	JavaScript Object Notation
JWT	JSON Web Token
KoDF	Korean DeepFake (dataset
LLM	Large Language Model
LM	Language Model
MAD-TSC	Multilingual Aligned Dataset for Target-dependent Sentiment Classification
MED	Multimedia Event Detection
MRI	Magnetic Resonance Imaging
MSD	Million Song Dataset
N/A	Non applicable
NEC	Non European Countries
NEFER	Neuromorphic Event-based Facial Expression Recognition dataset
NGO	Non-Governmental Organization
NIR	Near InfraRed
NLP	Natural Language Processing
NN	Neural Network
OCR	Optical Character Recognition
ODSS	Open Dataset of Synthetic Speech
OTP	One Time Password
PII	Personally Identifiable Information
POPD	Protection of Personal Data
RDBMS	Relational DataBase Management System
RGB	Red Green Blue
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
SR	Super Resolution
SSH	Social Sciences and Humanities
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
WAV	Waveform Audio file format
WDF	WildDeepFake dataset

Acronym	Meaning
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

1. Executive Summary

Various datasets have been 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 Description of Action (DoA). A variety of data (videos, images, audio, text, social media, user profile data, questionnaires, contact info, etc.) have been 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 the seven use cases;
- assess the effectiveness of the developed AI4Media technologies in a series of evaluation activities involving end-users;
- establish and support the International AI Doctoral Academy (AIDA); and
- establish the 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 were collected at different stages of the project:

- First, the leaders of the seven AI4Media use cases collected the **use case requirements**. This involved feedback from end users via questionnaires, interviews, focus groups and other similar methods, aiming to identify user needs with regard to the functionalities of the AI tools to be integrated in each use case. The objective of collecting such data was to define the user and use case requirements and guide the technical development of the AI4Media tools.
- In order to **develop novel AI methodologies and tools** to advance AI research for the media and also support the seven AI4Media use cases, multiple datasets (existing or new ones) were used or created by technical partners in the context of WP3, WP4, WP5 and 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 organisation and moderation), heterogeneous datasets were used or generated, including **video, images, audio, text, social media posts, user activity data,** etc. To this end, several existing datasets were deployed, either owned by AI4Media technical and use-case partners or openly shared by third parties (usually academic or research institutions). In addition, several new datasets were also 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 was 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 and use case demonstrators.

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

- Personal data of lecturers, researchers and post-graduate students collected as part of registration processes, course attendance or instruction, and curation of educational resources in the context of the **International AI Doctoral Academy** (AIDA).
- **Applicant data** collected by the platform used for submitting applications to get funding via the two AI4Media open calls. Data for the selected **sub-granted projects** and information of **internal experts and evaluators** involved in this process was also gathered.
- Information about AI4Media's **Associate Members** collected to facilitate their involvement in the project and also about researchers contributing to the **AI Media Observatory.** Also, contact information of people participating in AI4Media's events, workshops or other **dissemination & communication 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 used/collected/generated in 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. An initial version of the DMP (D1.2) was submitted at the beginning if the project, in M6, reflecting progress and expectations at that point in time. The initial DMP offered information on the general data management policy to 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 answered explicitly the questions listed there. Moreover, it listed the datasets that had already been collected or used by the partners at that point, as well as datasets that project partners foresaw to use or generate during the course of the project. For each dataset, explicit information was 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 final version of the DMP offers an update of the initial plan and extends the initial list of research and non-research datasets by adding several new datasets that were used, collected or generated in the project between M7 and M48. The initial version of the DMP discussed 82 datasets in total: 70 research datasets (19 created within the project by project partners and 51 third-party research datasets used by project partners) as well as 12 non-research datasets. **This final version of the DMP presents 166 datasets: 142 research datasets** (**61 created within**

the project by project partners and 81 third-party research datasets used by the partners) **and 24 non-research datasets**.

This deliverable was 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 activities and deliverables.

2. Introduction

This deliverable presents the final Data Management Plan (DMP) of the AI4Media project and aims to provide a detailed description of the datasets that were collected, processed or generated during the course of the 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 information about which datasets are or will be openly accessible and how they are or 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 was adopted by the AI4Media consortium with regard to all the datasets that were used in or generated by the project. It reflects the consortium's comprehensive approach towards data management. The DMP is a living document which evolved during the lifetime of the project. An initial version (D1.2) was delivered in M6 while this final version (D1.6) is delivered at the end of the project, in M48, and reflects the following changes: i) an update of the gereral data management strategy of the project (section 3) and especially the data summary (section 3.1), data security (section 3.7), and ethical & legal aspects (section 3.8); ii) the collection, creation and use of new datasets that were not included or foreseen in the initial DMP; more specifically, D1.2 presented 82 (research and non-research) datasets while D1.6 presents 166 datasets in total (sections 4-6).

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. 61 research datasets are presented in total.

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 the partners of WPs 3, 4, 5, and 6 to develop and test new AI methodologies, algorithms and tools. 81 third-party research datasets are presented in total.

Section 6 describes how we handle non-research data collected within AI4Media, e.g. data collected from open calls, in the context of AIDA, from dissemination activities, etc. 24 non-research datasets are presented in total.

Finally, **Section 7** concludes the deliverable.

3. Data management methodology

The methodological approach that has been used for the compilation of both D1.2 and D1.6 follows the "*Template for HORIZON 2020 DATA MANAGEMENT PLAN (DMP)*", version 1.0, released on 13.10.2016 by the EC. 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.

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 is achieved through six AI4Media pillars:

- The **European Media AI Observatory**, which sets and maintains 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 AIoD platform (WP3, 4, 5, 6).
- A **portfolio of use-cases**, aiming to provide direct application of the AI4Media technologies developed within WP3-WP6 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 (AIDA)**, which will nurture a new generation of PhD talent and ensure young skilled researchers remain in Europe (WP9).
- The **AI4Media Virtual Center of Excellence**, which will function as a portal and network nexus for all Media AI research and innovation activities in Europe (WP11).

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

- **Use case requirements data** (in the form of questionnaires, interviews, focus groups, etc.) is collected by 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 seven AI4Media use cases:
    - *AI for Social Media and against Disinformation*: 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;
    - *AI in Vision*: AI algorithms and tools for high quality video production and video content automation in TV production;
    - *AI for Social Sciences and Humanities*: AI 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 tools to advance game design and development process, focusing on i) improved music analysis and synthesis for games, and ii) game testing and quality assurance;
    - *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 and AI technologies to effectively

organize vast digital archives and collections of images and videos.

Details on these datasets can be found in section 4.64.5.5. Compliance with the legal obligations for processing personal data of users in user research activities are ensured. For more information about this, please have a look at Ethics deliverables D12.1 and D12.2 and the Ethics Management Plan (D1.3, D1.5).

- **Evaluation data** is collected from the end users of the proposed AI solutions in the context of the use cases and 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 feedback 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 4.6.
- **Media-related datasets** (existing or new ones) are used or collected by technical partners in the context of WPs 3, 4, 5, 6 in order to develop and test novel AI algorithms and methods for the media, with a focus on the seven AI4Media use cases. To develop algorithms and tools to address the different needs of each use case, heterogeneous datasets are used or generated, including video, images, audio, text, social media posts, user profiles, user/user group activity data, etc. Several existing datasets are also 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 were also created and used, e.g., datasets including tweets for disinformation detection or datasets including news items for local news analysis. Details on these datasets can be found in sections 4.2-4.5.
- **User survey data** is 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 policy recommendations. Details on these datasets can be found in section 4.1.
- **Personal data** and other information collected from researchers contributing to the AI Media Observatory (WP2). Details on these datasets can be found in section 6.2.
- **Personal data (e.g., name, email, affiliation, etc.) of lecturers, researchers and students** is collected in the context of the AIDA (WP9) as part of their registration process and course attendance (for students), course or lecture delivery (for lecturers) and curation of educational resources (researchers/curators). Details on these datasets can be found in section 6.5.
- **Data of applicants** (individual researchers and organizations) that used the platform set up by F6S in the context of WP10 to submit applications to get funding from **AI4Media's two open calls**. Data for the selected sub-granted projects and information of internal experts and evaluators is also gathered. Details on these datasets can be found in section 6.6 .
- **Personal data of participants in AI4Media's dissemination and community building events** (WP11, WP7) were also collected where necessary to allow better organization of these events and better services to attendees. Data of associate members are also

collected in the context of WP11 via online questionnaire forms. Details on these datasets can be found in sections 6.7 and 6.3.

- **Personal data of members of the consortium** to facilitate project management and coordination. Details on these datasets can be found in section 6.1.

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

As mentioned above, the project uses different types of data (video, images, audio, social media posts, system log data, questionnaires, applicant personal data, etc.) from various sources and also generated datasets including data of the aforementioned types. The data is available 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 WPs 3, 4, 5, 6) 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 (e.g. Twitter, etc.).
- Web (e.g., online 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 of WPs 3,4,5,6; and assessing the effectiveness of these tools in media use cases involving end users. Also, it is useful for the better design, organization, and execution of activities related to AIDA, AI Media Observatory, 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, deepfake detection, 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 are created within AI4Media and are made publicly available by partners are uploaded to open repositories like Zenodo or GitHub, thus making this data both easily discoverable and identifiable from the outside (since they are assigned a DOI). These datasets are also shared through the AIoD platform. Some datasets are 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 partners' servers. This data will only be discoverable and identifiable from registered users, using simple queries with keywords. Other internal data such as user requirement data and evaluation data are stored on user partners' servers and access is provided only to selected institutional users involved in the processing of this data.

What naming conventions are followed?
A specific naming convention is 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*[1]*, 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 evaluation of the use case 3 demonstrator can be named *AI4Media_Data_47_WP8_QUESTIONNAIRE_UseCase3-UserReqCollection2021_v1*. For a third-party dataset which is named by its creators SumMe[2] and includes videos and relevant annotation data, we can use the following name: *AI4Media_Data_91_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.

Are clear version numbers provided?
For datasets that are 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 are shared via open repositories, the metadata standards used by these repositories will be used.

[1] *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.
[2] https://gyglim.github.io/me/vsum/index.html#benchmark

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), several of these datasets are already publicly shared on open repositories like Zenodo. They can also be accessed through links available on the project website[3].

In addition to open data, there are also privately owned datasets. Most of these datasets are owned by the media companies and news organizations involved in AI4Media as well as by

[3] AI4Media open datasets: https://www.ai4media.eu/open-datasets/

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, but 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 is 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 4.5.5), end users (to evaluate the AI4Media technologies, see section 4.5.5), and other parties (e.g. applicant data requesting AI4Media funding, see section 6.6), 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) are processed and analysed aiming to achieve the project objectives. Where appropriate, the analysis results are 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 GitHub or GitLab. The datasets will also be shared through the AIoD platform. Some datasets will also be shared on AI4Media partners' institutional repositories. Links to all open datasets are included in the project website in a dedicated page. 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 are required to access the data depending on the dataset as is 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 were 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 AIoD 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, 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 focuses on uploading AI4Media resources (including datasets) on the AIoD platform (see 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 try to collect existing and new data in standardized formats, following well-known data representation models and metadata vocabularies. Standard data vocabularies are adopted for different types of datasets (social media data, audiovisual data, user analytics, etc.) while a common approach is used for metadata creation based on OpenAIRE guidelines[4]. More information can be found in sections 4 and 5.

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 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 is examined on a case-by-case basis.

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?

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

With regard to the individual questions about increasing data re-use, our general 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 is 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 selected wherever possible. This license 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 is examined on a case-by-case basis (see sections 4, 5 and 6). In general, effort is 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 is examined on a case-by-case basis (see sections 4, 5 and 6). The datasets that we openly share will be re-usable after the end of the project through the AoD platform and open repositories like Zenodo and GitHub.

How long is it intended that the data remains re-usable?
This is examined on a case-by-case basis (see sections 4, 5 and 6). The datasets that we 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 are 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 is performed by partners to ensure consistency of replies. These quality assurance processes, including data cleaning are discussed in sections 4, 5 and 6.

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 AIoD 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 is responsible for closely monitoring the execution of the data management plan and ensuring that project partners handle project datasets appropriately. The DPO works in the central administration of CERTH and has significant experience in data management and data protection as the DPO of many Horizon and 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 and the Ethics Managemnt Plan (D1.3, D1.5) 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 principal, data used in or generated by the project will be preserved in 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 like Zenodo 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 are designed to safeguard collected data against unauthorized use and to comply with all national and EU regulations. Engineering best practices and state-of-the-art data security measures are incorporated as well as GDPR considerations, and respective guidelines and principles.

As explained above, AI4Media datasets are either be openly shared (by uploading them in open repositories) or shared internally among specific partners (e.g. stored in 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 AIoD platform.

AI4Media wiki

The data is stored on the project wiki[5], 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

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

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 EU 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 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 EMEA SARL Greek Branch, 59-61 Agiou Konstantinou St., Marousi, 15124, Athens, Greece: Storage for all multimedia files uploaded and hosted by the platform.

- Heroku, Inc., 1 Market St., Suite 300, San Francisco, CA 94105, United States: Hosting of backend processes.
- MongoDB Limited, Building Two, Number One Ballsbridge, Ballsbridge, Dublin 4, Ireland: Platform data storage.

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) is part of the consortium, and provided guidance and relevant expertise throughout the project, including

presentations on data protection and privacy and a background note on processing personal data for scientific research purposes. The background notes explored the ethical and legal consideration of dataset re-use. In addition, the organizational structure of AI4Media includes an external Ethics Advisory Board. This Board advises on ethics issues as well as on the data processing procedures adopted in AI4Media and offers expertise on the matter.

A dedicated task deals specifically with such issues: Task 1.3 "*Ethical issue management"*, led by KUL. T1.3 delivered an initial Ethics Management Plan (D1.3) in M12 and the final Ethics Management Plan (D1.5) in M48. The Ethics Management Plan includes 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. It also includes elements reporting on KUL's online session focusing on ethics impact assessments in the context of technology development. The session provided an overview of various types of impact assessments, shared practical challenges, and offered tips for conducting them effectively.

Moreover, the following tasks provided analysis of relevant legal frameworks (GDPR, AI Act, Digital Services Act, Data Governance Act, Data Act) and offered clarifications with regard to AI ethics, privacy and data governance. Ethics considerations related to the use of data in Generative AI were also addressed in:

- Task T2.1 *"Analysis of the EU policy on AI and the forthcoming Commission's legislative proposal on AI regulation*" as part of deliverable D2.1;
- Task 4.1 "*Legal and ethical frameworks for trusted AI*" as part of deliverables D4.3 and D4.4.

In addition to the aforementioned tasks, WP12 "*Ethics requirements*" also deals with specific ethics and legal requirements. These included 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*).

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 processed personal data collected during the project or personal data and anonymized data collected during other previous professional experiences external to the project. In other words, partners of the consortium had in some case processed data already stored in 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 EU'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 were used to safeguard the personal data processed. Where relevant, personal data were pseudonymised according to the current state of the art, and the additional information necessary for re-identifying the individual was kept separately (according to Art. 4(5) of the GDPR). Pseudonymisation is expressly mentioned as a measure within Art. 89 GDPR.

Collected personal data 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 respected and fully complied 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 all the personal data processing activities reported by partners in the final Ethics Management Plan (D1.5).

Moreover, for personal data processed in the context of the AI4Media use cases, end users were provided with informed consent forms and information sheets when personal data had 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 was 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, transfer of data, including personal data, between EU and non-EU countries was also at stake (this issue is 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 involve these

partners include the import/export of personal data to these countries. As regards Switzerland, the European Commission has 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.[6] In January 2024, the EC published its report on the first review of the functioning of the eleven adequacy decisions.[7] The report investigated if these adequacy decisions stood the test of time and address new developments and challenges. Based on the report findings, the EC concludes that Switzerland continues to provide an adequate level of protection for personal data transferred from the EU.

As regards the United Kingdom, after the Brexit, the European Commission has also concluded an adequacy decision with the UK.[8] "Personal data can now flow freely from the European Union to the United Kingdom where it benefits from an essentially equivalent level of protection to that guaranteed under EU law."[9]

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 participate in use case activities like evaluation were 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 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 sought informed consent of all research participants. In order to obtain informed consent, the consortium provided prospective participants sufficient opportunity to consider whether or not to participate and this under circumstances that minimise the possibility of coercion or undue influence. 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

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

[7] Report on the first review of the functioning of the adequacy decisions adopted pursuant to Article 25(6) of Directive 95/46/EC, https://commission.europa.eu/document/f62d70a4-39e3-4372-9d49-e59dc0fda3df_en

[8] Commission Implementing Decision (EU) 2021/1772 of 28 June 2021 pursuant to Regulation (EU) 2016/679 of the European Parliament and of the Council on the adequate protection of personal data by the United Kingdom (notified under document C(2021)4800) (Text with EEA relevance), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021D1772

[9] https://ec.europa.eu/commission/presscorner/detail/%20en/ip_21_3183

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 was 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 are used solely for the research purposes of AI4Media, are not 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 is 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[10] 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 was 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 gave 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 was entirely voluntary and the participants had the right to withdraw from the research at any time without any adverse consequences.

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.

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

4. Data management plan for research datasets created within AI4Media

This section presents the research datasets that partners of the AI4Media consortium created during the project lifetime. An initial list of 19 datasets was presented in the initial DMP (D1.2) delivered in M6. This list has since been updated to include additional datasets created or collected between M7 and M48. **61 datasets in total are presented in this section.**

In the following sub-sections, we present the DMP plan for the different research datasets that were 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 (Table 1). It is the same template also used in D1.2. 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 its relation to project objectives) in the context of AI4Media? 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)

In the initial DMP, 19 research datasets were presented. Since then the number of datasets created within AI4Media has risen to 61 datasets, 42 more than in D1.2. **34 out of these 61 datasets are openly shared**. The remaining 27 datasets cannot be openly shared due to

various ethical and legal aspects that prevent/prohibit sharing, examined on a dataset-by-dataset basis (e.g. some social media data cannot be shared due to the platform's terms of use, use case evaluation data cannot be shared because they usually include personal data of end users, datasets of news articles cannot be shared because the copyright is owned by the relevant media organisations, etc.)

Below, we provide a Table that briefly summarizes the 61 datasets created within AI4Media and offers a glance at the structure of this section and its subsections. New datasets that were not included in the initial DMP are indicated with yellow.

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-impact-policy_v1	WP2	Workshop data for AI technology impact and policy recommendations	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
AI4Media_Data_04_WP3_Video_Mixamo_v1	WP3	Mixamo-Kinetics dataset	4.2.2
AI4Media_Data_05_WP3_Image_100Driver_v1	WP3	100-Driver dataset for distracted driver classification	4.2.3
AI4Media_Data_06_WP3_TEXT_LeQua_v1	WP3	LeQua 2022 datasets	4.2.4
AI4Media_Data_07_WP3_TEXT_ProductReviewsForOrdinalQuantification_v1	WP3	Product reviews for ordinal quantification dataset	4.2.5
AI4Media_Data_08_WP3_TEXT_ProductReviewsDataset_v1	WP3	Product reviews dataset	4.2.6
AI4Media_Data_09_WP3_UCIOpenMLdatasetsOrdinalQuantification_v2	WP3	UCI and OpenML datasets for ordinal quantification	4.2.7
AI4Media_Data_10_WP3_CherenkovTelescopeDataOrdinalQuantification_v2	WP3	Cherenkov telescope data for ordinal quantification	4.2.8
Data collected in WP4 (Explainability, Robustness and Privacy in AI)			4.3
AI4Media_Data_11_WP4_MRI_MultipleSclerosisLesionSegmentation_v1	WP4	White matter multiple sclerosis lesion segmentation datasets	4.3.1
Data collected in WP5 (Content-centered AI)			4.4
AI4Media_Data_12_WP5_Video_ObyGaze12_v1	WP5	ObyGaze12 dataset for detection of visual objectification in films	4.4.1
AI4Media_Data_13_WP5_TEXT_MAD-TSC_v1	WP5	MAD-TSC Multilingual Aligned Dataset for Target-dependent Sentiment Classification	4.4.2
AI4Media_Data_14_WP5_Video_OWFOD_v1	WP5	ÖWF Object Detection dataset	4.4.3
AI4Media_Data_15_WP5_Video_People@Places_v1	WP5	4PeopleAtPlaces dataset	4.4.4

DMP component	WP	Short summary	Relevant sub-section
AI4Media_Data_16_WP5_Video_ToDY_v1	WP5	ToDY dataset for visual time of day and season classification	4.4.5
AI4Media_Data_17_WP5_ImageFeatures_VisioneFeatureRepository_v1	WP5	VISIONE Feature Repository	4.4.6
AI4Media_Data_18_WP5_TEXT_COCO - LVIS-Open ImagesV4-classes Mapping_v1	WP5	COCO, LVIS, Open Images V4 classes mapping	4.4.7
AI4Media_Data_19_WP5_Video_BusViolence_v1	WP5	Bus violence dataset	4.4.8
AI4Media_Data_20_WP5_Image_PestStickyTraps_v1	WP5	Pest Sticky Traps dataset	4.4.9
AI4Media_Data_21_WP5_IMAGE_VWFP_v1	WP5	Virtual World Fallen People dataset	4.4.10
AI4Media_Data_22_WP5_VIDEO_SR-BVI-DVC_v1	WP5	SR_BVI-DVC super-resolution dataset	4.4.11
AI4Media_Data_23_WP5_VIDEO_BSC4K_v1	WP5	BSC4K super-resolution dataset	4.4.12
AI4Media_Data_24_WP5_IMAGE_NDISPark_v1	WP5	Night and Day Instance Segmented Park dataset	4.4.13
AI4Media_Data_25_WP5_TEXT_ArXiv Abstracts_v1	WP5	ArXiv abstracts for authorship analysis dataset	4.4.14
AI4Media_Data_26_WP5_4D_Florence4D_v1	WP5	Florence4D facial expression dataset	4.4.15
AI4Media_Data_27_WP5_EventRGB_NEFER_v1	WP5	Neuromorphic Event-based Facial Expression Recognition dataset	4.4.16
AI4Media_Data_28_WP5_Video_PEM360_v1	WP5	PEM360 dataset of 360° videos	4.4.17
AI4Media_Data_29_WP5_VIDEO-TEXT_VRT-Sum_v1	WP5	VRT-Sum video summarization dataset	4.4.18
AI4Media_Data_30_WP5_MODEL_CA-SUM-models_v1	WP5	CA-SUM pretrained video summarization models	4.4.19
AI4Media_Data_31_WP5_Audio_AudioPhylogeny_v1	WP5	Audio phylogeny dataset	4.4.20
AI4Media_Data_32_WP5_Video_CMM-Doc-dataset_v1	WP5	RAI CMM documentaries dataset	4.4.21
AI4Media_Data_33_WP5_Video_CMM-ANTS-dataset_v1	WP5	RAI CMM-ANTS newscasts dataset	4.4.22
AI4Media_Data_34_WP5_Video_CMM-dataset_v1	WP5	RAI CMM mixed dataset DMP component	4.4.23
AI4Media_Data_35_WP5_Text_Cross-lingual-news_v1	WP5	Cross lingual news dataset	4.4.24
AI4Media_Data_36_WP5_Misc_YouTubeDataset_v1	WP5	YouTube RAI channel dataset	4.4.25
Data collected in WP6 (Human- and Society-centred AI)			4.5
AI4Media_Data_37_WP6_SOCIALMEDIA_GreekTwitterPolitics_v1	WP6	Greek Politics Twitter dataset	4.5.1
AI4Media_Data_38_WP6_SOCIALMEDIA_Covid19Twitter_v1	WP6	Covid-19 Twitter dataset	4.5.2
AI4Media_Data_39_WP6_TEXT_TextTwitter_v1	WP6	Twitter text dataset	4.5.3

DMP component	WP	Short summary	Relevant sub-section
AI4Media_Data_40_WP6_TEXT_Covid 19DiscussionTopicsTwitter_v1	WP6	Twitter COVID-19 discussions topics dataset	4.5.4
AI4Media_Data_41_WP6_AUDIO_OD SS_v1	WP6	ODSS Open Dataset of Synthetic Speech	4.5.5
AI4Media_Data_42_WP6_IMAGE_ CelebHQGaze _v1	WP6	CelebHQGaze image dataset for gaze estimation	4.5.6
AI4Media_Data_43_WP6_TEXT_Europ ean-news-covid-vaccination_v1	WP6	European news about Covid vaccination dataset	4.5.7
AI4Media_Data_44_WP6_TEXT_Suiss eRomandeLocalNews_v1	WP6	Suisse Romande local news dataset	4.5.8
AI4Media_Data_45_WP6_TEXT_Lausa nneNews_v1	WP6	Lausanne news dataset	4.5.9
AI4Media_Data_46_WP6_TEXT_Suiss eAllemandeNews_v1	WP6	Suisse Allemande local news dataset	4.5.10
AI4Media_Data_47_WP6_TEXT_YouT ubeReferences_v1	WP6	References on YouTube dataset	4.5.11
AI4Media_Data_48_WP6_TEXT_Politi calBarometer_v1	WP6	Political Barometer dataset	4.5.12
AI4Media_Data_49_WP6_TEXT_ElecD eb60To16_v1	WP6	ElecDeb60To16-fallacy dataset	4.5.13
Data collected in WP8 (Use cases & demonstrators in media, society and politics)			4.6
AI4Media_Data_50_WP8_USER-RESEARCH_UseCase1_DW_v1	WP8	Data from user research activities in Use Case 1	4.6.1
AI4Media_Data_51_WP8_USER-RESEARCH_UseCase2_VRT_v1	WP8	Data from user research activities in Use Case 2	4.6.2
AI4Media_Data_52_WP8_QUESTION NAIRE_UseCase3-UserReqCollection2021_v1	WP8	Questionnaires for the collection of user requirements for Use Case 3	4.6.3
AI4Media_Data_53_WP8_QUESTION NAIRE_UseCase3Evaluation_v1	WP8	Questionnaires for the evaluation of Use Case 3	4.6.4
AI4Media_Data_54_WP8_USER-RESEARCH_UseCase4_NISV_v1	WP8	Data from user research activities in Use Case 4	4.6.5
AI4Media_Data_55_WP8_QUESTION NAIRE_UseCase5Evaluation_v1	WP8	Questionnaires for the evaluations of Use Case 5-B	4.6.6
AI4Media_Data_56_WP8_QUESTION NAIRE_VIDEO_UseCase7-UserFeedbackData_v1	WP8	Data from user research activities in Use Case 7	4.6.7
AI4Media_Data_57_WP8_Text_Truly MediaDataset-UseCase1-ATC_v1	WP8	TrulyMedia social media + web dataset	4.6.8
AI4Media_Data_58_WP8_Video_Gam eGlitches-UC5-MODL_v1	WP8	Game glitches dataset for Use Case 5	4.6.9
AI4Media_Data_59_WP8_Audio_RAW compositions_v1	WP8	Musical production for AI co-creation dataset	4.6.10
AI4Media_Data_60_WP8_TEXT_Curre ntAffairsTranscripts-UC4-NISV_v1	WP8	Current affairs transcripts dataset for Use Case 4	4.6.11
AI4Media_Data_61_WP8_ QUESTIONNAIRE_AI-IndustrialNeeds-T8.4_v1	WP8	User survey data for AI industrial needs (for T8.4)	4.6.12

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 was developed by CERTH to collect information on the evolving landscape of AI technologies for the media sector as part of an anonymous public survey. The questionnaire was addressed to both AI researchers working on multimedia AI but also to people working in the media industry or whose work is closely related to this industry (e.g. researchers studying the media, media regulators, people working in relevant NGOs, etc.). The survey aimed to collect their opinions on the benefits, risks, technological trends and challenges of AI use in the media industry as well as their experience on AI strategies and AI skills in media organisations, their insights on the most promising ways to facilitate AI adoption and knowledge transfer and, finally, their perceptions about ethical use of AI. The online survey was filled anonymously by 150 people. The collected questionnaires were then analyzed and the results of this analysis were included in the roadmap for AI technologies and applications in the media sector (D2.3) in the context of T2.3. Type/format: Excel files containing questions and user responses. Re-use of existing data: No. Data origin: Questionnaires filled online by AI researchers, media professionals, media regulators, etc. Expected size: A few MBs 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 survey replies are stored in CERTH's servers. No DOI is assigned. 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 since they are used as part of an internal exercise that facilitated the authoring of D2.3. However, a summary of this data is presented in D2.3, which is available on the project website. 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 CERTH servers, it is only internally accessible by CERTH members. Methods/software tools to access data: Web-browser Repository: CERTH's internal servers. Restrictions on access: Shared with selected project partners upon request. 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 CERTH servers 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 a server in CERTH's premises. Access requires username/password authentication. CERTH fully complies with the applicable national, European and International framework, and the GDPR. 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 Workshop data for AI technology impact and policy

DMP component	AI4Media_Data_02_WP2_SURVEY_AI- impact-policy_v1 Partner: UvA
Data Summary	Purpose: Six workshops have been conducted by UvA with other partners (KUL & NISV) to collect information on the social and economic impact of AI for media technologies in the context of T2.4. The data collected was analyzed and the results of this analysis were used for drafting the white papers on the social, economic, and political impact of media AI Technologies (D2.5) as well as other shorter report formats for dissemination. Type/format: Minutes of workshops (online and in person). Re-use of existing data: No. Data origin: Input/feedback from workshops from consortium partners, associate members and invited external participants. Expected size: A few MBs. Data utility: Only internal for WP2 research. The collected data was used for the

	authoring of D2.5, D2.4 and D2.6 as well as shorter reports and factsheets disseminated on the AI Media Observatory.
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 been presented in D2.5 (and the methodology used) and relevant short reports, which are available on the project website. No personal information will be published. How it will be accessible: Stored in UvA servers. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: Shared among UvA, KUL and NISV
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 UvA's servers that use state-of-the-art IT security measures and company policies to mitigate the risk of illegitimate access while also complying with the GDPR.
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, UPB**

Data Summary	Purpose: This dataset ("Face Verification with Challenging Imposters and Diversified Demographics 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 is available at https://github.com/AIMultimediaLab/FaVCI2D-Face-Verification-with-Challenging-Imposters-and-Diversified-Demographics Search keywords: N/A Versioning: GitHub supports versioning Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is 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 interoperable	Interoperability: The file structure makes the use of the dataset easy. 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 UPB's servers. UPB 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

4.2.2 Mixamo-Kinetics dataset

DMP component	AI4Media_Data_04_WP3_Video_ Mixamo _v1 Partner: UNITN
Data Summary	Purpose: The first large scale dataset for benchmarking domain adaptation methods for action recognition in the challenging task of transferring knowledge from the synthetic to the real domain. Mixamo-Kinetics is used in WP3 (T3.3) for video domain adaptation. Type/format: mp4 Re-use of existing data: No, the dataset is created within AI4Media. Data origin: The dataset comprises 36,195 videos, divided into 14 action categories and two domains, i.e., the source domain (synthetic videos generated from Mixamo) and the target domain (real videos from Kinetics (https://github.com/cvdfoundation/kinetics-dataset). Expected size: ~80 GB Data utility: It is useful to AI4Media partners and other researchers that want to perform video domain adaptation.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available at https://drive.google.com/drive/folders/1vGeWBXO5hcFlEll-PLw6zbg-5QSauQGi?usp=sharing Search keywords: N/A Versioning: GitHub supports versioning Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is openly accessible via GitHub at https://github.com/vturrisi/CO2A

	How it will be accessible: The data can be downloaded from an online archive. Methods/software tools to access data: N/A Repository: GitHub Restrictions on access: N/A
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: The dataset comprises 36,195 videos, divided into 14 action categories and two domains, i.e., the source domain (synthetic videos from the Mixamo dataset) and the target domain (real videos from the Kinetics dataset). THe source dataset (Mixamo) consists of 24,533 videos synthetically generated using the 3D characters from Mixamo. The dataset comprises videos depicting actions performed by 6 distinct avatars, with different backgrounds, camera positions and random 3D objects in the scene. Also, key-points are provided for each character following the scheme from the COCO dataset but without the keypoints for eyes and ears. Each frame is generated with a resolution of 512 by 512 and the mean length of the videos is 138 frames. The target dataset (Kinetics) was created considering 11,662 videos from 14 action categories extracted from the Kinetics dataset. The overlapping actions between the source and the target datasets are swing dancing, breakdancing, salsa dancing, throwing, capoeira, jogging, shouting, side kick, clapping, texting, golf putting, squat, punching and backflip. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data and the code are 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 is shared in a google drive repository.
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

4.2.3 100-Driver dataset for distracted driver classification

DMP component	AI4Media_Data_05_WP3_Image_100Driver _v1 Partner: UNITN
Data Summary	Purpose: 100-Driver is a large-scale, diverse posture-based distracted driver dataset. 100-Driver involves different types of variations that closely meet real-world applications, including changes in the vehicle, person, camera view, lighting, and

	modality. It includes 4 settings for investigating practical problems of DDC, including the traditional setting without domain shift and 3 challenging settings (i.e., cross-modality, cross-view, and cross-vehicle) with domain shifts. Type/format: mp4. Re-use of existing data: No, the dataset is created within AI4Media. Data origin: The dataset comprises more than 470K images taken by 4 cameras observing 100 drivers over 79 hours from 5 vehicles. Expected size: ~60 GB Data utility: It is useful to AI4Media partners to perform behavior understanding in different driving domains.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available at https://100-driver.github.io/ Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is openly accessible at https://100-driver.github.io/ How it will be accessible: The data can be downloaded from an online archive after completing a licensing form. Methods/software tools to access data: N/A Repository: https://100-driver.github.io/ 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: During data collection, we elaborately controled the diversity of raw data in terms of vehicles (5 vehicles, Mazda 3 axela, Lynk&co 03, Toyota C-HR, Hyundai X25 and Ankai A6), camera locations (4 Xiaomi-C1 cameras in front-left, front, front-right, and side-right, modalities (RGB and NIR), lighting conditions (from morning to afternoon, from spring to winter, and different weather conditions), drivers (100 participants), appearance variations (changing clothes, wearing mask, hat and sunglasses). The RGB modality has been captured in daytime while the Near Infrared (NIR) modality was collected in nighttime. To ensure the appearance variations, a part of the drivers (25% in daytime and 15% in nighttime) were recorded over multiple time periods, leading to huge appearance variations especially in clothes and lighting. Following the collection setting, we initially obtained 79.34 hours of video. The overall annotation process was conducted by 20 experts. To boost the efficiency of data annotation, we first grouped the data by drivers. In addition, we aligned the start and end times for the 4 cameras of the same driver, so that we could label each predefined class for all 4 cameras at once based on the timestamp. Each individual behaviour is labelled with behaviour class, modality type, driver ID, camera ID, vehicle ID and scene ID. Given the labelled video clips, we conducted a down sampling to generate more diverse data considering the high similarity between adjacent frames. We further removed outliers with very different content from the labelled class. We finally produce a total of 470,208 samples to form our 100-Driver dataset.

	Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data is protected by copyright.The user acquires no ownership, rights, or title of any kind in all or any parts of the dataset. Terms of use clarified at https://100-driver.github.io/img/100_Driver_Licensing.pdf Availability for re-use: N/A Usable by third parties after end of project: Data already publicly shared for research purposes. 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 is available on request.
Ethical aspects	Possible ethical and legal aspects preventing sharing: During collection, we equipped a safety officer giving relevant instructions according to the road condition. Each participant was informed of the risks involved in data collection and has signed a GDPR informed consent to allow the data to be publicly available for research study. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.2.4 LeQua 2022 datasets

DMP component	**AI4Media_Data_06_WP3_TEXT_LeQua_v1 Partner: CNR**
Data Summary	Purpose: The aim of LeQua 2022 (the 1st edition of the CLEF "Learning to Quantify" lab) is to allow the comparative evaluation of methods for "learning to quantify" in textual datasets, i.e., methods for training predictors of the relative frequencies of the classes of interest in sets of unlabelled textual documents. These predictors (called "quantifiers") will be required to issue predictions for several such sets, some of them characterised by class frequencies radically different from the ones of the training set. LeQua 2022 offers two tasks (T1 and T2), each admitting two subtasks (A and B): T1A: This task is concerned with evaluating binary quantifiers; T1B: This task is concerned with evaluating single-label multi-class quantifiers; T2A: Like Task T1A, this task is concerned with evaluating binary quantifiers. Unlike in Task T1A, participants will be provided with the raw text of the documents; T2B: Like Task T1B, this task is concerned with evaluating single-label multi-class quantifiers; like in Task T2A, participants will be provided with the raw text of the documents. Type/format: Raw text and already processed vector formats Re-use of existing data: Yes, we use an existing dataset Data origin: https://jmcauley.ucsd.edu/data/amazon/

	Expected size: 7.1GB Data utility: Benchmark for learning to count tasks used in experiments in T3.7.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available online (https://zenodo.org/records/6546188) and indexed in Google Search keywords: quantification, prevalence estimation Versioning: Yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: Available on Zenodo https://zenodo.org/records/6546188 Methods/software tools to access data: Web browser Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: Text data is in simple plain text format. Vector data is in csv format. No metadata. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: Creative Commons Attribution 4.0 International Availability for re-use: Data is already available online. Usable by third parties after end of project: Yes 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 Zenodo servers. Zenodo 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

4.2.5 Product reviews for ordinal quantification dataset

DMP component	**AI4Media_Data_07_WP3_TEXT_ProductReviewsForOrdinalQuantification_v1 Partner: CNR**
Data Summary	Purpose: This data set comprises a labeled training set, validation samples, and testing samples for ordinal quantification. The goal of quantification is not to predict the class

	label of each individual instance, but the distribution of labels in unlabeled sets of data. The data is extracted from the McAuley data set of product reviews in Amazon, where the goal is to predict the 5-star rating of each textual review. We have sampled this data according to three protocols that are designed for the evaluation of quantification methods. Used for experiments in T3.7. Type/format: Text processed vector format by means of a Large Language Model (Roberta) Re-use of existing data: Yes, we use an existing dataset Data origin: https://jmcauley.ucsd.edu/data/amazon/ Expected size: 41.9GB Data utility: Benchmark for Ordinal Quantification tasks used in T3.7.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available online (https://zenodo.org/records/8405476) and indexed in Google. Search keywords: ordinal quantification, prevalence estimation Versioning yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: Available on Zenodo at https://zenodo.org/records/8405476 Methods/software tools to access data: Web browser Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: Text data is represented in vector format, with one vector per line, comma-separated values. No metadata. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: Creative Commons Attribution 4.0 International Availability for re-use: Data is already available online. Usable by third parties after end of project: Yes 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 Zenodo servers. Zenodo 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

4.2.6 Product reviews dataset

DMP component	AI4Media_Data_08_WP3_TEXT_ProductReviewsDataset_v1 Partner: CNR
Data Summary	Purpose: This data set comprises a labelled training set used in the experimentation of the paper "Binary Quantification and Dataset Shift: An Experimental Investigation". The data is extracted from the McAuley data set of product reviews on Amazon. Type/format: Raw text partitioned in samples. A file per sample Re-use of existing data: Yes, we use an existing dataset Data origin: https://jmcauley.ucsd.edu/data/amazon/ Expected size: 5.2GB Data utility: Benchmark for quantification under dataset shift, used for experiments in T3.7
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available online (https://zenodo.org/records/8421611) and indexed in Google. Search keywords: quantification, dataset shift, prevalence estimation Versioning yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: Available on Zenodo at https://zenodo.org/records/8421611 Methods/software tools to access data: Web browser Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: Text data is represented in vector format, with one vector per line, comma-separated values. No metadata. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: Creative Commons Attribution 4.0 International Availability for re-use: Data is already available online. Usable by third parties after end of project: Yes 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 Zenodo servers. Zenodo 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

4.2.7 UCI and OpenML datasets for ordinal quantification

DMP component	AI4Media_Data_09_WP3_UCIOpenMLdatasetsOrdinalQuantification_v2 Partner: CNR
Data Summary	Purpose: This dataset comprises four different labelled datasets, all targeted at the evaluation of ordinal quantification algorithms (i.e., algorithms that estimate the distribution of the class labels in a set of unseen data items, where there is a total order defined on the set of classes). Used for experiments in T3.7. Type/format: The datasets are represented as csv files. Each of the four datasets consists of a set of samples to be used in the experimentation. Each sample is represented as a set of indices into the original UCI or OpenML dataset, so that the sample can be extracted unambiguously from the original dataset. Re-use of existing data: Yes, we use existing UCI and OpenML datasets; what we do is providing scripts to extract, from each of them, sets of samples that are suitable for testing ordinal quantification algorithms. Data origin: https://archive.ics.uci.edu/ml/index.php and https://www.openml.org Expected size: 25.6MB Data utility: These datasets will be useful to anyone wishing to test the accuracy of ordinal quantification algorithms.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is available online (https://zenodo.org/records/8177302) and indexed in Google. Search keywords: Ordinal quantification, dataset shift, prevalence estimation Versioning: Yes (the current version is v0.2.0) Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: The data is and will always be accessible on Zenodo (https://zenodo.org/records/8177302) Methods/software tools to access data: Via a web browser. Detailed instructions for correct use are given at https://zenodo.org/records/8177302. Repository: Zenodo Restrictions on access: No

Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Detailed instructions for correct use are given at https://zenodo.org/records/8177302. Data and metadata vocabularies: Data is simply a set of pointers to individual data items in the original UCI or OpenML datasets, using comma-separated values. No metadata. Use of standard vocabularies: No. Mappings to commonly used vocabularies: No.
Increase data re-use	Licence: Creative Commons Attribution 4.0 International. Availability for re-use: Data is already available online at https://zenodo.org/records/8177302 Usable by third parties after end of project: Yes 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 downloadable from the Zenodo servers. Zenodo 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: No ethical or legal issues that can have an impact on data sharing and that we can think of. Is informed consent for data sharing and long term preservation given: No informed consent for data sharing and long term preservation included in questionnaires dealing with personal data, since no personal data are involved in these datasets.
Other Issues	N/A

4.2.8 Cherenkov telescope data for ordinal quantification

DMP component	AI4Media_Data_10_WP3_CherenkovTelescopeDataOrdinalQuantification_v2 Partner: CNR
Data Summary	Purpose: This dataset is targeted at the evaluation of ordinal quantification algorithms (i.e., algorithms that estimate the distribution of the class labels in a set of unseen data items, where there is a total order defined on the set of classes). Used for experiments in T3.7. Type/format: The dataset is represented as csv files. It consists of a set of samples to be used in the experimentation. Each sample is represented as a set of indices into the original FACT database of telescope observations, so that the sample can be extracted unambiguously from the original dataset. Re-use of existing data: Yes, we use the existing FACT dataset; what we do is providing scripts to extract from it sets of samples that are suitable for testing ordinal quantification algorithms. Data origin: https://factdata.app.tu-dortmund.de/

	Expected size: 46.7MB Data utility: While not a dataset pertaining to media, this dataset will be useful to anyone wishing to test the accuracy of ordinal quantification algorithms, which are indeed very important in the world of media.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is available online (https://zenodo.org/records/8172813) and indexed in Google. Search keywords: Ordinal quantification, dataset shift, prevalence estimation Versioning: Yes (the current version is v0.2.0) Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: The data is and will always be accessible on Zenodo (https://zenodo.org/records/8177302) Methods/software tools to access data: Web browser. Detailed instructions for correct use are given at https://zenodo.org/records/8177302. Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Detailed instructions for correct use are given at https://zenodo.org/records/8172813. Data and metadata vocabularies: Data is simply a set of pointers to individual data items in the original FACT dataset, using comma-separated values. No metadata. Use of standard vocabularies: No. Mappings to commonly used vocabularies: No.
Increase data re-use	Licence: Creative Commons Attribution 4.0 International. Availability for re-use: Data is already available online at https://zenodo.org/records/8172813 Usable by third parties after end of project: Yes 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 downloadable from the Zenodo servers. Zenodo 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: No ethical or legal issues that can have an impact on data sharing and that we can think of. Is informed consent for data sharing and long term preservation given: No informed consent for data sharing and long term preservation included in questionnaires dealing with personal data, since no personal data are involved in these datasets.

Other Issues	N/A

4.3 Datasets collected in the context of WP4

4.3.1 White matter multiple sclerosis lesion segmentation datasets

DMP component	AI4Media_Data_11_WP4_MRI_MultipleSclerosisLesionSegmentation_v1 Partner: HES-SO
Data Summary	Purpose: Segmentation of white matter lesions in Multiple Sclerosis patients, addressing variability in imaging conditions across multiple centers. Used in AI4Media WP4 to promote research in model robustness assessment, interpretability and uncertainty estimation on real-world complex data with a 3D structure. Type/format: 3D MRI scans with T1-weighted and FLAIR contrasts, pre-processed to 1mm isovoxel space. Re-use of existing data: Yes, the dataset reuses scans contributed by partners in another project Data origin: MRI scans collected from multiple clinical centers: Rennes, Bordeaux, Lyon, Ljubljana, Best, and Lausanne. Expected size: - Data utility: Benchmark for uncertainty estimation and model robustness. Used for testing relevant techniques in WP4.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, avalaible in Zenodo. Search keywords: MRI data, Uncertainty Estimation, Robustness, Distributional Shift, Semantic Segmentation, NeuroImaging Versioning: Yes Metadata creation: Metadata follow the Zenodo standards and include information about the imaging centers, scan parameters, and pre-processing steps clearly defined in the reference paper.
Making data openly accessible	Data openly accessible: Partially. The in-domain training and development data will be openly accessible in Zenodo at https://zenodo.org/records/7051658 and https://zenodo.org/records/7051692. The Lausanne evaluation set will be restricted and accessible via a Grand-Challenge. How it will be accessible: Data deposited in Zenodo. The Lausanne data will be accessible via Grand-Challenge. Methods/software tools to access data: Standard MRI viewing software Repository: Data in Zenodo at https://zenodo.org/records/7051658 and https://zenodo.org/records/7051692); Software and models on github at https://github.com/Shifts-Project/shifts Restrictions on access: Lausanne data restricted due to privacy and licensing agreements
Making data interoperable	Interoperability: Data adheres to standard formats (NIfTI for MRI images) Data and metadata vocabularies: Use of standard medical imaging vocabularies and

	Zenodo metadata standards. Use of standard vocabularies: Yes, standard medical imaging Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Commons CC BY NC SA 4.0. Data can be downloaded after signing OFSEP data usage agreement. Availability for re-use: Lausanne data will be accessible post-publication via controlled access Usable by third parties after end of project: Yes, all data except for the restricted Lausanne dataset Re-use timeframe: Indifinitely Data quality assurance process: Expert validation of segmentation masks and consistency checks across centers
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data anonymization, controlled access, user authentication
Ethical aspects	Possible ethical and legal aspects preventing sharing: Privacy concerns for patient data Is informed consent for data sharing and long term preservation given: Yes
Other Issues	N/A

4.4 Datasets collected in the context of WP5

4.4.1 ObyGaze12 dataset for detection of visual objectification in films

DMP component	AI4Media_Data_12_WP5_Video_ObyGaze12_v1 Partner: UCA
Data Summary	Purpose: The ObyGaze12 dataset is made of 1,914 movie clips densely annotated by experts for objectification concepts identified in film studies and psychology. The purpose is to reveal and quantify the usage of complex temporal patterns operated in cinema to produce the cognitive perception of objectification. Type/format: csv Re-use of existing data: No Data origin: Films are annotated from scratch by experts Expected size: 2,000 clips densely annotated with 8 visual concepts Data utility: This data is useful to AI researchers to tackle a new AI task, that of detecting visual objectification of characters in videos, and to humanities researchers who can analyze quantitatively and qualitatively the data annotations.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, soon to be formatted with the Croissant ML format. Already online on Github, freely accessible and usable. Search keywords: Yes, through the Croissant ML format. Versioning: No Metadata creation: Croissant ML is going to be added to the formatting to have a

	machine-readable version of the dataset, including the metadata.
Making data openly accessible	Data openly accessible: This dataset produced and used in the project is made openly available at https://github.com/husky-helen/ObyGaze12 How it will be accessible: Deposition in an open repository Methods/software tools to access data: The data is readable by custom code, and soon described in the Croissant ML format for easier access. As of May 2024, it is publicly hosted on Github with code examples to read and use in an ML environment. Repository: Currently on Github and soon on Zenodo Restrictions on access: N/A
Making data interoperable	Interoperability: Will soon be with Croissant ML formatting Data and metadata vocabularies: Will soon be with Croissant ML formatting Use of standard vocabularies: Will soon be with Croissant ML formatting Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: CC-BY v4.0 Availability for re-use: Immediately Usable by third parties after end of project: Yes Re-use timeframe: Not determined Data quality assurance process: The validity of the dense annotations has been evaluated with different forms of inter-annotator agreement (see publication)
Allocation of resources	Costs for making data FAIR: Already covered by another project (engineers on data formatting) Costs for long-term preservation: Zenodo is used, there are no additional costs as data is light to host
Data security	Security measures: N/A
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given:The data corresponds to annotations of films by experts. Experts are project partners. Films can be used for research as per the Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights translated into the French law in 2021.
Other Issues	Refer to other national/funder/sectorial/departmental procedures for data management that you may be using (if any)

4.4.2 MAD-TSC Multilingual Aligned Dataset for Target-dependent Sentiment Classification

DMP component	AI4Media_Data_13_WP5_TEXT_MAD-TSC_v1 Partner: CEA
Data Summary	Purpose: MAD-TSC (Multilingual Aligned Dataset for Target-dependent Sentiment Classification) enables sentiment classification in political news. Sentiment is annotated for individual entities occuring in sentences sampled from geographically and politically diversified news sources. The dataset includes 5,110 annotated entity mentions from 4,714 unique sentences. The sentences are aligned across eight

	languages since all sentences have professional translations in all languages. The dataset facilitates the comparison of sentiment classification approaches across languages. MAD-TSC enables a fine-grained understanding of political news when combined with structured knowledge bases such as Wikidata and ParlGov. The predictions can be easily aggregated to assess the orientation of news sources, the public discussion of impactful topics, or the opinions expressed about demographic segments. Type/format: The dataset includes textual data and associated annotations. It is distributed in JSON format. Re-use of existing data: The sentences are sampled from news articles available via the VoxEurop website. Data origin: https://voxeurop.eu/en/ Expected size: 5MB Data utility: The dataset is useful to multiple stakeholders. NLP research groups can use it to train and evaluate fine-grained sentiment analysis methods in the eight languages included. Social and/or political scientists can use models pre-trained with the dataset to automatically analyze political texts shared on the Web. Media organizations, such as AI4Media industrial partners, can use the MAD-TSC to analyze the positioning of their own content and that of other news organizations toward different political tendencies and the potential biases toward demographic segments. CEA and VRT exemplified the utility of the dataset via an analysis of Dutch-language news sources from Belgium. Relevant to both WP5 and WP6.
Making data findable, incl. provisions for metadata	Is data discoverable: MAD-TSC has a DOI reference obtained via Zenodo (https://doi.org/10.5281/zenodo.7940057). The dataset is referenced in the associated publication (https://doi.org/10.18653/v1/2023.acl-long.461). It is also distributed on Github, freely accessible and usable. Search keywords: Yes, the dataset is searchable with its title as keywords in any public search engine. Versioning: No Metadata creation: Metadata include the annotations associated with the entity mentions from the sentences. They are stored in JSON format using a usual three-class sentiment scale (-1=negative; 0 = neutral; 1=positive)
Making data openly accessible	Data openly accessible: Yes, the dataset is available via Github and Zenodo How it will be accessible: https://github.com/EvanDufraisse/MAD_TSC and https://zenodo.org/records/7940057 Methods/software tools to access data: The dataset is readable by custom code. It is hosted on Github, along with detailed usage examples. Repository: Github and Zenodo Restrictions on access: No restriction
Making data interoperable	Interoperability: The dataset is distributed following common practices in the sentiment classification sub-community. Data and metadata vocabularies: Metadata are formatted according to common

	practices in the sentiment classification sub-community Use of standard vocabularies: Standard vocabulary is used for formatting annotations. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: MIT Availability for re-use: Immediate Usable by third parties after end of project: Yes Re-use timeframe: Not determined Data quality assurance process: Each entity mention was annotated by at least three annotators. Only mentions with sufficient inter-annotator agreement were included.
Allocation of resources	Costs for making data FAIR: The annotation costs were covered via AI4Media and another project. Costs for long-term preservation: None. The dataset is hosted on Zenodo
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: No Is informed consent for data sharing and long term preservation given: The data corresponds to annotations of films by experts. Experts are project partners. Films can be used for research as per the Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights translated into the French law in 2021.
Other Issues	Following the guidelines and requirements of the French national research agency (ANR)

4.4.3 ÖWF Object Detection dataset

DMP component	AI4Media_Data_14_WP5_Video_OWFOD_v1 Partner: JR
Data Summary	Purpose: This dataset contains object detection annotations and quality metadata of historic film with scientific and educational content. Type/format: JSON and CSV files, and Python code to access the MPEG-4 video files Re-use of existing data: No Data origin: The content is from the collection "Österreichische Bundesinstitut für den Wissenschaftlichen Film (ÖWF)" of the Austrian Mediathek. Expected size: 100MB Data utility: Researchers/developers working on object detection.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: No Versioning: Yes Metadata creation: Annotations are in JSON format, adhering to the widely used MS COCO format.
Making data	Data openly accessible: Data available on Zenodo at

openly accessible	https://zenodo.org/records/8400030 How it will be accessible: Downloadable from Zenodo Methods/software tools to access data: Python code for working with the data is included Repository: Data is available in Zenodo. Code is available at GitHub Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: Data Card is provided Use of standard vocabularies: No Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: MIT License Availability for re-use: Available Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: Annotations have been manually checked by two experts
Allocation of resources	Costs for making data FAIR: 1,000€, covered by AI4Media Costs for long-term preservation: None, using public repositories
Data security	Security measures: Zenodo has appropriate security mechanisms
Ethical aspects	Possible ethical and legal aspects preventing sharing: Persons are depicted, but the content has been cleared to be provided on the web Is informed consent for data sharing and long term preservation given: No
Other Issues	N/A

4.4.4 PeopleAtPlaces dataset

DMP component	AI4Media_Data_15_WP5_Video_People@Places_v1 Partner: JR
Data Summary	Purpose: The PeopleAtPlaces dataset provides annotations per image for training classifiers for bustle (i.e., more or less populated) and cinematographic shot type (from extreme close-up to extreme long shot). The annotations are provided for images of the Places365-Standard training set, and have been generated using an automatic pipeline. 100 images per class are provided as validation set, and the annotations of these images have been manually checked and corrected.t. Type/format: JSON and CSV files, and Python code to access the JPEG images Re-use of existing data: Yes Data origin: Places365-Standard dataset Expected size: ~1GB Data utility: Researchers/developers working video shot classification

Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: No Versioning: Yes Metadata creation: Annotations are in JSON format (following human pose description formats) and CSV
Making data openly accessible	Data openly accessible: Yes, available in Zenodo at https://zenodo.org/records/8398916 How it will be accessible: Downloadable from Zenodo Methods/software tools to access data: Python code for reproducing the dataset is included Repository: Data is available in Zenodo. Code is available at GitHub Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: No Use of standard vocabularies: No Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: CC-BY-4.0 License Availability for re-use: Available Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: Annotations have been manually checked
Allocation of resources	Costs for making data FAIR: 1,000€ covered by a national research project Costs for long-term preservation: None, using public repositories
Data security	Security measures: Zenodo has appropriate security mechanisms
Ethical aspects	Possible ethical and legal aspects preventing sharing: Persons are depicted, but the content has been cleared to be provided with the source dataset. Is informed consent for data sharing and long term preservation given: No
Other Issues	N/A

4.4.5 ToDY dataset for visual time of day and season classification

DMP component	**AI4Media_Data_16_WP5_Video_ToDY_v1 Partner: JR**
Data Summary	Purpose: The dataset provides training and validation data for classifying images by time of day and season (time of year). The images are taken from the Skyfinder dataset, containing webcam images along with timestamps and geolocation. The annotations have been automatically derived from the metadata, which is sufficiently precise for season. For time of day, the annotation have been manually checked and corrected. The dataset contains 2,790 training files and 311 validation files per class for

	season, and 986 training files and 110 validation files per class for time of day. Type/format:CSV files, and Python code to access the MPEG-4 video files Re-use of existing data: Yes Data origin: Skyfinder dataset Expected size: 500MB Data utility: Researchers/developers working on visual classification of seasons and times of day
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: No Versioning: Yes Metadata creation: Annotations are in CSV format, partly mined from existing metadata, partly manually annotated
Making data openly accessible	Data openly accessible: Available on Zenodo at https://zenodo.org/records/8398861 How it will be accessible: Downloadable from Zenodo Methods/software tools to access data: Python code for recreating the dataset is provided Repository: Data available on Zenod and code is available on GitHub at https://github.com/wbailer/ToDY Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: N/A Use of standard vocabularies: No Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: CC-BY-4.0 License Availability for re-use: Available Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: Annotations have been manually checked
Allocation of resources	Costs for making data FAIR: 1,000€ covered by a national research project Costs for long-term preservation: nNone, using public repositories
Data security	Security measures: Zenodo has appropriate security mechanisms
Ethical aspects	Possible ethical and legal aspects preventing sharing: n/a Is informed consent for data sharing and long term preservation given: No
Other Issues	N/A

4.4.6 VISIONE Feature Repository

DMP component	AI4Media_Data_17_WP5_ImageFeatures_VisioneFeatureRepository_v1 Partner: CNR
Data Summary	Purpose: Three datasets containing a diverse set of features extracted from the V3C1+V3C2, VBSLHE, and MVK video datasets, respectively. These features were extracted by CNR and employed in the VISIONE video retrieval system during the latest editions of the Video Browser Showdown (VBS) competition (https://www.videobrowsershowdown.org/). This repository comprises the following files related to V3C1+V3C2, VBSLHE, and MVK datasets: ● Tab-separated files reporting the segmentation of each video into shots. ● A Python script designed to extract the middle frame of each video segment ● An archive of image features extracted using the ALADIN model for all the segment's middle frames. ● An archive of image features extracted using the CLIP ViT-H/14 - LAION-2B model for all the segment's middle frames. ● An archive of image features extracted using the CLIP ViT-L/14 model for all the segment's middle frames. . ● An archive of image features extracted using the CLIP2Video model for all the segment's middle frames. ● An archive of objects detected using the Faster R-CNN+Inception ResNet (trained on the Open Images V4) model for all the segment's middle frames. ● An archive of objects detected using Mask R-CNN (trained on LVIS) model for all the segment's middle frames. . ● An archive of objects detected using VfNet (trained on COCO dataset) model for all the segment's middle frames . Type/format: CSV files, HDF5 files and a Python Script, Re-use of existing data: Yes, we are reusing the original videos or keyframes from the three datasets: V3C1+V3C2, VBSLHE, and MVK. These were used solely for feature extraction and are not redistributed. Data origin: V3C1+V3C2, VBSLHE and MVK datasets Expected size: 90GB Data utility: These datasets are essential for testing and enhancing video retrieval systems, particularly for research groups involved in the VBS competition. They also provide researchers with pre-extracted image features for more than 3M video keyframes, saving the time and effort required for feature extraction.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: Yes Versioning: Yes Metadata creation: N/A.
Making data openly accessible	Data openly accessible: Yes, available in Zenodo How it will be accessible: Downloadable from Zenodo Methods/software tools to access data: The data can be accessed using any software

	that supports HDF5 and CSV file formats, such as Python or Matlab. Repository: The data has been deposited in three Zenodo repositories: https://zenodo.org/records/8188570 (features extracted from V3C1+V3C2) https://zenodo.org/records/8355037 (features extracted from MVK), https://zenodo.org/records/10013329 (Features extracted from VBSLHE) Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: N/A Use of standard vocabularies: Yes, we will use standard vocabularies for all data types in our dataset to allow inter-disciplinary interoperability. This includes using the HDF5 format for structured data storage, and standard CSV formatting for tabular data. Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Commons Attribution 4.0 International. Availability for re-use: Yes Usable by third parties after end of project: Yes Re-use timeframe: undetermined Data quality assurance process: checks for data integrity and completeness
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: none, using public repositories
Data security	Security measures: Zenodo has appropriate security mechanisms
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: No
Other Issues	N/A

4.4.7 COCO, LVIS, Open Images V4 classes mapping

DMP component	AI4Media_Data_18_WP5_TEXT_COCO- LVIS-Open ImagesV4-classes Mapping_v1 Partner: CNR
Data Summary	Purpose: This repository contains a mapping between the classes of COCO, LVIS, and Open Images V4 datasets into a unique set of 1,460 classes. COCO contains 80 classes, LVIS contains 1,460 classes, Open Images V4 contains 601 classes. The mapping of these classes was obtained using a semi-automatic procedure in order to have a unique final list of 1,460 classes. Moreover, a hierarchy for each class, using wordnet, has been generated. Type/format: CSV and TXT file Re-use of existing data: Using the class names of COCO, LVIS, and Open Images V4 datasets Data origin: COCO, LVIS, and Open Images V4 datasets Expected size: 1 MB

	Data utility: This data will be useful for researchers and developers in AIMedia as well as third parties. It can be used, for example, for systems that aim to utilize multiple object detectors trained on these datasets simultaneously by leveraging the unified class mapping and hierarchy /as done in the VISIONE video retrieval system.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, available on Zenodo Search keywords: Yes Versioning: Yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, available in Zenodo: https://zenodo.org/records/7194300 How it will be accessible: Downloadable from Zenodo Methods/software tools to access data: Any software supporting TXT and CSV files Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Commons Attribution 4.0 International Availability for re-use: Yes Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: Manually checks for integrity, completeness
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Zenodo has appropriate security mechanisms
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

4.4.8 Bus violence dataset

DMP component	AI4Media_Data_19_WP5_Video_BusViolence_v1 Partner: CNR
Data Summary	Purpose: The Bus Violence dataset is a large-scale collection of videos depicting violent and non-violent situations in public transport environments. This benchmark was gathered from multiple cameras located inside a moving bus where several people simulated violent actions, such as stealing an object from another person, fighting between passengers, etc. It contains 1,400 video clips manually annotated as having or not violent scenes, making it one of the biggest benchmarks for video violence

	detection in the literature. Specifically, videos are recorded from three cameras at 25 frames per second - two cameras located in the corners of the bus (with resolution 960x540 px) and one fisheye in the middle (1280x960 px). The clips have a minimum length of 16 frames and a maximum of 48 frames, capturing a very precise action (either violence or non-violence). The dataset is perfectly balanced, containing 700 videos of violence and 700 videos of non-violence. Type/format: MP4 files Re-use of existing data: No, this is original data. Data origin: Surveillance cameras mounted on public transporttaion vehicles Expected size: 500MB Data utility: This dataset is useful to researchers working in AI and Computer Vision to train and test solutions for visual detection of violence scenes.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: Yes Versioning: Yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, avialbel on Zenodo https://zenodo.org/records/7044203 How it will be accessible: From Zenodo Methods/software tools to access data: The data can be accessed using any software that supports MP4 standard video files. Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: N/A Use of standard vocabularies: Yes, vocabulary is the standard binary one for violence detection (1 – video containing violence, 0 – video without violence). Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Commons Attribution 4.0 International Availability for re-use: Yes Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: Checks for data integrity and completeness
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: none, using public repositories
Data security	Security measures: Zenodo has appropriate security mechanisms
Ethical	Possible ethical and legal aspects preventing sharing: No

aspects	Is informed consent for data sharing and long term preservation given: No
Other Issues	N/A

4.4.9 Pest Sticky Traps dataset

DMP component	AI4Media_Data_20_WP5_Image_PestStickyTraps_v1 Partner: CNR
Data Summary	Purpose: The Pest Sticky Traps dataset is a collection of yellow chromotropic sticky trap pictures specifically designed for training/testing deep learning models to automatically count insects and estimate pest populations. Images were manually annotated by some experts of the Department of Agriculture, Food and Environment of the University of Pisa (Italy) by putting a dot over the centroids of each identified insect. Specifically, we labeled insects as belonging to the category "whitefly" considering two different species, i.e., the sweet potato whitefly (Bemisia tabaci) (Gennadius) and the greenhouse whitefly (Trialeurodes vaporariorum) (Westwood). Type/format: JPG and CSV files Re-use of existing data: No, this is original data. Data origin: Pictures from smartphone Expected size: 150MB Data utility: The dataset can be useful to researchers working in AI and Computer vision to develop and tests solutions for visual counting. In particular, density estimation of insects.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: Yes Versioning: Yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, available in Zenodo: https://zenodo.org/records/7801239 How it will be accessible: From Zenodo Methods/software tools to access data: The data can be accessed using any software that supports JPG and JSON standard files. Repository: Zenodo repository Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Commons Attribution 4.0 International Availability for re-use: Yes

	Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: Checks for data integrity and completeness
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: None, using public repositories
Data security	Security measures: Zenodo has appropriate security mechanisms
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: No
Other Issues	N/A

4.4.10 Virtual World Fallen People dataset

DMP component	AI4Media_Data_21_WP5_IMAGE_VWFP_v1 Partner: CNR
Data Summary	Purpose: A synthetic dataset for visual fallen people detection comprising images extracted from the highly photo-realistic video game Grand Theft Auto V developed by Rockstar North. Each image is labeled by the game engine, which provides bounding boxes and statuses (fallen or non-fallen) of people present in the scene. The dataset comprises 6,071 synthetic images depicting 7,456 fallen and 26,125 non-fallen pedestrian instances in various looks, camera positions, background scenes, lightning, and occlusion conditions. VWFP aims to improve fallen-people detection systems that often lack difficult-to-gather training data (e.g., people in high-risk environments) using synthetic data from virtual worlds. Type/format: PNG images and annotations in CSV file. Re-use of existing data: No, the dataset is created within AI4Media. Data origin: Grand Theft Auto V developed by Rockstar North Expected size: 11.5 GB Data utility: It is useful to researchers working on AI and computer Vision to develop and test solutions to detect fallen people.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available at https://zenodo.org/records/6394684 Search keywords: N/A Versioning: Zenodo supports versioning Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is available on Zenodo at https://zenodo.org/records/6394684 How it will be accessible: The data can be downloaded freely. Methods/software tools to access data: N/A Repository: Zenodo

	Restrictions on access: None.
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: Annotations contain bounding box coordinates and class (fallen/non-fallen) for each person in the images. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: GNU General Public License (GPL) v3 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: Zenodo has appropriate security mechanisms
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

4.4.11 SR_BVI-DVC super-resolution dataset

DMP component	AI4Media_Data_22_WP5_VIDEO_SR-BVI-DVC_v1 Partner: BSC
Data Summary	Purpose: This dataset is comprised of 2,000 sequences of 64 frames each, all in 4K resolution. The source material for these sequences is 200 original clips from the BVI-DVC public dataset. Each of these clips has been upscaled using nine distinct super-resolution algorithms, encompassing both traditional and Deep Learning-based methods. The traditional methods employed include nearest-neighbor interpolation, bicubic interpolation, and bilinear interpolation. The Deep Learning-based methods utilized are Real-BasicVSR, RVRT, BasicVSR, SwinIR-Real, SwinIR-Classical, and Real-ESRGAN. The objective of this dataset is to create pairs of "real" and "fake" or upscaled video content in 4K resolution, featuring identical scenes. This dataset has been used for the training and evaluation of a Super-Resolution detection model specifically designed for 4K videos. Type/format: Each video is represented by a sequence of 64 .png frames at 3840 x 2160 resolution Re-use of existing data: The original 4K clips are being used as “real” or natively 4K content. Data origin: The original dataset is a collection from different sources: Videvo Free Stock Video Footage set, IRIS32 Free 4K Footage set, Harmonics database, BVI-Texture database, MCML 4K video quality database, BVI-HFR database, SJTU 4K video database, LIVE-Netflix database, Mitch Martinez Free 4K Stock Footage set, Dareful Free 4K Stock Video data set, MCL-V database, MCL-JCV database, Netflix Chimera,

	TUM HD databases, Ultra Video Group-Tampere University database Expected size: Around 2TB Data utility: The dataset can be employed to train and evaluate Super-Resolution models, Super-Resolution detection models or image/video quality assessment models.
Making data findable, incl. provisions for metadata	Is data discoverable: No, because the dataset described is stored on BSC's internal corporate systems, although there are plans for future sharing. Search keywords: N/A Versioning: N/A Metadata creation: The README includes information about the original video sources, as well as details about upscaling algorithms and their parameters.
Making data openly accessible	Data openly accessible: The data is not openly accessible, although there are plans for future sharing. How it will be accessible: Via a third-party repository link, only shared via explicit request. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: The BSC research team has access to the dataset.
Making data interoperable	Interoperability: The data is preserved in two formats: individual .png files and compressed files for each upscaling method. This allows for flexibility in accessing and utilizing the data. 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 original database was compiled by the University of Bristol, Bristol, UK, comprising sequences originally generated by various sources. In the same manner, all intellectual property rights remain with the originators of each sequence. The test sequences shall only be used for academic research (no commercial use). Availability for re-use: Yes, in the future, for academic research only. Usable by third parties after end of project: Yes, in the future, for academic research only. 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 in secure BSC machines, with user control access via unique credentials and VPN. BSC 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: Source video license incompatibilities. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.4.12 BSC4K super-resolution dataset

DMP component	AI4Media_Data_23_WP5_VIDEO_BSC4K_v1 Partner: BSC
Data Summary	Purpose: This dataset contains paired video sequences at 1080p and 4K resolution recorded simultaneously. The motivation of the dataset is to overcome the challenges introduced by artificially downscaling high-resolution content to obtain the low-resolution counterparts, a common method in Super-Resolution tasks. The first version of the dataset contains 33 4K and 33 1080p videos, cut to 64 frames each, recorded indoor and outdoor with a single DSLR camera. The 1080p versions are upscaled to 4K to create pairs of original/synthetic data. The traditional methods employed include nearest-neighbor interpolation, bicubic interpolation, and bilinear interpolation. The Deep Learning-based methods utilized are Real-BasicVSR, RVRT, BasicVSR, SwinIR-Real, SwinIR-Classical, and Real-ESRGAN. Type/format: Each video is represented by a sequence of 64 .png frames at 3840 x 2160 resolution or 1920 x 1080. Re-use of existing data: All data is manually collected and original. Data origin: All data is manually collected and original. Expected size: Around 300GB Data utility: The dataset can be employed to train and evaluate Super-Resolution models, Super-Resolution detection models or image/video quality assessment models.
Making data findable, incl. provisions for metadata	Is data discoverable: The dataset is not publicly available. There are no plans for data sharing. Search keywords: N/A Versioning: N/A Metadata creation: The dataset includes information and metadata about the 4K and 1080p video sources, as well as details about upscaling algorithms and their parameters.
Making data openly accessible	Data openly accessible: The data is not openly accessible. There are no plans for data sharing. How it will be accessible: N/A Methods/software tools to access data: N/A Repository: N/A Restrictions on access: The BSC research team has access to the dataset.
Making data interoperable	Interoperability: The data is preserved in two formats: individual .png files and compressed files for each upscaling method. This allows for flexibility in accessing and

	utilizing the data. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: TBD 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 in secure BSC machines, with user control access via unique credentials and VPN. BSC 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: Every individual who is present in the video sequences has provided their explicit consent for their recording. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.4.13 Night and Day Instance Segmented Park dataset

DMP component	AI4Media_Data_24_WP5_IMAGE_NDISPark_v1 Partner: CNR
Data Summary	Purpose: A collection of images of parking lots for vehicle detection, segmentation, and counting. Each image is manually labeled with pixel-wise masks and bounding boxes localizing vehicle instances. The dataset includes about 250 images depicting several parking areas describing most of the problematic situations that we can find in a real scenario: seven different cameras capture the images under various weather conditions and viewing angles. Another challenging aspect is the presence of partial occlusion patterns in many scenes such as obstacles (trees, lampposts, other cars) and shadowed cars. The main peculiarity is that images are taken during the day and the night, showing utterly different lighting conditions. Type/format: JPG and JSON files Re-use of existing data: No, this is original data. Data origin: Surveillance cameras Expected size: 120MB Data utility: This dataset can be used by researchers working on AI and Computer Vision to develop and test solutions for visual parking space analysis.

Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: Yes Versioning: Yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, available on Zenodo: https://zenodo.org/records/6560823 How it will be accessible: From Zenodo Methods/software tools to access data: The data can be accessed using any software that supports JPG and JSON standard files. Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Commons Attribution 4.0 International. Availability for re-use: Yes Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: Checks for data integrity and completeness
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: None, using public repositories
Data security	Security measures: Zenodo has appropriate security mechanisms
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: No
Other Issues	N/A

4.4.14 ArXiv abstracts for authorship analysis dataset

DMP component	**AI4Media_Data_25_WP5_TEXT_ArXivAbstracts_v1 Partner: CNR**
Data Summary	Purpose: This data set comprises a labelled training set used in the experimentation of the paper "Binary Quantification and Dataset Shift: An Experimental Investigation" from CNR. The data is extracted from the McAuley data set of product reviews on Amazon. Type/format: Raw text partitioned in samples. A file per sample Re-use of existing data: This is sample from ArXiv data Data origin: https://info.arxiv.org/help/bulk_data/index.html

	Expected size: 1.3GB Data utility: Benchmark for Authorship Analysis
Making data findable, incl. provisions for metadata	Is data discoverable: Data is available online (https://zenodo.org/records/7404702) and indexed in Google. Search keywords: author analysis, text classification Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, available in Zenodo: https://zenodo.org/records/7404702 How it will be accessible: From Zenodo Methods/software tools to access data: Web browser Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: The file structure makes the use of the dataset easy. Data and metadata vocabularies: A single csv file with abstract of a paper and a unique author's identifier. No metadata. Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: Creative Commons Attribution 4.0 International Availability for re-use: Data is already available online. Usable by third parties after end of project: Yes 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 Zenodo servers. Zenodo 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

4.4.15 Florence4D facial expression dataset

DMP component	AI4Media_Data_26_WP5_4D_Florence4D_v1 Partner: UNIFI
Data Summary	Purpose: Collection of 3D sequences representing transitions between a diverse set of facial expressions. This dataset aims at filling the gap in the availability of 4D data.

	Type/format: Sequences of point cloud data in OBJ format Re-use of existing data: Part of the dataset contains CoMA meshes. Data origin: CoMA, synthetic models from the web, real 3D Scans. Expected size: 30 GB Data utility: The dataset is useful for all research and industrial teams working on 3D representation learning and facial expression recognition.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, available in UNIFI website (https://www.micc.unifi.it/resources/datasets/florence-4d-facial-expression/) Search keywords: Florence 4D Facial Expression Dataset Versioning: No Metadata creation: N/A
Making data openly accessible	Data openly accessible: Accessible from UNIFI website https://www.micc.unifi.it/resources/datasets/florence-4d-facial-expression/ after completing a form https://forms.gle/AgLNSjjMZYXdUAK68 How it will be accessible: From UNIFI website Methods/software tools to access data: OBJ is an open format. Repository: Google drive Restrictions on access: Data accessible after form completion
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: OBJ format, annotation in plain text Use of standard vocabularies: Expressions are classified according to Plutchik's wheel of emotions Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Common Attribution License-CC-BY v4.0 Availability for re-use: Yes Usable by third parties after end of project: Free for non-commercial purposes Re-use timeframe: N/A Data quality assurance process: Manual checking
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: N/A
Ethical aspects	Possible ethical and legal aspects preventing sharing: Written consents were collected from subjects for using their 3D face scans. Is informed consent for data sharing and long term preservation given: Yes.
Other Issues	N/A

4.4.16 Neuromorphic Event-based Facial Expression Recognition dataset

DMP	AI4Media_Data_27_WP5_EventRGB_NEFER_v1

component	Partner: UNIFI
Data Summary	Purpose: Paired RGB and event videos representing human faces labeled with the respective emotions and also annotated with face bounding boxes and facial landmarks. Currently, the only event+RGB emotion dataset. Type/format: RGB, Raw Event Data, Reconstructed Event Frames Re-use of existing data: No Data origin: Acquisition Expected size: 10 Gb Data utility: The dataset is useful for all research and industrial teams working on neuromorphic representation learning and facial expression recognition.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, DOI: https://doi.org/10.1109/CVPRW59228.2023.00432 Search keywords: NEFER Versioning: No Metadata creation: N/A
Making data openly accessible	Data openly accessible: Accessible on GitHub https://github.com/miccunifi/NEFER How it will be accessible: From GitHub Methods/software tools to access data: Web browser Repository: GitHub, Google drive Restrictions on access: No
Making data interoperable	Interoperability: Yes Data and metadata vocabularies: RGB+CSV Use of standard vocabularies: Expressions are classified according to Ekman's emotion classification Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Creative Common Attribution License-CC-BY v4.0 Availability for re-use: Yes Usable by third parties after end of project: Free for non-commercial purposes Re-use timeframe: N/A Data quality assurance process: Manual checking
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: N/A
Ethical aspects	Possible ethical and legal aspects preventing sharing: Written consents were collected from subjects for using their 3D face scans Is informed consent for data sharing and long term preservation given: Yes
Other Issues	N/A

4.4.17 PEM360 dataset of 360° videos

DMP component	AI4Media_Data_28_WP5_Video_PEM360_v1 Partner: UCA
Data Summary	Purpose: PEM360 includes user head movements and gaze recordings in 360° videos, along with self-reported emotional ratings of valence and arousal, and continuous physiological measurement of electrodermal activity and heart rate. Its purpose is to understand the connection between user attention, user emotions and immersive content. Type/format: csv Re-use of existing data: No Data origin: The 360° video stimuli are selected from an existing publicly available dataset Expected size: 34 users on 7 360° videos each, amounting to 238 traces of gaze motion and emotion ratings and physiological measures Data utility: This data is useful to better understand non-motion factors (such as emotion or video content) that can non-deterministically impact models to predict user attention.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is publicly available, but no metadata file can be automatically parsed. Search keywords: PEM360 Versioning: No Metadata creation: N/A
Making data openly accessible	Data openly accessible: This dataset produced and used in the project is made openly available at https://gitlab.com/PEM360/PEM360/ How it will be accessible: Deposition in an open repository Methods/software tools to access data: The data is readable by custom code made available in the dataset repository Repository: Gitlab 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: CC-BY Availability for re-use: Immediately Usable by third parties after end of project: Yes Re-use timeframe: Not determined Data quality assurance process: The validity of the collected user traces has been analyzed through EDA and known correlations have been retrieved.

Allocation of resources	Costs for making data FAIR: Not engaged Costs for long-term preservation: Gitlab is used, there are no additional costs as data is light to host
Data security	Security measures: N/A
Ethical aspects	Possible ethical and legal aspects preventing sharing: No Is informed consent for data sharing and long term preservation given: Yes, the user experiment carried out to collect the data has been approved by an IRB and informed consent has been formally collected (and user retribution financially).
Other Issues	N/A

4.4.18 VRT-Sum video summarization dataset

DMP component	AI4Media_Data_29_WP5_VIDEO-TEXT_VRT-Sum_v1 Partner: VRT, CERTH
Data Summary	Purpose: This dataset is composed of a set of videos of varying content (e.g. sports, news, interviews), the professionally-edited video summaries (one per video), and a short text that is used as a voice-over during the presentation of the video summary (one or two per video). It will be used in T5.1 for training and evaluation purposes, assisting the development of text-driven methods for video summarization, Type/format: The video files are in MP4 format, and the text descriptions (in English and Dutch) are in WORD files. Re-use of existing data: No Data origin: The VRT partner of AI4Media Expected size: ~12.6GB 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: No 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 and the textual content of the description, or data about the shot-level structure of the videos) will be stored in HDF5 files; a documentation of these metadata will be created to facilitate their re-use.
Making data openly accessible	Data openly accessible: No. The videos and the associated text descriptions are proprietary data of VRT that we do not have the right to make publicly available. 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. The videos and the associated text descriptions are proprietary data of VRT that we do not have the right to share.

	Data and metadata vocabularies: Any metadata generated to assist training and evaluation of text-driven video summarization methods (e.g. deep feature vectors representing the visual content of video frames and the textual content of the description, 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 and the associated text descriptions are proprietary data of VRT that we do not have the right to grant a licence. Availability for re-use: N/A Usable by third parties after end of project: The videos and the associated text descriptions are proprietary data of VRT, that we do not have the right to make available for re-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 and any created metadata (e.g. deep feature vectors representing the visual content of video frames and the textual content of the description, or data about the shot-level structure of the videos) will be hosted on CERTH's servers. CERTH 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. Regular rolling daily backups are scheduled to minimise the risk of data loss.
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	No

4.4.19 CA-SUM pretrained video summarization models

DMP component	AI4Media_Data_30_WP5_MODEL_CA-SUM-models_v1 Partner: CERTH
Data Summary	Purpose: This dataset contains pretrained models of the CA-SUM network architecture for video summarization, that is presented in CERTH's work titled "Summarizing Videos using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames", in Proc. ACM ICMR 2022. It will be used in T5.1 for comparison with other summarization approaches, and for assisting the implementation of the relevant use cases of WP8. Type/format: The model files are in PT format. Re-use of existing data: No

	Data origin: CERTH Expected size: ~195MB Data utility: It will be useful to WP5 partners working on the development of video summarization methods, for comparison purposes. It will be useful also to WP8 partners involved in use cases that require the production of video summaries.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: CA-SUM pretrained models, video summarization Versioning: N/A Metadata creation: N/A.
Making data openly accessible	Data openly accessible: Yes, available on Zenodo How it will be accessible: The pretrained models are publicly available on Zenodo at https://zenodo.org/records/6562992 Methods/software tools to access data: The pretrained models can be found using any web browser and the above listed search keywords Repository: Zenodo at https://zenodo.org/records/6562992 Restrictions on access: No
Making data interoperable	Interoperability: Yes. The pretrained models can be used both on Windows and Linux machines. Data and metadata vocabularies: A documentation of this dataset is available on Zenodo (https://zenodo.org/records/6562992). Details about the CA-SUM video summarization method are available on the relevant scientific publication (https://zenodo.org/records/6759007) Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset is provided for academic, non-commercial use only. Availability for re-use: Yes Usable by third parties after end of project: Yes. The dataset is publicly available for use. Re-use timeframe: N/A Data quality assurance process: N/A
Allocation of resources	Costs for making data FAIR: No. The dataset is publicly available on the Zenodo platform that is cost-free. Costs for long-term preservation: No. The dataset is publicly available on the Zenodo platform that is cost-free.
Data security	Security measures: The dataset is hosted on Zenodo, which is a EU open research repository. Zenodo is powered by CERN Data Centre and the Invenio digital library framework and is fully run on open source products all the way through. Zenodo servers are managed via OpenStack and Puppet configuration management system which ensures that the servers always have the latest security patches applied. The

	dataset is hosted also on CERTH's servers. CERTH 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. Regular rolling daily backups are scheduled to minimise the risk of data loss.
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	No

4.4.20 Audio phylogeny dataset

DMP component	AI4Media_Data_31_WP5_Audio_AudioPhylogeny_v1 Partner: FhG-IDMT
Data Summary	Purpose: The IDMT Audio Phylogeny Dataset contains audio phylogeny trees for evaluation of audio phylogeny algorithms. It includes two different sets of phylogeny trees with 60 trees each, where every tree contains 20 nodes (audio files). The main difference between these two sets is in the set of transformations T. Type/format: audio files (.wav) and phylogeny trees metadata as text (.txt) Re-use of existing data: Yes. 3 existing audio files are used as roots for the processing. Data origin: VCTK corpus version 0.92. Under Creative Commons License: Attribution 4.0 International Expected size: 1.5 GB Data utility: This dataset is useful in the context of T5.6 for the evaluation of algorithms for audio phylogeny analysis.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable. The dataset is hosted on Zenodo: https://zenodo.org/records/8135331 Search keywords: Audio phylogeny dataset IDMT Versioning: N/A Metadata creation: For every audio phylogeny tree generated metadata is provided describing the process of generations by indicationg parent-child couples and transformations applied.
Making data openly accessible	Data openly accessible: The dataset is available on Zenodo at https://zenodo.org/records/8135331 How it will be accessible: Downloadable from Zenodo Methods/software tools to access data: Web-browser to download the data as zip file Repository: Zenodo: https://zenodo.org/records/8135331 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: Creative Commons Attribution 4.0 International License. Availability for re-use: Yes Usable by third parties after end of project: Yes, it's 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: Zenodo has appropriate security mechanisms
Ethical aspects	Possible ethical and legal aspects preventing sharing: N/A Is informed consent for data sharing and long term preservation given: One half of the audio data originates from an open dataset of speech samples whose speakers have given a permission for data sharing and reuse. Th other half of audio data is AI generated music with no ethical or licensing concerns.
Other Issues	N/A

4.4.21 RAI CMM documentaries dataset

DMP component	AI4Media_Data_32_WP5_Video_CMM-Doc-dataset_v1 Partner: RAI
Data Summary	Purpose: The CMM Doc dataset is a collection of selected documentaries with their segmentation from the Rai archive. The documentaries are used to experiment with the video segmentation techniques considered within the project and to compare the quality of these segmentations with the ones present in the archive (ground truth). Type/format: mp4, csv, wav Re-use of existing data: No Data origin: Rai Archive Expected size: 12GB, around 15 hours of audio/video content plus the csv files describing the segmentations. Data utility: It is useful to evaluate the performance of different video segmentation approaches, comparing each of them with a ground truth.
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: No, data is stored on RAI's private 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: 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: N/A
Ethical aspects	Possible ethical and legal aspects preventing sharing: Copyright owned by RAI Is informed consent for data sharing and long-term preservation given: N/A
Other Issues	N/A

4.4.22 RAI CMM-ANTS newscasts dataset

DMP component	AI4Media_Data_33_WP5_Video_CMM-ANTS-dataset_v1 Partner: RAI
Data Summary	Purpose: The CMM-ANTS dataset is a collection of selected newscasts with their segmentation from the Rai archive. The programmes are used to experiment with the video segmentation techniques considered within the project and to compare the quality of these segmentations with the ones present in the archive (ground truth). Type/format: mp4, csv, wav Re-use of existing data: No Data origin: Rai Archive Expected size: 5GB, around 15 hours of audio/video content plus the csv files describing the segmentations. Data utility: It is useful to evaluate the performance of different video segmentation approaches, comparing each of them with a ground truth.
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: No, data stored on RAI's private 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: 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: N/A
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

4.4.23 RAI CMM mixed dataset

DMP component	**AI4Media_Data_34_WP5_Video_CMM-dataset_v1 Partner: RAI**
Data Summary	Purpose: CMMdataset is a collection of selected content of different kinds with their segmentation (if available) from the Rai archive. These contents are used to experiment with the video segmentation techniques considered within the project and to compare the quality of these segmentations with the ones present in the archive (ground truth). Type/format: mp4, wav, csv Re-use of existing data: No Data origin: Rai Archive Expected size: 70GB, around 90 hours of audio/video content plus the csv files describing the segmentations. Data utility: It is useful to evaluate the performance of different video segmentation approaches, comparing each of them with a ground truth.

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: No, data stored on RAI's private 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: 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: N/A
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

4.4.24 Cross lingual news dataset

DMP component	AI4Media_Data_35_WP5_Text_Cross-lingual-news_v1 Partner: RAI
Data Summary	Purpose: The cross lingual news dataset is a collection of news items published by international online newspapers and press agencies. It is used by RAI and CNR to train and test CNR's Generalized Funneling component. Type/format: json Re-use of existing data: No Data origin: News items from RSS feeds and press agencies Expected size: 20 MB Data utility: It is useful to WP5 partners to evaluate and benchmark text classification

	models.
Making data findable, incl. provisions for metadata	Is data discoverable: No, this is a private dataset. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: No 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: 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: N/A
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

4.4.25 YouTube RAI channel dataset

DMP component	AI4Media_Data_36_WP5_Misc_YouTubeDataset_v1 Partner: RAI
Data Summary	Purpose: A collection of URLs to RAI's YouTube channel content items + chapterisation of data. Type/format: json Re-use of existing data: No Data origin: RAI YouTube channel Expected size: 1.3 MB

	Data utility: It is useful to WP5 partners to evaluate and benchmark media editorial segmentation.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: N/A Versioning: Yes, provided by hosting platform Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: https://huggingface.co/datasets/raicrits/YouTube_RAI_dataset Methods/software tools to access data: git Repository: https://huggingface.co/datasets/raicrits/YouTube_RAI_dataset 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: Apache 2.0 Availability for re-use: Yes Usable by third parties after end of project: Yes 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
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

4.5 Datasets collected in the context of WP6

4.5.1 GreekPolitics Twitter dataset

DMP component	AI4Media_Data_37_WP6_SOCIALMEDIA_GreekPolitics_v1 Partner: AUTH
Data Summary	Purpose: The AUTH GreekPolitics Dataset contains 2,578 tweet IDs from Twitter posts with politically charged content in the Greek language, spanning the period January 2014 – March 2021. Manually annotated ground-truth labels along 4 sentiment dimensions are provided for each tweet: polarity ('1' = positive, '0' = neutral, '-1' =

	negative), figurativeness ('1' = figurative, '0' = literal), aggressiveness ('1' = aggressive, '0' = non-aggressive) and bias ('1' = partisan, '0' = non-partisan). This data is used in Task 6.4 "*AI for Healthier Political Debate*" to investigate political public opinion mining and/or monitoring. The dataset is 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 use data mined from Twitter using the official Twitter API. Data origin: GreekPolitics Twitter posts were collected based on specific query hashtags related to the Greek political scene, using the official Twitter API. These hashtags are mainly related to the names of the various political parties and popular politicians represented in the Greek parliament over the past decade, while variants of them (in both the Greek and the Latin alphabet) were also exploited. Expected size: ~ 10 GB Data utility: The data is 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 dataset is stored at an internal AUTH server and it is discoverable via AUTH's webpage https://aiia.csd.auth.gr/auth-greekpolitics-dataset/. The data comprising the dataset (i.e. the tweets) are discoverable via the tweet ID; anyone can access the associated tweet text through the Twitter API. There is a chance that a subset of the GreekPolitics tweets may have been deleted, in the time interval that has passed since the dataset was originally constructed by AUTH. Search keywords: AUTH GreekPolitics Dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: In order to access the GreekPolitics Dataset (https://aiia.csd.auth.gr/auth-greekpolitics-dataset/), interested parties should complete and sign a license agreement: http://aiia.csd.auth.gr/wp-content/uploads/2023/10/AUTH_GreekPolitics_Dataset_License.pdf. How it will be accessible: After completing the license agreement, interested parties receive FTP credentials for downloading the dataset from AUTH's servers. Methods/software tools to access data: Download via FTP Repository: AUTH servers Restrictions on access: Only available upon signing a license agreement.
Making data interoperable	Interoperability: Using Twitter's standard data representation model. Data and metadata vocabularies: The following metadata vocabulary is 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: Data is shared after interested parties should and sign a license agreement: http://aiia.csd.auth.gr/wp-content/uploads/2023/10/AUTH_GreekPolitics_Dataset_License.pdf 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: Yes Re-use timeframe: N/A Data quality assurance process: Data quality assurance is 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 is stored at an internal AUTH server. Appropriare security measures are implemented and compliance with GDPR is considered.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Twitter data cannot be redistributed directly because of the terms of service of Twitter APIs. We only share the Tweet IDs. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

4.5.2 Covid-19 Twitter dataset

DMP component	AI4Media_Data_38_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 spans the period from March 2020 to March 2021. 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: 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 is 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	Possible ethical and legal aspects preventing sharing: Twitter data cannot be

aspects	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

4.5.3 Twitter Text dataset

DMP component	AI4Media_Data_39_WP6_TEXT_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 4.5.2. 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. Twets were collected from March 2020 to March 2021. 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 4.5.2, 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

4.5.4 Twitter COVID-19 discussions topics dataset

DMP component	AI4Media_Data_40_WP6_TEXT_Covid19DiscussionTopicsTwitter_v1 Partner: BSC
Data Summary	Purpose: Dataset of tweet IDs and timestamps in CSV format, organized by COVID-19-related discussion topics they belong to. It contains a total of 3,423,260 entries between 10 topics and covers 2 time periods: March to August 2020 and August 2020 to March 2021. The tweet IDs correspond to the tweets included in the COVID-19 Twitter dataset described in section 4.5.2 The topics include: 1. Azithromycin and covid 2. Bleach and covid 3. 5G and covid 4. Covid is artificial 5. Amrmy mobilizations during covid 6. Invermictin and covid 7. Moderna vaccine 8. Astrazeneca vaccine 9. Blood types and covid susceptibility 10. H1N1 and covid All timestamps are in UNIX seconds. 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: CSV Re-use of existing data: The original tweets were stored by Twitter. This dataset includes tweet IDs and timestamps and is based on the dataset described in section 4.5.2, originally collected by BSC. Data origin: Twitter API Expected size: ∽ 100 MB Data utility: The data is 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 researchers to perform research in similar topics.
Making data findable, incl. provisions for metadata	Is data discoverable: The dataset is discoverable through the AI4Media website (https://www.ai4media.eu/open-datasets/), which includes a link to the datast repository (https://docs.hpai.cloud/s/A6EwRCrswm3B5Q8) The original tweets are discoverable via twitter.com and via twitter API using the

	tweet IDs included in the datset Search keywords: AI4Media Twitter Discussions Topics dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Dataset available on the cloud at https://docs.hpai.cloud/s/A6EwRCrswm3B5Q8 How it will be accessible: Access via web-browser - direct download of data Methods/software tools to access data: Web browser Repository: https://docs.hpai.cloud/s/A6EwRCrswm3B5Q8 Restrictions on access: N/A
Making data interoperable	Interoperability: Use of Twitter IDs and UNIX timestamps Data and metadata vocabularies: The following metadata vocabulary is used: *id: Tweet ID* *timestamp_ms*: Unix timestamp (in miliseconds) of when the Tweet was created at Use of standard vocabularies: Use of Twitter IDs and UNIX timestamps Mappings to commonly used vocabularies: Use of Twitter IDs and UNIX timestamps
Increase data re-use	Licence: MIT license Availability for re-use: Yes Usable by third parties after end of project: Yes Re-use timeframe: N/A Data quality assurance process: Yes
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data stored on BSC-HPAI's repo on Nextcloud Hub. Nextcloud features a host of innovative security technologies from brute force protection to advanced server side and integrated end-to-end, client side encryption with enterprise-grade key handling and a wide range of security hardenings.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Twitter data cannot be redistributed because of the terms of service of Twitter APIs. Thus we wonly share tweetIDs Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.5.5 ODSS Open Dataset of Synthetic Speech

DMP component	AI4Media_Data_41_WP6_AUDIO_ODSS_v1 Partner: FhG-IDMT, CERTH
Data Summary	Purpose: ODSS is a multilingual, multispeaker dataset of synthetic and natural speech, designed to foster research and benchmarking of novel studies on synthetic speech detection. ODSS comprises audio utterances generated from text by state-of-the-art synthesis methods, paired with their corresponding natural counterparts. The synthetic audio data includes several languages, with an equal representation of genders. This dataset is required to learn the statistical distribution of features derived from synthetic speech data – to be distinguished, e.g., from pristine speech data. Type/format: Audio (wav and csv files) Re-use of existing data: Yes, re-utterances and transcripts of speech corpora of pristine speech. Data origin: The data was generated using Text-to-Speech synthesis methods. Expected size: 2.4GB Data utility: Useful in the context of T6.2 for developing and testing new algorithms for content manipulation/synthesis detection.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, data is hosted on Zenodo and has a DOI reference (https://doi.org/10.5281/zenodo.8370668) Search keywords: ODSS, Open Dataset of Synthetic Speech Versioning: Yes Metadata creation: Metadata is provided in CSV format, which has the advantage of being both machine-readable and human-readable.
Making data openly accessible	Data openly accessible: Yes, data available on Zenodo at https://zenodo.org/records/8370669 How it will be accessible: Downloadable via zenodo.org. Methods/software tools to access data: Any software supporting WAV and CSV files. Repository: Zenodo (https://zenodo.org/records/8370669) Restrictions on access: No
Making data interoperable	Interoperability: Audio files are present in WAV format, which is largely 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: CC-BY-SA v4.0 Availability for re-use: Yes Usable by third parties after end of project: Yes Re-use timeframe: Undetermined Data quality assurance process: N/A
Allocation of	Costs for making data FAIR: N/A

resources	Costs for long-term preservation: None
Data security	Security measures: N/A
Ethical aspects	Possible ethical and legal aspects preventing sharing: Requirements of the CC-BY-SA 4.0 licence need to be fulfilled. Is informed consent for data sharing and long term preservation given: All recordings used to train the neural networks involved in the data generation, as well as all pristine recordings included in the dataset were acquired in controlled conditions and with the informed consent of all the speakers involved, in compliance with the GDPR.
Other Issues	N/A

4.5.6 CelebHQGaze image dataset for gaze estimation

DMP component	AI4Media_Data_42_WP6_IMAGE_ CelebHQGaze _v1 Partner: UNITN
Data Summary	Purpose: Common gaze-correction methods usually require annotating training data with precise gaze, and head pose information. Solving this problem using an unsupervised method remains an open problem, especially for high-resolution face images in the wild, which are not easy to annotate with gaze and head pose labels. To address this issue, we created a high-resolution CelebHQGaze dataset of faces with corrected gazes. CelebHQGaze consists of 29,255 high resolution celebrity images that are collected from CelebAHQ. It consists of 21,005 face images with the eyes staring at the camera and 8,250 face images with eyes looking somewhere else. Similarly to CelebGaze, we extract facial landmarks and generate the mask. All images are cropped to 512 × 512, and the mask size is fixed to 46×80. Useful to WP6 for gaze correction applications. Type/format: JPEG images and accompanying metadata in json format. Re-use of existing data: No, the dataset is created within AI4Media. Data origin: CelebA dataset Expected size: ~3 GB Data utility: It is useful to WP6 partners (especially T6.2) to correct the gaze of portraits.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, it is discoverable via GitHub and the project website Search keywords: CelebHQGaze Versioning: GitHub supports versioning Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is openly accessible via GitHub at https://github.com/zhangqianhui/GazeAnimationV2 How it will be accessible: The data can be downloaded from an online archive after completing a form. Methods/software tools to access data: Web browser Repository: GitHub 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: The repository consists of two datasets: CelebGaze and CelebHQGaze. CelebGaze consists of 25,283 celebrity images, most of which have been collected from CelebA and a minority from the Internet. Specifically, there are 21,832 face images with the eyes staring at the camera and 3,451 face images with the eyes looking somewhere else. We cropped all the images to 256 x 256 and computed the eye mask region using Dlib. Specifically, we used Dlib to extract 68 facial landmarks, and we computed the mean of 6 points near the eye region, which is the center point of the mask. The size of the mask is fixed to 30 x 50. CelebHQGaze consists of 29,255 high resolution celebrity images that are collected from CelebAHQ. It consists of 21,005 face images with the eyes staring at the camera and 8,250 face images with eyes looking somewhere else. Similarly to CelebGaze, we extract facial landmarks and generate the mask. All images are cropped to 512 x 512, and the mask size is fixed to 46 x 80. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The data and the code are 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 is shared in a google drive repository.
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

4.5.7 European news about Covid vaccination dataset

DMP component	AI4Media_Data_43_WP6_TEXT_European-news-covid-vaccination_v1 Partner: IDIAP
Data Summary	Purpose: The dataset consists of 50,000+ full newspaper articles in five countries (France, Italy, Spain, Switzerland, and UK), and four languages (French, Spanish, Italian, and English) over a period of 22 months, as well as a translated English version for the full dataset. The common theme of the articles is about Covid vaccination and the objective of this dataset was to study the local component and the peculiarities of each country addressing the common theme of vaccination. From this original dataset, a subsample of articles focused on the No-Vax movement (1,786 articles) has been used to carry out two additional lines of research, one concerning the identification of frames in these articles and the other concerning the study of quality articles on the subject of disinformation. The dataset is used in T6.5 for the analysis of hyper-local news.

	Type/format: csv file Re-use of existing data: No Data origin: All the data is collected from the 19 webpages of the newspapers. Expected size: 460 MB Data utility: The dataset contains articles from European newspapers with a great tradition and reputation, with the content in their original language, as well as in English. This dataset can be of interest for research lines such as disinformation detction since this content could be labelled as truthful, as well as for political analysis of how information is presented by different ideologies or countries.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is not openly accessible. There are no plans for data sharing. Search keywords: N/A Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: The data is not openly accessible. There are no plans for data sharing. 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: 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 IDIAP's servers, and can only be accessed by authorized users.
Ethical aspects	Possible ethical and legal aspects preventing sharing: We have not obtained permission from the newspapers to share the data. We have obtained authorization only for data extraction and use for research purposes. Is informed consent for data sharing and long term preservation given: N/A

Other Issues	N/A

4.5.8 Suisse Romande local news dataset

DMP component	AI4Media_Data_44_WP6_TEXT_SuisseRomandeLocalNews_v1 Partner: IDIAP
Data Summary	Purpose: The dataset contains 130,155 articles sourced from the websites of three Swiss francophone newspapers: Arc Info, La Cote, and Le Nouvelliste, spanning the time period from 01/01/2015 to 30/06/2022. The dataset, compiled from the temporary data feeds provided by the press agency, consists of the articles in their entirety including the title, headline, and content, along with metadata for each article. The collected articles are primarily in French language and are categorically sorted by topics and region, everything encoded in the JSON format. The dataset is used in T6.5 for the analysis of hyper-local news. Type/format: JSON Re-use of existing data: No Data origin: Local newpapers websites (Arc Info, La Cote, and Le Nouvelliste) Expected size: 1 GB Data utility: Used in T6.5 for the analysis of hyper-local news. Useful to researchers studying local news press.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, with a unique identifier: https://zenodo.org/records/10256911 Search keywords: Suisse Romande local news Versioning: Only one release version Metadata creation: The metadata are the ones available on the websites associated to their respective articles
Making data openly accessible	Data openly accessible: The dataset is accessible via Zenodo upon request (the authors' names are anonymized). How it will be accessible: From Zenodo at https://zenodo.org/records/10256911 Methods/software tools to access data: Only need for a json reader, the documentation explains the process in Python. Repository: Available on Zenodo Restrictions on access: Access upon request
Making data interoperable	Interoperability: The identifier and the metadata provide unique sets of attributes for each article, which can be identified this way. Data and metadata vocabularies: Date and newspapers. Use of standard vocabularies: No use of standard vocabularies Mappings to commonly used vocabularies: The mapping is explained in the documentation.
Increase data re-use	Licence: Worldwide distribution is restricted to academic institutions / non profit

	research institutions. Availability for re-use: Yes, no embargo. Usable by third parties after end of project: Available for research purposes Re-use timeframe: N/A Data quality assurance process: Data quality assessment + back-up system
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: Marginal (small dataset without updates)
Data security	Security measures: Accessible only upon request. Dataset stored in IDIAP's servers and alla propriate security measures are implemented.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The press agency authorized the use for research pruposes only. Is informed consent for data sharing and long term preservation given: The consent was given by the owner of the rights to the articles
Other Issues	N/A

4.5.9 Lausanne news dataset

DMP component	AI4Media_Data_45_WP6_TEXT_LausanneNews_v1 Partner: IDIAP
Data Summary	Purpose: Dataset of 2,666 articles scraped from the newspaper Lausanne CItée, all published in French, from the time period between January 1, 2019 and November 30, 2022. The dataset is used in T6.5 for the analysis of hyper-local news. Type/format: pickle Re-use of existing data: No Data origin: All the data is collected from the webpage of the newspaper Lausanne CItée Expected size: 8.6 MB Data utility: Used in T6.5 for the analysis of hyper-local news. Useful to researchers studying local news press.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is not openly accessible. There are no plans for data sharing. Search keywords: N/A Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: The data is not openly accessible. There are no plans for data sharing. 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: 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 IDIAP's servers, and can only be accessed by authorized users.
Ethical aspects	Possible ethical and legal aspects preventing sharing: We have not obtained permission from the newspaper to share the data. We have obtained authorization only for data extraction and use for research purposes by IDIAP. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.5.10 Suisse Allemande local news dataset

DMP component	AI4Media_Data_46_WP6_TEXT_SuisseAllemandeNews_v1 Partner: IDIAP
Data Summary	Purpose: The dataset consists of around 14,500 full articles from 2 local newspapers in the canton of Zurich, Tagblatt der Stadt Zürich and Winterthurer Zeitung, with all the content in German. These are articles between the years 2012 and 2022. The dataset is used in T6.5 for the analysis of hyper-local news. Type/format: csv file Re-use of existing data: No Data origin: All the data is collected from the webpages of the newspapers Tagblatt der Stadt Zürich and Winterthurer Zeitung. Expected size: Few MBs Data utility: Used in T6.5 for the analysis of hyper-local news. Useful to researchers studying local news press.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is not openly accessible. There are no plans for data sharing. Search keywords: N/A Versioning: No

	Metadata creation: No
Making data openly accessible	Data openly accessible: The data is not openly accessible. There are no plans for data sharing. 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: 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 IDIAP's servers, and can only be accessed by authorized users.
Ethical aspects	Possible ethical and legal aspects preventing sharing: We have not obtained permission from the newspapers to share the data. We have obtained authorization only for data extraction and use for research purposes by IDIAP. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.5.11 References on YouTube dataset

DMP component	**AI4Media_Data_47_WP6_TEXT_YouTubeReferences_v1 Partner: IDIAP**
Data Summary	Purpose: The dataset describes how YouTube knowledge communications videos put and use references to specify external information sources that are used to create the videos. 44 English-speaking YouTube channels that specialize in knowledge communication were manually chosen, and one video per channel was chosen for the analysis. For each video, the referencing scheme (*in-video* and *bibliography*) and general context about the channel/creators (visual style, upload date, the number of subscribers, topic, details of the creator group, running time of the video) were summarized and categorized. Additionally, a sample of 129 references was collected to verify whether the resources were actually available or not. The dataset is used in T6.5 for the analysis of referencing practices in YouTube.

	Type/format: csv file Re-use of existing data: No Data origin: The channel list was collected based on past literature that listed a selection of knowledge communication channels on YouTube and extended by using the recommendation of the YouTube platform. The video and reference list items were arbitrarily selected from each channel. Expected size: 1 MB Data utility: To the best of our knowledge, this dataset is the first systematic observation of referencing practices in YouTube. By linking the accessibility and availability of external references included in online videos, we can extend the research questions to automatic validation of video content with regard to references, or towards the fundamental debate about information accessibility and information quailty.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is not openly accessible. There are no plans for data sharing. Search keywords: N/A Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: The data is not openly accessible. There are no plans for data sharing. 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: 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 IDIAP's servers, and can only be accessed by authorized users.
Ethical	Possible ethical and legal aspects preventing sharing: The data is not publicly available

aspects	as that was not requested in the Institutional Ethical approval of the protocol. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

4.5.12 Political Barometer dataset

DMP component	AI4Media_Data_48_WP6_TEXT_PoliticalBarometer_v1 Partner: AUTH
Data Summary	Purpose: This Dataset contains text from Twitter posts with politically charged content in Greek language to be used in the Political Barometer software (https://icarus.csd.auth.gr/political-barometer/) to predict the outcome of the Greek elections of June 2024. Due to Twitter's change in API access policy and limitations, this dataset remains private. Twitter IDs and URLs weren't stored in a consistent manner and format. The dataset is used for the research activities of T6.4, focusing on online political debate analysis. Type/format: Text Re-use of existing data: No Data origin: Tweets in Greek language collected from Twitter, spanning the period June 2022 - March 2024. Expected size: 0.25 GB Data utility: The data is useful for text and sentiment analysis and analysis of online political debate.
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: No 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: 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: No

	Availability for re-use: No 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: Data is stored on the Intranet of the AIIA Laboratory (AUTH). Appropraite security measures are implemented.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Due to Twitter's change in API access policy and limitations, this dataset remains cannot be shared. Vaild identifiers to make data eligible for sharing weren't store in a consistent manner. Is informed consent for data sharing and long-term preservation given: N/A
Other Issues	N/A

4.5.13 ElecDeb60To16-fallacy dataset

DMP component	AI4Media_Data_49_WP6_TEXT_ElecDeb60To16_v1 Partner: UCA
Data Summary	Purpose: The ElecDeb60To16-fallacy dataset collects televised debates of the presidential election campaigns in the U.S. from 1960 to 2016. It contains political debates annotated with argument components (evidence, claim), relations (support, attack), 6 fallacy types, and 14 sub-fallacy types. This data is meant to be used in Task 6.4 "AI for Healthier Political Debate" to aid in the construction of text machine-learning models that classify fallacious arguments. Type/format: csv files Re-use of existing data: We extended the ElecDeb60To16 dataset by adding the annotations for fallacious arguments. Data origin: The political debates' text was collected from the website of the Commission on Presidential Debates, which provided transcripts of the debates broadcasted on TV and held among the leading candidates for the presidential and vice presidential nominations in the US. Expected size: 25 MB Data utility: Useful in the context of T6.4 for political debate analysis. Also, useful to any researcher working on falacious argument analysis and debate analysis.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is publicly available and discoverable from the project website and GitHub, but no metadata file can be automatically parsed. Search keywords: ElecDeb60To16-fallacy dataset Versioning: N/A Metadata creation: N/A
Making data	Data openly accessible: The dataset is publicly available on GitHub at

openly accessible	https://github.com/pierpaologoffredo/IJCAI2022/tree/main/dataset How it will be accessible: Downloadable via GitHub Methods/software tools to access data: Given that the data's format is csv, any spreadsheet software or text editor can be used to access it. Repository: GitHub (https://github.com/pierpaologoffredo/IJCAI2022) Restrictions on access: No
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: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Availability for re-use: The data is available for re-use Usable by third parties after end of project: The dataset is publicly available for non-commercial purposes Re-use timeframe: N/A Data quality assurance process: Inter-annotator agreement analysis
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: Data is publicly available on GitHub. GitHub has appropriate security mechanisms
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

4.6 Datasets collected in the context of WP8

4.6.1 Data from user research activities in Use Case 1

DMP component	**AI4Media_Data_50_WP8_USER-RESEARCH_UseCase1_DW_v1 Partner: DW**
Data Summary	Purpose: In order to realise Use Case 1 in WP8, partner DW has designed various research activities with *professional* users, who work at DW or fact-checking organisations (journalists, verification experts or managers). This research is required for the definition of user requirements (D8.1), further user requirements gathered for the first White Paper (D8.3) as well as the user evaluation of new functions integrated into the Truly Media demonstrator and related trustworthy AI elements. The following, mainly qualitative research methods, have been conducted with small, selected target groups: tailored questionnaire-based surveys, 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 AI4Media project (aspects such as usefulness, usability, business and journalistic value,

	or performance). By preparing or conducting this research with professional users, DW has generatet a dataset containing very limited *Participant Personal Data* and *User Responses related to surveys with* questionnaires that were conducted anonymously. Any data will be stored internally on DW's systems/servers and will not be shared with external organisations or AI4Media project partners. The limited Participant Personal Data included information such as Name, Company Email and Company. 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 has been used in the context of WP8 to define the user requirements and to evaluate the AI-enhanced demonstrator (see above). Analysed, aggregated results have been 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 is 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. Is informed consent for data sharing and long-term preservation given: Questionnaire-based surveys were conducted anonymously where users did not provide their personal data such as Name, Company, or email. They were informed about the purpose of this research, the anonymous approach and how responses are used/reported.
Other Issues	No

4.6.2 Data from user research activities in Use Case 2

DMP component	AI4Media_Data_51_WP8_USER-RESEARCH_UseCase2_VRT_v1 Partner: VRT
Data Summary	Purpose: In order to realise Use Case 2, VRT has conducted qualitative research activities with journalists, who work at VRT and other European media organisations. This research was required for the definition of the user requirements (D8.1), further user requirements gathered for the first White Paper (D8.3) as well as the user evaluation of new AI functionalities as part of the Smart News Assistant (D8.3, D8.6). Mainly qualitative research methods have been conducted such as interviews, focus groups or demo/evaluation sessions with individual users. By preparing or conducting this research with professional users, VRT has generated a dataset containing Online meeting recordings and User Responses related to surveys. Questionnaires were conducted anonymously. Any data will be stored internally on VRT's systems/servers and will not be shared. Recordings are removed after the project. The limited 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 VRT and the production of aggregated results summaries in WP8 deliverables. Type/format: Documents containing Participant Personal Data and documents containing questions/user responses from user research activities. Video recordings of demo sessions and interviews. 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 has been used in the context of WP8 to define the final set of user requirements for UC2 and to evaluate the AI functionalities in the UC2

	demonstrators. Analysed, aggregated results have been 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 is stored on VRT's internal corporate systems and there are no plans for data sharing. The WP8 deliverables containing aggregated, analysed results are classified as confidential 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: The data will only be accessible by VRT. 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 VRT. Data handling and protection follows standard VRT 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: Questionnaire-based surveys were conducted anonymously where users did not provide their personal data such as Name, Company, or email. They were informed data was only used for research analysis purposes.
Other Issues	No

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

DMP component	AI4Media_Data_52_WP8_QUESTIONNAIRE_UseCase3-UserReqCollection2021_V1 Partner: RAI
Data Summary	Purpose: A structured questionnaire has been developed by RAI to collect user opinions about end user requirements proposed by UC3. The questionnaire includes questions about each presented feature/epic/user story;its goal is to understand if and how each feature can be useful in a production workflow and to highlight a possible development priority. The questionnaire was filled by RAI journalist/editorial staff people/archivists and by people with the same profiles belonging to other broadcasters. The collected questionnaires have been analysed and the results of this analysis have been used to drive UC3 features development. The results and related analyses are 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 has been used in the context of WP8 to evaluate requirements proposed by UC3. The consortium used 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 are stored on RAI servers. A summary of the requirements evaluation results is available in D8.1. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Raw data is not openly accessible since it contains personal information. It is protected and shared only among project partners and a summary of this data has been presented in D8.1. In case of further reports or papers submitted for publication, all findings will be integrated into the reports or papers. Datasets will not be added to the publication. How it will be accessible: The dataset is 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 is not licensed since it can 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 is 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 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: An Informed Consent Form was prepared for the participation in the requirements collection activity to manage treatment of personal data. The questionnaires 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

4.6.4 Questionnaires for the evaluation of Use Case 3

DMP component	**AI4Media_Data_53_WP8_QUESTIONNAIRE_UseCase3Evaluation_V1 Partner: RAI**
Data Summary	Purpose: Structured questionnaires have been developed by RAI for the evaluation of the developed AI services/tools in the context of UC3 pilot trials. The questionnaires 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 is used in the context of WP8 to evaluate different versions of AI4Media services/tools for UC3. The evaluation results of each pilot phase have been

	used by the technical partners to improve and extend the functionalities of the developed services/tools during the following 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) are stored on RAI servers. A summary of the evaluation results have been 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 is not openly accessible since it contains personal information. It is protected and shared only among project partners. However, a summary of the evaluation results has been presented in relevant WP8 deliverables. In case of further reports or papers submitted for publication, all research findings will be integrated into the reports or papers. Datasets will not be added to the publication. How it will be accessible: The dataset is 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: Dataset is not licensed since it can 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 is 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 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: An Informed Consent Form has been prepared for the participation of end-users in the evaluation activities to manage treatment of personal data. The questionnaires 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

4.6.5 Data from user research activities in Use Case 4

DMP component	AI4Media_Data_54_WP8_USER-RESEARCH_UseCase4_NISV_v1 Partner: NISV
Data Summary	Purpose: In order to realise Use Case 4 in WP8, partner NISV performed various qualitative research activities with researchers from social sciences and humanities (SSH). This research was required to define the user requirements brought forward in the project, as well as evaluate the CLARIAL Media Suite demonstrator. The following, mainly qualitative research methods were executed with small, selected target groups: tailored questionnaires, interviews, focus groups or demo/evaluation sessions with individual users or small groups. By conducting this research with professional users, NISV generated a dataset containing *Participant Personal Data* and *User Responses*. This data is stored internally on NISV's systems/servers and it is not planned to share this data with external organisations or AI4Media project partners. Participant Personal Data include information such as name, company email, position, and area of work/expertise. The data will be used for internal analysis purposes by NISV and the production of aggregated results summaries in WP8 deliverables. 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 was used in the context of WP8 to define the final set of user requirements and to evaluate the demonstrator for Use Case 4. Analysed, aggregated results were shared with technical partners to develop/improve functionalities, the demonstrators and the plans for commercial exploitation. Data presented in the deliverables has been anonymised.
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 WP8 deliverables containing aggregated, analysed results are classified as confidential 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: 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 is only 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 is 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 was prepared for potential participants in the user research activities described above. This includes 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

4.6.6 Questionnaires for the evaluations of Use Case 5-B

DMP component	**AI4Media_Data_55_WP8_QUESTIONNAIRE_UseCase5Evaluation_v1 Partner: IRCAM**
Data Summary	Purpose: For the evaluation of UC5-B (music for games), online questionnaires have been prepared by IRCAM, and filled by the participants of the UC5 tests. The questions are about: the usability and readability of the frontend, the usefulness of the prototype, efficiency of the algorithms, and the relevancy of the approach. The data were first collected in an pseudonymised via Google Forms (https://docs.google.com/forms), then downloaded on a private directory of the experimentator (at IRCAM) for analysis and statistics. This data was then summarised

	to be published in WP8 deliverables. Type/format: Google forms, and Microsoft Excel files. Re-use of existing data: No. Data origin: Google forms filled by participants of the UC5-B demonstrator evaluation, and Excel documents filled by the experimentator of IRCAM during the interviews. Expected size: few MBs. Data utility: These data were used to improve the UC5-B demonstrator at each phase of its development along the project lifetime.
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: The raw (pseudonymised) data are stored locally on a private directory at IRCAM and are not shared publicly. A summary and statistics are published in relevant WP8 deliverables. How it will be accessible: The raw datasets are not accessible, their summary and statistics are published in relevant WP8 deliverables. 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: No, internal use only 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 forms were filled by participants during the online tests, then an interview with the experimentator of IRCAM made it possible to confirm the replies, and to obtain additional details, comments and opinions.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The raw datasets are stored locally at IRCAM on a private directory of the experimentator. They are secured by the Computer and Network System of IRCAM. The questionnaires have been done in a pseudonymised way in order to find the response of each participant during the interview, but the association between the pseudo name and true identity of the participant are not stored.

Ethical aspects	Possible ethical and legal aspects preventing sharing: The datasets are not published since the contain personal data. Is informed consent for data sharing and long term preservation given: The questionnaires began after an Informed Consent question about the use the collected data. All the evaluation participants validated this question.
Other Issues	No

4.6.7 Data from user research activities in Use Case 7

DMP component	AI4Media_Data_56_WP8_QUESTIONNAIRE_VIDEO_UseCase7-UserFeedbackData_v1 Partner: IMG
Data Summary	Purpose: IMG performed interviews with users of the demonstrators in Use Case 7. They also filmed test users in the Living Lab trials they conducted (content managers and team members) using the dmonstrators. A structured questionnaire was developed by IMG to collect user opinions about the demonstrators in UC7. Along with the questionnaire, images, videos and heatmaps of how users interact with the demonstrators were captured. The collected feedback has been analysed, and the results of this analysis were used to drive UC7 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 demonstrators. Re-use of existing data: No Data origin: Questionnaires filled by end-users in the context of UC7. 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 UC7 demonstrators with real users from media companies, in Living Lab trials, were used to validate their performance and help tailor them further to meet user requirements.
Making data findable, incl. provisions for metadata	Is data discoverable: The user feedback data for UC7 was securely stored on IMG servers. Summaries of the feedback data and relevant analyses were included in 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. A summary of this data will only be presented in WP8 deliverables. 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 is 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 are 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: Consent forms were provided to people that took part in the trials and shared this data with IMG for the purpose of improving our technology and software demonstrators. A Privacy Notice specified that the treatment of the data is confidential, complies with GDPR and is carried out exclusively for analytical and statistical purposes.
Other Issues	No

4.6.8 Truly Media dataset

DMP component	**AI4Media_Data_57_WP8_Text_TrulyMediaDataset-UseCase1-ATC_v1 Partner: ATC**
Data Summary	Purpose: Truly Media is used in the context of Use Case 1 as the base demonstrator. The dataset was created for testing purposes. It contains a selection of social media posts from different social media platforms and other web content (news articles, blog posts) depending on the interests and searches made by Truly Media users. The dataset contains 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/X 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 may 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 as part of UC1. 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. provisions for metadata	Is data discoverable: That data will be discoverable only by AI4Media partners that will develop tools for Use Case 1 and that will take part in evaluation activities. The dataset is accessible only through Truly Media's environment and is 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 is 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

4.6.9 Game glitches dataset for Use Case 5

DMP component	AI4Media_Data_58_WP8_Video_GameGlitches-UC5-MODL_v1 Partner: MODL
Data Summary	Purpose: This dataset will contain several short videos of games' glitches and corresponding game logs, generated using the modl.ai platform. This dataset will be used in the context of Use Case 5 (feature 5A) to train an AI capable of automatically recognizing different kinds of game glitches. Type/format: movie data (.mp4), Game log files (.txt) Re-use of existing data: No. Data origin: Videos of game glitches created internally in the company (MODL). Expected size: ~4 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: Yes, via Github and project website Search keywords: MODL, AI4Media, Game glitches dataset Versioning: Yes Metadata creation: No
Making data openly accessible	Data openly accessible: A non-personally identifiable version of the data set has been made generally available on GitHub at https://github.com/modl-ai/ai4media. How it will be accessible: Via GitHub Methods/software tools to access data: Github.com / git Repository: GitHub: https://github.com/modl-ai/ai4media Restrictions on access: No
Making data interoperable	Interoperability: No Data and metadata vocabularies: No

	Use of standard vocabularies: No Mappings to commonly used vocabularies: No
Increase data re-use	Licence: Open use, no liability, warranty, or trademark usage. Availability for re-use: Yes Usable by third parties after end of project: Yes Re-use timeframe: At least 1 year Data quality assurance process: No
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: The dataset is entirely synthetic and openly available. It contains no identifying information. Stored in GitHub.
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

4.6.10 Musical production for AI co-creation dataset

DMP component	AI4Media_Data_59_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 UC6 aims to produce a set of tools to assist artists - especially music composers - 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 an 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 consistent way for exploration and analysis by the artist and developers. Along with the platform for audio generation 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	Is data discoverable: Data is not uniquely identifiable, but can be labeled according to

findable, incl. provisions for metadata	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

4.6.11 Current affairs transcripts dataset for Use Case 4

DMP component	**AI4Media_Data_60_WP8_ TEXT_CurrentAffairsTranscripts-UC4-NISV_v1 Partner: NISV**
Data Summary	Purpose: Within the context of Use Case 4, NISV and IDIAP developed and validated a framework that leverages LLMs for automated frame detection in TV transcripts. For this purpose, NISV created a dataset of transcripts of two current affairs programmes

	on Dutch television. Specifically, it contains a selection of 2,000 news media items from broadcasts of news programs EenVan-daag (1,000 items) and Nieuwsuur (1,000 items). EenVandaag 2 (OneToday) is a daily evening program broadcasted on Dutch public television channel NPO1. EenVandaag has the format of a news programme with current issues and background information behind the news. The programme is about 30 minutes long and deals with various news topics during an episode. It has multiple presenters introducing various news items, and also interview experts live in the studio. Nieuwsuur 3 (News Hour) is also an evening programme and is broadcasted on NPO2, a Dutch public television channel. The broadcasts are between 30 and 45 minutes long, and also have the format of a news programme with current issues and background information behind the news. This programme also has multiple presenters and live interviews with experts. We chose these two current newscasts as they provide a good overview of Dutch news items on a daily basis. These shows also provided a large corpus of items over the years, with little change in the show format. Due to this, the data is very consistent. Type/format: CSV file containing transcripts. Re-use of existing data: The transcripts are extracted from television programmes in the NISV archive. Data origin: Transcripts created within AI4Media by NISV Expected size: A few KBs in total Data utility: For analysis of the framing detection framework, the spoken words in the video recordings of these programmes were transcribed. This was done with the open-source, Kaldi automatic transcriber that NISV uses on audiovisual materials in its archive. This software can automatically transcribe Dutch spoken language into text. This pre-processing step resulted in a dataset of 2,000 texts covering news between 2014 and 2018, varying in length, with an average number of 499 words for Nieuwsuur and 664 words for EenVandaag. The data was then translated from Dutch to English using the Deepl API. The resulting dataset was classified by a human annotator and GPT-3.5 using a prompt composed by the definition of different frame types.
Making data findable, incl. provisions for metadata	Is data discoverable: No. The dataset is stored on NISV's internal server. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: It is not possible to make this dataset openly accessible since the Dutch public broadcaster owns copyright for the transcripts and has not given a permission to make this dataset accessible. NISV is in ongoing negotiations with the broadcaster to make exceptions for this dataset and others created in similar research contexts. How it will be accessible: The data is stored on NISV's internal server. Based on a research agreement, temporary access to the dataset was given to IDIAP for the purposes of the experiment.

	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 copyright restrictions prevent reuse. 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 is 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: It is not possible to make this dataset openly accessible since the Dutch public broadcaster owns copyright for the transcripts and has not given a permission to make this dataset accessible. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	No

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

DMP component	**AI4Media_Data_61_WP8_ QUESTIONNAIRE_AI-IndustrialNeeds-T8.4_v1 Partner: ATC**
Data Summary	Purpose: Structured questionnaires were 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 include questions that aim to gather input and insights from industry stakeholders regarding the AI needs of media organisations. This dataset includes questionnaires filled by the industry stakeholders. Data collected through the questionnaires is 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 is 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 is 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) are stored on the servers of the partners that participate in the relevant T8.4 activities. A summary of the results is included in D8.3 and D8.6, which are confidential and will be accessible only by 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 is 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 are 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: Informed Consent Forms were prepared for the participation in T8.4 validation activities. The questionnaires 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 were used by AI4Media partners to develop and test new AI algorithms, methodologies, and tools in the context of WPs 3, 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 exclusively.

In the initial DMP, 51 third-party research datasets were identified. Since then the number of third-party datasets used within AI4Media has risen to **81 datasets** in total. In the following subsections, we present the DMP plan for these datasets using the same template as in Section 4 (see Table 1).

The following Table briefly summarizes the 81 datasets presented in this section and offers a glance at the structure of the section and its subsections. New datasets that were not included in the initial DMP are indicated with yellow.

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_62_WP3_EMAIL_Enron_v1	WP3	Enron email dataset	5.1.1
AI4Media_Data_63_WP3_SOCIALMEDIA_FacebookWall_v1	WP3	Facebook Wall dataset	5.1.2
AI4Media_Data_64_WP3_IMAGE_AGAIN _v1	WP3	Affect Game Annotation (AGAIN) dataset	5.1.3
AI4Media_Data_65_WP3_TEXT_IMDBreviews_v1	WP3	IMDB movie reviews dataset	5.1.4
AI4Media_Data_66_WP3_TEXT_MHAD2DPose_v1	WP3	MHAD 2D pose dataset	5.1.5
AI4Media_Data_67_WP3-5_TEXT_20Newsgroups_v1	WP3,5	20Newsgroups dataset	5.1.6
AI4Media_Data_68_WP3-5_TEXT_HPAmazonReviews_v1	WP3,5	HP Amazon reviews dataset	5.1.7
AI4Media_Data_69_WP3-5_TEXT_JRCAcquis_v1	WP3,5	JRCAcquis legislative text dataset	5.1.8
AI4Media_Data_70_WP3-5_TEXT_Kindle_v1	WP3,5	Kindle document dataset	5.1.9
AI4Media_Data_71_WP3-5_TEXT_OHSUMED_v1	WP3,5	OHSUMED MEDLINE document dataset	5.1.10
AI4Media_Data_72_WP3-5_TEXT_RCV1-Reuters_v1	WP3,5	RCV1 Reuters stories dataset	5.1.11
AI4Media_Data_73_WP3-5_TEXT_RCV1RCV2-Reuters_v1	WP3,5	RCV1RCV2 Reuters stories dataset	5.1.12
AI4Media_Data_74_WP3-5_TEXT_Reuters-21578_v1	WP3,5	Reuters-21578 dataset	5.1.13
AI4Media_Data_75_WP3-	WP3,5	11 Tweet Sentiment datasets	5.1.14

DMP component	WP	Short summary	Relevant sub-section
5_TEXT_11TweetSentiment_v1			
AI4Media_Data_76_WP3-5_TEXT_WipoGamma_v1	WP3	WipoGamma patent document dataset	5.1.15
AI4Media_Data_77_WP3_Image_WIDER_FACE _v1	WP3	WIDER FACE face detection dataset	5.1.16
AI4Media_Data_78_WP3_IMAGE_Caltech101 _v1	WP3	Caltech 101 dataset	5.1.17
AI4Media_Data_79_WP3_IMAGE_CUB-200-2011 _v1	WP3	CUB-200-2021 image dataset	5.1.18
AI4Media_Data_80_WP3_IMAGE_Describable-Textures_v1	WP3	Describable Textures dataset	5.1.19
AI4Media_Data_81_WP3_IMAGE_Food101_v1	WP3	Food-101 dataset	5.1.20
AI4Media_Data_82_WP3_IMAGE_MAMe_v1	WP3	MAMe dataset	5.1.21
AI4Media_Data_83_WP3_IMAGE_MIT-ISR_v1	WP3	MIT-ISR dataset	5.1.22
AI4Media_Data_84_WP3_IMAGE_OuluKnots_v1	WP3	Oulu Knots dataset	5.1.23
AI4Media_Data_85_WP3_IMAGE_OxfordFlowers_v1	WP3	Oxford Flower dataset	5.1.24
AI4Media_Data_86_WP3_IMAGE_OxfordIIITPet_v1	WP3	Oxford-IIIT Pet dataset	5.1.25
AI4Media_Data_87_WP3_IMAGE_StanfordDogs_v1	WP3	Stanford Dogs dataset	5.1.26
Data used in WP4 (Explainability, Robustness and Privacy in AI)			5.2
AI4Media_Data_88_WP4_IMAGE_Imagenet_01	WP4	ImageNet-ILSVRC2012 image classification dataset	5.2.1
AI4Media_Data_89_WP4_IMAGE_FFHQ_v1	WP4	FFHQ dataset for GAN training	5.2.2
AI4Media_Data_90_WP4_IMAGE_MNIST_v1	WP4	MNIST image dataset	5.2.3
AI4Media_Data_91_WP4_IMAGE_VIDEO_Interestingness10k	WP4	Interestingness10k image +video dataset	5.2.4
AI4Media_Data_92_WP4_VIDEO_Memorability2020	WP4	MediaEval Memorability 2020 dataset	5.2.5
AI4Media_Data_93_WP4_IMAGE_drawnUI2021	WP4	ImageCLEF DrawnUI 2021 dataset	5.2.6
AI4Media_Data_94_WP4_Video_FCVID_v1	WP4	FCVID event recognition dataset	5.2.7
AI4Media_Data_95_WP4_Video_YLI-MED _v1	WP4	YLI-MED event recognition dataset	5.2.8
Data used in WP5 (Content-centered AI)			5.3
AI4Media_Data_96_WP5_VIDEO_SumMeGycli14_v1	WP5	SumMe video summarization dataset	5.3.1
AI4Media_Data_97_WP5_VIDEO_TVSumSong15_v1	WP5	TVSum video summarization dataset	5.3.2
AI4Media_Data_98_WP5_VIDEO_MonumentsOfItaly_v1	WP5	RAI Monuments of Italy dataset	5.3.3

DMP component	WP	Short summary	Relevant sub-section
AI4Media_Data_99_WP5_IMAGE_LVIS_v1	WP5	LVIS image dataset	5.3.4
AI4Media_Data_100_WP5_IMAGE_CIFAR_v1	WP5	CIFAR10/100 image dataset	5.3.5
AI4Media_Data_101_WP5_IMAGE_STL10_v1	WP5	STL-10 image dataset	5.3.6
AI4Media_Data_102_WP5_TEXT_CCNET_v1	WP5	CCNet text dataset for multilingual representation learning	5.3.7
AI4Media_Data_103_WP5_VIDEO_360VideoViewingHeadMountedVR_v1	WP5	360 Video Viewing Dataset in Head Mounted Virtual Reality	5.3.8
AI4Media_Data_104_WP5_VIDEO_HeadMovementinPanoramicVideo_v1	WP5	Predicting Head Movement in Panoramic Video Dataset	5.3.9
AI4Media_Data_105_WP5_VIDEO_GazePredictionDynamic360ImmersiveVideos_v1	WP5	Gaze prediction in Dynamic 360° Immersive Videos Dataset	5.3.10
AI4Media_Data_106_WP5_VIDEO_YourAttentionIsUnique_v1	WP5	Your Attention is Unique Dataset	5.3.11
AI4Media_Data_107_WP5_VIDEO_HeadEyeMovements360Videos_v1	WP5	Dataset of Head and Eye Movements for 360° Videos	5.3.12
AI4Media_Data_108_WP5_Audio_dim-sim music_v1	WP5	Dim-sim dataset for music similarity search	5.3.13
AI4Media_Data_109_WP5_Audio_spam_music_v1	WP5	SPAM dataset for music segmentation	5.3.14
AI4Media_Data_110_WP5_Audio_salami_music_v1	WP5	SALAMI dataset for music segmentation	5.3.15
AI4Media_Data_111_WP5_Audio_harmonix_music_v1	WP5	Harmonix dataset for music segmentation	5.3.16
AI4Media_Data_112_WP5_Audio_FMA_v1	WP5	Free Music Archive dataset	5.3.17
AI4Media_Data_113_WP5_Audio_LAKH-MIDI_v1	WP5	LAKH MIDI music dataset	5.3.18
AI4Media_Data_114_WP5_Audio_MIDI_Piano_v1	WP5	Piano Audio and MIDI music datasets	5.3.19
AI4Media_Data_115_WP5_Audio_GiantSteps _v1	WP5	GiantSteps music datasets	5.3.20
AI4Media_Data_116_WP5_IMAGE_MSCOCO_v1	WP5	MS COCO dataset	5.3.21
AI4Media_Data_117_WP5_VIDEO_BVI-DVC _v1	WP5	BVI-DVC dataset	5.3.22
AI4Media_Data_118_WP5_Image_Adience_v1	WP5	Adience dataset	5.3.23
AI4Media_Data_119_WP5_Image_IMDB-Wiki_v1	WP5	IMDB-Wiki dataset	5.3.24
Data used in WP6 (Human- and Society-centred AI)			5.4
AI4Media_Data_120_WP6_VIDEO_Deepfake-Detection-Challenge-Dataset_v1	WP6	Deepfake Detection Challenge video dataset	5.4.1
AI4Media_Data_121_WP6_VIDEO_FaceForensics++_v1	WP6, WP4	FaceForensics++ video dataset	5.4.2
AI4Media_Data_122_WP4_IMAGE_ImageCLEFaware-dataset_v1	WP6	Visual profile impact rating and ranking – ImageCLEFaware dataset	5.4.3

DMP component	WP	Short summary	Relevant sub-section
AI4Media_Data_123_WP6_EEG_DEAP_v1	WP6	DEAP EEG dataset	5.4.4
AI4Media_Data_124_WP6_EEG_SEED_v1	WP6	SEED EEG dataset	

	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 findable, incl. provisions for metadata	Is data discoverable: Data is discoverable. The dataset is hosted in the Network 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_63_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	Licence: The dataset is already publicly available under an attribution licence

re-use	(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_64_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/1f4eO0A0JH6FE8n5v7ozjJztObPcdLzoJ 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_65_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 metadata	Is data discoverable: Data is available online (https://ai.stanford.edu/~amaas/data/sentiment/) and indexed in Google. It is also 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_66_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_67_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 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 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_68_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_69_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_70_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_71_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_72_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_73_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_74_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_75_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_76_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

5.1.16 WIDER FACE face detection dataset

DMP component	AI4Media_Data_77_WP3_Image_WIDER_FACE _v1 Partner: CERTH
Data Summary	Purpose: WIDER FACE is a public face detection benchmark dataset, of which images are selected from the publicly available WIDER dataset. The dataset features a high degree of variability in scale, pose and occlusion in the depicted faces. It is used by CERTH to evaluate the face detection models in WP3. Type/format: jpg Re-use of existing data: Yes Data origin: The dataset contains more than 32K images available at http://shuoyang1213.me/WIDERFACE/. Expected size: 1.8 GB Data utility: It is useful to WP3 partners to evaluate and benchmark face detection models.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable and publicly available. Search keywords: wider face dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible at http://shuoyang1213.me/WIDERFACE/. 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: http://shuoyang1213.me/ 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 Creative Common License. Availability for re-use: The dataset is already publicly available. Usable by third parties after end of project: Yes. 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 is 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

5.1.17 Caltech 101 dataset

DMP component	**AI4Media_Data_78_WP3_IMAGE_Caltech101 _v1 Partner: BSC**
Data Summary	Purpose: This public dataset contains images of 101 different types of objects, with boundary annotations. Each image has a class label associated. It is used for training models for image classification or object segmentation tasks. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The dataset is a set of images collected via Google Image search. Expected size: 151MB total Data utility: It is useful to WP3 partners working in image classification or object segmentation tasks.
Making data	Is data discoverable: The data can be acquired through the website

findable, incl. provisions for metadata	https://data.caltech.edu/records/mzrjq-6wc02. Search keywords: caltech 101 Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://data.caltech.edu/records/mzrjq-6wc02. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as zip file. Repository: https://data.caltech.edu/records/mzrjq-6wc02 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: Creative Commons Attribution 4.0 International. Allows re-distribution and re-use of the work on the condition that the creator is appropriately credited. Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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

5.1.18 CUB-200-2021 image dataset

DMP component	AI4Media_Data_79_WP3_IMAGE_CUB-200-2011 _v1 Partner: BSC
Data Summary	Purpose: This public dataset contains 11,788 images of 200 bird species, each associated with a Wikipedia article and organized by taxonomic classification. Annotations in the form of image locations are included for several body parts of the birds.

	Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The dataset is a collection of images collected via Flickr Image search. Expected size: Around 2GB total Data utility: It is useful to WP3 partners working in image classification or object detection/location tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website https://www.vision.caltech.edu/datasets/cub_200_2011/. Search keywords: cub200, cub-200-2011 Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://www.vision.caltech.edu/datasets/cub_200_2011/. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as zip file. Repository: https://www.vision.caltech.edu/datasets/cub_200_2011/ 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: No license found. Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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

5.1.19 Describable Textures dataset

DMP component	AI4Media_Data_80_WP3_IMAGE_ Describable-Textures_v1 Partner: BSC
Data Summary	Purpose: The Describable Textures Dataset contains 5,640 textured images, organized according to a list of 47 terms, inspired from human perception. The images contain at least 90% of the surface representing the category attribute. Each image has a class label associated. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The dataset is a collection of images collected via Google and Flickr Image searches. Expected size: Around 1.1GB total Data utility: It is useful to WP3 partners working in image classification tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website https://www.robots.ox.ac.uk/~vgg/data/dtd/. Search keywords: describable textures dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://www.robots.ox.ac.uk/~vgg/data/dtd/. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as tar.gz file. Repository: https://www.robots.ox.ac.uk/~vgg/data/dtd/ 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: Data only available for research purposes. Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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

5.1.20 Food-101 dataset

DMP component	AI4Media_Data_81_WP3_IMAGE_Food101_v1 Partner: BSC
Data Summary	Purpose: The Food-101 dataset contains 101,000 images of food dishes, separated in 101 food categories. Each image has a class label associated. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The images that form the dataset come from individual posts on Foodspotting, a (now defunct) social network for food picture sharing. Expected size: 8.6GB total Data utility: It is useful to WP3 partners working in image classification tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/. Search keywords: food 101 dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as tar.gz file. Repository: https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/ 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: Data available for research purposes. Other uses may be negotiated with the original owners of each picture. Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject

	to licenses from creators. 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 BSC servers. BSC 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

5.1.21 MAMe dataset

DMP component	AI4Media_Data_82_WP3_IMAGE_MAMe_v1 Partner: BSC
Data Summary	Purpose: The dataset contains images of pieces of art from three museums, organized in 29 classes representing different mediums and techniques the art pieces are made on/with. Each image has a class label associated. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The dataset is a collection of images published by the Metropolitan Museum of Art of New York, the Los Angeles County Museum of Art and the Cleveland Museum of Art. All images are under a CC0 license. Expected size: The low-resolution version of the dataset (all images resized to 256x256px) takes 765MB. Data utility: It is useful to WP3 partners working in image classification tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website https://hpai.bsc.es/MAMe-dataset/. Search keywords: mame dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://hpai.bsc.es/MAMe-dataset/. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as a zip file. Repository: https://hpai.bsc.es/MAMe-dataset/

	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: CC BY-NC-ND 4.0 Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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

5.1.22 MIT-ISR dataset

DMP component	**AI4Media_Data_83_WP3_IMAGE_MIT-ISR_v1 Partner: BSC**
Data Summary	Purpose: The dataset contains 15,620 images representing 67 different types of indoor scenes. Each image has a class label associated. A subset of the images are also segmented and annotated with the objects that they contain. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The dataset is a collection of images obtained via Google, AltaVista and Flickr image searches, and from the LabelMe dataset. Expected size: 6.5GB total Data utility: It is useful to WP3 partners working in image classification tasks and object segmentation/recognition tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website https://web.mit.edu/torralba/www/indoor.html. Search keywords: mit indoor scenes dataset Versioning: N/A

	Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://web.mit.edu/torralba/www/indoor.html. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as a tar file. Repository: https://web.mit.edu/torralba/www/indoor.html 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: Research purposes only. Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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

5.1.23 Oulu Knots dataset

DMP component	AI4Media_Data_84_WP3_IMAGE_OuluKnots_v1 Partner: BSC
Data Summary	Purpose: The Oulu Knots dataset contains images depicting wood and various defects on it. Each image has a class label associated. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The images in the dataset were produced by members of the University of Oulu.

	Expected size: 11MB Data utility: It is useful to WP3 partners working in image classification tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data appears to not be reachable currently. Search keywords: N/A Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: This is a public dataset, although it appears to not be reachable currently. 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: 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: No license found. Availability for re-use: N/A Usable by third parties after end of project: The data is not 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 was downloaded from the original source and was stored on BSC servers. BSC 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

5.1.24 Oxford Flower dataset

DMP component	AI4Media_Data_85_WP3_IMAGE_OxfordFlowers_v1 Partner: BSC
Data Summary	Purpose: The Oxford Flower dataset contains images depicting 102 types of flowers. Each image has a class label and a segmentation annotation.

	Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The images in the dataset are a mix of own production from the Visual Geometry Group at University of Oxford and images collected from the web. Expected size: Around 0.7GB total Data utility: It is useful to WP3 partners working in image classification tasks and object segmentation tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website https://www.robots.ox.ac.uk/~vgg/data/flowers/102/. Search keywords: flowers 102, oxford flowers dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://www.robots.ox.ac.uk/~vgg/data/flowers/102/. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as a tgz file. Repository: https://www.robots.ox.ac.uk/~vgg/data/flowers/102/ 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: No license found. Availability for re-use: N/A Usable by third parties after end of project: The data is openly 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 BSC servers. BSC 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

5.1.25 Oxford-IIIT Pet dataset

DMP component	AI4Media_Data_86_WP3_IMAGE_OxfordIIITPet_v1 Partner: BSC
Data Summary	Purpose: The dataset contains images depicting 37 breeds of cats and dogs. Each image has a class label, bounding box of the animal's head and a segmentation annotation. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The images in the dataset have been collected from Catster, Dogster, Flickr and Google image search. Expected size: Around 1.3GB total Data utility: It is useful to WP3 partners working in image classification tasks and object segmentation tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website https://www.robots.ox.ac.uk/~vgg/data/pets/. Search keywords: oxford iiit pet Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from https://www.robots.ox.ac.uk/~vgg/data/pets/. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as a tar.gz file. Repository: https://www.robots.ox.ac.uk/~vgg/data/pets/ 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: Creative Commons Attribution-ShareAlike 4.0 International License Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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

5.1.26 Stanford Dogs dataset

DMP component	AI4Media_Data_87_WP3_IMAGE_StanfordDogs_v1 Partner: BSC
Data Summary	Purpose: The dataset contains images depicting 120 breeds of dogs. Each image has a class label and bounding box. Type/format: Each image is stored as a .jpg file. Re-use of existing data: Yes. Data origin: The images in the dataset come from the ImageNet1K dataset. Expected size: Around 1.2GB total Data utility: It is useful to WP3 partners working in image classification tasks and object detection tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired at the website http://vision.stanford.edu/aditya86/ImageNetDogs/. Search keywords: oxford iiit pet Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is accessible from http://vision.stanford.edu/aditya86/ImageNetDogs/. It is a public dataset. How it will be accessible: The dataset is accessible via direct download. Methods/software tools to access data: Web-browser to download the data as a tar.gz file. http://vision.stanford.edu/aditya86/ImageNetDogs/ 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: No license found.

	Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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	Refer to other national/funder/sectorial/departmental procedures for data management that you may be using (if any)

5.2 Datasets used in the context of WP4

5.2.1 ImageNet-ILSVRC2012 image classification dataset

DMP component	AI4Media_Data_88_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 openly accessible	Data openly accessible: No. The data is accessible through registration on the 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_89_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_90_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_91_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, interestingness 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. provisions for metadata	Is data discoverable: Data is discoverable. The dataset is hosted on the Interdigital 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_92_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_93_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	Data openly accessible: The data is not open yet. The data set is so far owned by

openly accessible	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

5.2.7 FCVID event recognition dataset

DMP component	AI4Media_Data_94_WP4_Video_FCVID_v1 Partner: CERTH
Data Summary	Purpose: Fudan-Columbia Video Dataset (FCVID) is a public video event recognition dataset, containing 91,223 Web videos annotated manually according to 239 categories. It is used by CERTH to evaluate the proposed video event recognition models. Type/format: .AVI/.FLV Re-use of existing data: Yes. Data origin: The dataset can be obtained by sending a request email to the authors at

	https://fvl.fudan.edu.cn/dataset/fcvid/. Expected size: 2TB Data utility: It is useful to WP4 partners to evaluate and benchmark video event recognition models.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable and publicly available after sending a request email to the authors. Search keywords: fcvid dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible upon email request at https://fvl.fudan.edu.cn/dataset/fcvid/. How it will be accessible: Shared through a third-party repository link after the email request has been approved by the authors. Methods/software tools to access data: Web-browser to download the data as zip file. Repository: https://fvl.fudan.edu.cn/ Restrictions on access: The dataset can be obtained after sending an email request to the authors.
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: proprietary, for research purposes only Availability for re-use: The dataset is already publicly available after sending a request email to the authors. Usable by third parties after end of project: Yes. 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 is 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: Data should not be shared without the approval of the authors. Is informed consent for data sharing and long term preservation given: N/A

Other Issues	N/A

5.2.8 YLI-MED event recognition dataset

DMP component	AI4Media_Data_95_WP4_Video_YLI-MED _v1 Partner: CERTH
Data Summary	Purpose: The YLI-MED corpus is an index of videos from the YFCC100M specialized for Multimedia Event Detection (MED) research. The videos are categorized as depicting one of 10 target events, or no target event. It is used by CERTH to evaluate the proposed event recognition models. Type/format: mp4. Re-use of existing data: Yes. Data origin: The dataset contains more than 50K videos available at https://multimediacommons.wordpress.com/yli-multimedia-event-detection-subcorpus/. Expected size: 90 GB Data utility: It is useful to WP4 partners to evaluate and benchmark event recognition models.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable and publicly available. Search keywords: YLI MED dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is already openly accessible at https://multimediacommons.wordpress.com/yli-multimedia-event-detection-subcorpus/. How it will be accessible: Shared through a Amazon S3 data bucket Methods/software tools to access data: Web-browser to download the data manually or to access the AWS CLI. Repository: https://multimediacommons.wordpress.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 Creative Common License. Availability for re-use: The dataset is already publicly available. Usable by third parties after end of project: Yes. 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 is 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: N/A 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_96_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_97_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_98_WP5_VIDEO_MonumentsOfItaly_v1** **Partner: RAI**
Data Summary	Purpose: A collection of videos depicting various Italian monuments. This dataset has been useful as a reference for developments in WP5 about landmark recognition. The processing included 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 a private repository and can 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	Security measures: Data is 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: Possible rights issues related to material included in news items for which license of usage has expired. Is informed consent for data sharing and long term preservation given: No
Other Issues	No

5.3.4 LVIS image dataset

DMP component	**AI4Media_Data_99_WP5_IMAGE_LVIS_v1 Partner: JRC**
Data Summary	Purpose: 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. Type/format: Images, segmentation mask + json annotations Re-use of existing data: Yes, we are reusing an existing dataset Data origin: https://www.lvisdataset.org/dataset Expected size: 30 GB Date utility: 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	Is data discoverable: Data is discoverable. The dataset is hosted in the Network Repository https://www.lvisdataset.org/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 at https://www.lvisdataset.org/dataset by its owners thus, we will not re-share it. 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 (https://www.lvisdataset.org/dataset) Restrictions on access: None
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: 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_100_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_101_WP5_IMAGE_STL10_v1 Partner: QMUL**
Data Summary	Purpose: STL-10 is a CIFAR-10 inspired dataset whose samples are drawn from the 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_102_WP5_TEXT_CCNET_v1 Partner: CEA**
Data	Purpose: CCNet is a large text dataset composed of high-quality monolingual datasets

Summary	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_103_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_104_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.	Is data discoverable: Yes, the data is hosted in the github repository of the authors

provisions for metadata	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_105_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_106_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_107_WP5_VIDEO_HeadEyeMovements360Videos_v1 Partner: UCA
Data Summary	Purpose: The dataset of Head and Eye Movements for 360° Videos is composed of 19 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	Costs for making data FAIR: N/A

resources	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_108_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[11] 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.

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

	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.
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_109_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[12]. 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)[13], audio file metadata (tsv), python scripts for dataset parsing Re-use of existing data: Yes

[12] https://github.com/urinieto/msaf-data/tree/master/SPAM
[13] https://ccrma.stanford.edu/~urinieto/SPAM/SPAM-features.tgz

	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 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_110_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 accessible	Data openly accessible: The dataset is openly 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[14]. 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

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

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_111_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 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	Interoperability: N/A

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 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_112_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_113_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_114_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_115_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

5.3.21 MS COCO dataset

DMP component	AI4Media_Data_116_WP5_IMAGE_MSCOCO_v1 Partner: QMUL
Data Summary	Purpose: MS COCO is a dataset for object detection, object segmentation and image captioning. It contains over 330,000 images, more than 200,000 of which are labeled, featuring 80 object categories. It is used in WP5 to evaluate self-supervised methods via object detection fine-tuning. Type/format: Image (jpeg) Re-use of existing data: Yes Data origin: https://cocodataset.org/ Expected size: Train partition: 18GB, validation partition: 1 GB, Unlabeled partition: 19GB Data utility: It is useful in the context of T5.3 for evaluating self-supervised methods in dense-prediction tasks. In general, this dataset is also useful for any researcher that wants to train deep learning models for object detection and segmentation or image

	captioning.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data are discoverable in the MS COCO website at https://cocodataset.org/ Search keywords: MS COCO dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data is available at the MS COCO website at https://cocodataset.org/ How it will be accessible: The data can be downloaded from the MS COCO website. Methods/software tools to access data: Web browser Repository: The data are hosted at the MS COCO website at https://cocodataset.org/ Restrictions on access: No
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: Each image in the dataset is associated with a unique ID and contains metadata such as the file name, license, and URL. Annotations are provided in json format and include bounding boxes and segmentation masks with object categories, keypoints and captions. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: The dataset's annotations belong to the COCO Consortium and are licensed under a Creative Commons Attribution 4.0 License. The COCO Consortium does not own the copyright of the images. Use of the images must abide by the Flickr Terms of Use. The users of the images accept full responsibility for the use of the dataset, including but not limited to the use of any copies of copyrighted images that they may create from the dataset. 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: 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: The dataset will be downloaded from the original source and will be stored on QMUL's servers. QMUL 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

5.3.22 BVI-DVC dataset

DMP component	AI4Media_Data_117_WP5_VIDEO_BVI-DVC _v1 Partner: BSC
Data Summary	Purpose: The dataset contains 200 video clips at various resolutions, including 4K. These clips are used to generate synthetic real and fake video pairs to train and evaluate Video Super-Resolution detection models. Type/format: Each video is represented by a sequence of 64 .png frames Re-use of existing data: The original 4K clips from BVI-DVC are being used as "real" content. The lower resolution video frames are used to generate the "fake" content, by upscaling them with different Super-Resolution algorithms. Data origin: The dataset is a collection from different sources: Videvo Free Stock Video Footage set, IRIS32 Free 4K Footage set, Harmonics database, BVI-Texture database, MCML 4K video quality database, BVI-HFR database, SJTU 4K video database, LIVE-Netflix database, Mitch Martinez Free 4K Stock Footage set, Dareful Free 4K Stock Video data set, MCL-V database, MCL-JCV database, Netflix Chimera, TUM HD databases, Ultra Video Group-Tampere University database Expected size: 300GB combined for the 200 4K clips Data utility: The dataset was originally created to train deep video compression algorithms. It can also be used to train and evaluate Super-Resolution models, as well as image/video quality assessment methods or other similar tasks.
Making data findable, incl. provisions for metadata	Is data discoverable: The data can be acquired from the website https://fan-aaron-zhang.github.io/BVI-DVC, after filling a registration form. Search keywords: bvi-dvc dataset, deep video compression dataset Versioning: There is one version of the dataset available Metadata creation: Video metadata is included in the README, including resolution and scanning format, chroma sampling, and bit depth
Making data openly accessible	Data openly accessible: The data is accessible from https://fan-aaron-zhang.github.io/BVI-DVC, but requires access granted after filling a form. How it will be accessible: The dataset is accessible via MS OneDrive. Methods/software tools to access data: The dataset is accessible via MS OneDrive. Repository: Data and relevant information are stored at https://fan-aaron-zhang.github.io/BVI-DVC. Restrictions on access: Access is granted after filling a form with personal information.
Making data interoperable	Interoperability: The data is interoperable, as long as original licenses and use restrictions are followed. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A

	Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: This database has been compiled by the University of Bristol, Bristol, UK, comprising sequences originally generated by various sources. All intellectual property rights remain with the originators of each sequence. The test sequences from source (15) shall only be used for academic research (no commercial use). Availability for re-use: N/A Usable by third parties after end of project: The data is openly available and is subject to licenses from creators. 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 BSC servers. BSC 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

5.3.23 Adience dataset

DMP component	AI4Media_Data_118_WP5_Image_Adience_v1 Partner: RAI
Data Summary	Purpose: Adience is a dataset for gender and age classification tasks. It is used by RAI to fine tune the gender classification component. Type/format: jpg Re-use of existing data: yes Data origin: The dataset contains about 27K images available at https://talhassner.github.io/home/projects/Adience/Adience-data.html Expected size: 1.34 GB Data utility: It is useful to WP5 partners to evaluate and benchmark face analysis models.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes Search keywords: adience dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: Third party website

	Methods/software tools to access data: Web browser to download tar.gz files Repository: https://talhassner.github.io/home/projects/Adience/Adience-data.html Restrictions on access: Name and email required
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: Academic research purpose 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: N/A
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

5.3.24 IMDB-Wiki dataset

DMP component	AI4Media_Data_119_WP5_Image_IMDB-Wiki_v1 Partner: RAI
Data Summary	Purpose: IMDB-Wiki is a publicly available dataset of face images of celebrities from IMDB and Wikipedia with age and gender labels. It is used by RAI to train the gender classification component used in UC1. Type/format: jpg Re-use of existing data: yes Data origin: The dataset contains more than 523K images available at https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/ Expected size: 270 GB Data utility: It is useful to WP5 partners to evaluate and benchmark face analysis models.
Making data findable, incl. provisions for metadata	Is data discoverable: Data is discoverable and publicly available. Search keywords: imdb wiki dataset Versioning: N/A Metadata creation: N/A

Making data openly accessible	Data openly accessible: Yes How it will be accessible: Third party website Methods/software tools to access data: Web browser to download tar and tar.gz files Repository: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/ 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: Academic research purpose 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: N/A
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

5.4 Datasets used in the context of WP6

5.4.1 Deepfake Detection Challenge dataset

DMP component	AI4Media_Data_120_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

This dataset is also used in WP4.

DMP component	AI4Media_Data_121_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 is used by CERTH is the context of T6.2 to develop and test new algorithms for the detection of deep fake videos on the web, focusing on facial manipulation, and in the context of T4.3 to facilitate the development and evaluation of explainable deepfake detection methods. 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. It is also useful in the context of T4.3 for producing explanations about the outcomes of deepfake detection. 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	Data openly accessible: The data is already openly accessible on GitHub at

openly accessible	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_122_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_123_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_124_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_125_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 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-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_126_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_127_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_128_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_129_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 openly accessible	Data openly accessible: The data is already openly accessible on the authors' 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

5.4.11 ForenSynths dataset

DMP component	AI4Media_Data_130_WP6_Image_ForenSynths_v1 Partner: CERTH
Data Summary	Purpose: This dataset consists of synthesized images from 11 models (GAN and Deepfakes). The ProGAN-generated images are used for training purposes while all 11 models are used for evaluation purposes. Type/format: Images (png) Re-use of existing data: We re-use the dataset introduced in https://arxiv.org/abs/1912.11035 Data origin: https://github.com/peterwang512/CNNDetection Expected size: 80GB Data utility: CERTH uses this dataset to train and evaluate synthetic image detectors in T6.2. Useful for researchers working on sythetic image detection.
Making data findable, incl. provisions for metadata	Is data discoverable: Data discoverable through this link https://github.com/peterwang512/CNNDetection. Search keywords: ForenSynths dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data are already publicly available in GitHub at https://github.com/peterwang512/CNNDetection How it will be accessible: The repository provides Google Drive links to download the data. Methods/software tools to access data: Web browser. Repository: GitHub (https://github.com/peterwang512/CNNDetection) Restrictions on access: No
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use.

	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 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 CERTH's password protected servers.
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

5.4.12 SynthBuster dataset

DMP component	AI4Media_Data_131_WP6_IMAGE_SynthBuster_v1 Partner: CERTH
Data Summary	Purpose: This dataset contains AI-generated images from 9 different models, namely DALL-E 2, DALL-E 3, Adobe Firefly, Midjourney v5, Stable Diffusion 1.3, Stable Diffusion 1.4, Stable Diffusion 2, Stable Diffusion XL, and Glide. It is combined with real images from RAISE (https://dl.acm.org/doi/10.1145/2713168.2713194) in order to evaluate binary synthetic image detection classifiers. Type/format: Images (png) Re-use of existing data: We re-use the dataset introduced in https://ieeexplore.ieee.org/document/10334046 Data origin: https://zenodo.org/records/10066460 Expected size: 12GB Data utility: CERTH uses this dataset to train and evaluate synthetic image detectors in T6.2. Useful for researchers working on sythetic image detection.
Making data findable, incl. provisions for metadata	Is data discoverable: Data discoverable through this link https://zenodo.org/records/10066460. Search keywords: Synthbuster dataset Versioning: N/A Metadata creation: N/A
Making data openly	Data openly accessible: The data are already publicly available in https://zenodo.org/records/10066460

accessible	How it will be accessible: The repository provides a link to download the data. Methods/software tools to access data: Web browser. Repository: Zenodo Restrictions on access: No
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. 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 shared in Zenodo. 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 CERTH's password protected servers.
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

5.4.13 Latent Diffusion Training-set

DMP component	AI4Media_Data_132_WP6_Image_LatentDiffusion_v1 Partner: CERTH
Data Summary	Purpose: This dataset contains AI-generated images from the Latent Diffusion model. We use it in order to train and evaluate binary synthetic image detection classifiers. Type/format: Images (png) Re-use of existing data: We re-use the dataset introduced in https://ieeexplore.ieee.org/document/10095167 Data origin: https://github.com/grip-unina/DMimageDetection Expected size: 21GB Data utility: CERTH uses this dataset to train and evaluate synthetic image detectors in T6.2. Useful for researchers working on sythetic image detection.
Making data findable, incl. provisions for metadata	Is data discoverable: Data discoverable through this link https://github.com/grip-unina/DMimageDetection. Search keywords: On the detection of synthetic images generated by diffusion models

	Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data are already publicly available in https://github.com/grip-unina/DMimageDetection How it will be accessible: The repository provides a link to download the data. Methods/software tools to access data: Web browser. Repository: GitHub Restrictions on access: No
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. 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 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 CERTH's password protected servers.
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

5.4.14 Diffusion datasets

DMP component	AI4Media_Data_133_WP6_Image_DiffusionDatasets_v1 Partner: CERTH
Data Summary	Purpose: This dataset contains AI-generated images from Guided, Latent Diffusion, Glide, and DALL-E models used in order to evaluate binary synthetic image detection classifiers. Type/format: Images (png) Re-use of existing data: We re-use the dataset introduced in https://arxiv.org/abs/2302.10174 Data origin: https://github.com/WisconsinAIVision/UniversalFakeDetect Expected size: 895MB

	Data utility: CERTH uses this dataset to train and evaluate synthetic image detectors in T6.2. Useful for researchers working on sythetic image detection.
Making data findable, incl. provisions for metadata	Is data discoverable: Data discoverable through this link https://github.com/WisconsinAIVision/UniversalFakeDetect. Search keywords: Towards Universal Fake Image Detectors that Generalize Across Generative Models Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: The data are already publicly available in https://github.com/WisconsinAIVision/UniversalFakeDetect How it will be accessible: The repository provides a link to download the data. Methods/software tools to access data: Web browser. Repository: GitHub Restrictions on access: No
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. 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 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 CERTH's password protected servers.
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

5.4.15 FakeAVCeleb dataset

DMP component	AI4Media_Data_134_WP6_VIDEO_FakeAVCeleb_v1 Partner: CERTH
Data Summary	Purpose: The Fake Audio Visual Celeb (FakeAVCeleb) dataset consists of 20,000 videos, of those only 500 are real and the remaining 19,500 are manipulated. The manipulation methods contained in the dataset are Faceswap, FSGAN, Wav2Lip and

	SV2TTS. We use this dataset for training and testing deepfake detection models in T6.2. Type/format: Videos (mp4) Re-use of existing data: Yes, we re-use already published data. Data origin: https://sites.google.com/view/fakeavcelebdash-lab/?pli=1 Expected size: ~163GB Data utility: Used in T6.2 for training and testing video deepfake detection models.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted here https://sites.google.com/view/fakeavcelebdash-lab/?pli=1 Search keywords: FakeAVCeleb Versioning: Yes Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, at https://sites.google.com/view/fakeavcelebdash-lab/?pli=1 How it will be accessible: The data is hosted here https://sites.google.com/view/fakeavcelebdash-lab/?pli=1. Methods/software tools to access data: Browser. Repository: N/A Restrictions on access: The user should accept a licence agreement first: https://docs.google.com/forms/d/e/1FAIpQLSfPDd3oV0auqmmWEgCSaTEQ6CGpFeB-ozQJ35x-B_0Xjd93bw/viewform
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. Data and metadata vocabularies: N/A. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: https://docs.google.com/forms/d/e/1FAIpQLSfPDd3oV0auqmmWEgCSaTEQ6CGpFeB-ozQJ35x-B_0Xjd93bw/viewform 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

5.4.16 ForgeryNet dataset

DMP component	AI4Media_Data_135_WP6_VIDEO_ForgeryNet_v1 Partner: CERTH
Data Summary	Purpose: The ForgeryNet dataset consists of 221,247 videos, of those only 99,630 are real and the remaining 121,617 are manipulated with 10 total manipulation methods. We use this dataset for training and testing deepfake detection models in T6.2. Type/format: Videos (mp4) Re-use of existing data: Yes, we re-use published data. Data origin: https://arxiv.org/pdf/2103.05630.pdf Expected size: 500GB Data utility: Used in T6.2 for training and testing video deepfake detection models.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is hosted here https://yinanhe.github.io/projects/forgerynet.html#download Search keywords: ForgeryNet dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes How it will be accessible: The data is hosted here https://yinanhe.github.io/projects/forgerynet.html#download Methods/software tools to access data: Browser. Repository: Google drive Restrictions on access: The user should accept a licence agreement first.
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. 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: 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

5.4.17 Korean DeepFake dataset

DMP component	AI4Media_Data_136_WP6_VIDEO_KoDF_v1 Partner: CERTH
Data Summary	Purpose: The Korean DeepFake (KoDF) dataset consists of 237,942 videos, of those, 62,166 are real and the remaining 175,776 are manipulated with 6 manipulation methods. We use this dataset for training and testing deepfake detection models in T6.2. Type/format: Videos (mp4) Re-use of existing data: Yes, we re-use published data. Data origin: https://deepbrainai-research.github.io/kodf/ Expected size: 4000GB (4TB) Data utility: Used in T6.2 for training and testing video deepfake detection models.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, The data can be found at https://deepbrainai-research.github.io/kodf/ Search keywords: Korean DeepFake dataset, KoDF Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, at https://deepbrainai-research.github.io/kodf/ How it will be accessible: The data can be found at https://deepbrainai-research.github.io/kodf/

	Methods/software tools to access data: Web browser. Repository: https://deepbrainai-research.github.io/kodf/ Restrictions on access: The user should accept a licence agreement first.
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. 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 user should accept this licence agreement first. 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

5.4.18 WildDeepFake dataset

DMP component	**AI4Media_Data_137_WP6_VIDEO_WildDeepFake_v1 Partner: CERTH**
Data Summary	Purpose: The WildDeepFake (WDF) dataset consists of 7,314 face sequences extracted from 707 deepfake videos; of those, 3,805 are real and the remaining 3,509 are fake. All of them were sourced from the internet. We use this dataset for training and testing deepfake detection models in T6.2. Type/format: Videos (mp4) Re-use of existing data: Yes, we re-use published data. Data origin: https://github.com/deepfakeinthewild/deepfake-in-the-wild Expected size: 100GB

	Data utility: Used in T6.2 for training and testing video deepfake detection models.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data can be found on GitHub at https://github.com/deepfakeinthewild/deepfake-in-the-wild Search keywords: WildDeepFake Dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, available on GitHub: https://github.com/deepfakeinthewild/deepfake-in-the-wild How it will be accessible: Can be downloaded from GitHub Methods/software tools to access data: Web browser Repository: GitHub: https://github.com/deepfakeinthewild/deepfake-in-the-wild Restrictions on access: The user should accept a licence agreement first: https://docs.google.com/forms/d/e/1FAIpQLSfN-CrnxDz_Furv0KzcNdO_Nzf_3Gpy4s-P4qRjKBJuD2CaEA/viewform
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. Data and metadata vocabularies: N/A. Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Licence prevents the user from re-sharing the dataset: https://docs.google.com/forms/d/e/1FAIpQLSfN-CrnxDz_Furv0KzcNdO_Nzf_3Gpy4s-P4qRjKBJuD2CaEA/viewform Availability for re-use: Yes 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 have NOT provided their consent for the usage of these videos)
Other Issues	N/A

5.4.19 DeepFakes from Different Models (DFDM) dataset

DMP component	AI4Media_Data_138_WP6_VIDEO_DFDM_v1 Partner: CERTH
Data Summary	Purpose: The DeepFakes From Different Models (DFDM) dataset consists of 6,450 fake videos sourced from YouTube and manipulated with 5 manipulation methods. We use this dataset for training and testing deepfake detection models in T6.2. Type/format: Videos (mp4) Re-use of existing data: Yes, we re-use published data. Data origin: https://github.com/shanface33/Deepfake_Model_Attribution Expected size: 183GB Data utility: Used in T6.2 for training and testing video deepfake detection models.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data can be found in GitHub at https://github.com/shanface33/Deepfake_Model_Attribution Search keywords: DeepFakes From Different Models, DFDM dataset, Model Attribution of Face-swap Deepfake Videos Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, available on GitHub at https://github.com/shanface33/Deepfake_Model_Attribution How it will be accessible: From GitHub Methods/software tools to access data: Browser Repository: GitHub Restrictions on access: The DFDM database is released only for academic research. Researchers from educational institute are allowed to use this database freely for noncommercial purpose.
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. 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 DFDM database is released only for academic research. Researchers from educational institute are allowed to use this database freely for noncommercial purpose. Users should sign a form: https://docs.google.com/forms/d/e/1FAIpQLSeM-1pJ13RyPVgF0bGRQtLiupwWDvALD6rKa_Oa8sIluIqtSA/viewform?vc=0&c=0&w=1&flr=0 Availability for re-use: Yes 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 have NOT provided their consent for the usage of these videos)
Other Issues	N/A

5.4.20 DF-Platter dataset

DMP component	AI4Media_Data_139_WP6_VIDEO_DF_Platter_v1 Partner: CERTH
Data Summary	Purpose: The DF-Platter dataset consists of 133,260 videos sourced from YouTube and manipulated with 3 manipulation methods. We use this dataset for training and testing deepfake detection models in T6.2. Type/format: Videos (mp4) Re-use of existing data: Yes, we re-use published data. Data origin: https://iab-rubric.org/df-platter-database# Expected size: ~500GB Data utility: Used in T6.2 for training and testing video deepfake detection models.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data can be found at https://iab-rubric.org/df-platter-database# Search keywords: DF-Platter dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: Yes, available at https://drive.google.com/drive/folders/1GeR-a2LfcMkcY6Qzpv2TP8utLtYFBmTs How it will be accessible: Download as a compressed file Methods/software tools to access data: Browser Repository: Google drive Restrictions on access: The user should accept a licence agreement first: https://iab-

	rubric.org/old/images/pdf/license/LICENSE_AGREE_DB_df-platter.pdf
Making data interoperable	Interoperability: Standard formats are considered in this dataset, the data are easy to download and use. Data and metadata vocabularies: N/A Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: Licence prevents the user from re-sharing the dataset: https://iab-rubric.org/old/images/pdf/license/LICENSE_AGREE_DB_df-platter.pdf Availability for re-use: Yes 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 have NOT provided their consent for the usage of these videos)
Other Issues	N/A

5.4.21 MAFW dataset

DMP component	AI4Media_Data_140_WP6_VIDEO_MAFW_v1 Partner: QMUL
Data Summary	Purpose: MAFW is a large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. Clips in this database come from China, Japan, Korea, Europe, America and India, and cover various themes, e.g., variety, family, science fiction, suspense, love, comedy, and interviews, encompassing a wide range of human emotions. Each clip has been independently labeled 11 times by 11 well-trained annotators. MAFW database has enormous diversities, large quantities, and rich annotations, including: 10,045 number of video clips from movies, TV dramas, and short videos, a 11-dimensional expression distribution vector for each video clip, three kinds of annotations: (1) single expression label; (2) multiple expression label; (3) bilingual emotional descriptive text, two subsets: single-expression set, including 11 classes of single emotions; multiple-expression set, including 32 classes of multiple emotions, three automatic annotations: the frame-level 68 facial landmarks, bounding boxes of face regions, and gender, four

	benchmarks: uni-modal single expression classification, multi-modal single expression classification, uni-modal compound expression classification, and multi-modal compound expression classification. Used in T6.4 for facial expression recognition in video. Type/format: mp4, png Re-use of existing data: We are reusing an existing dataset Data origin: https://mafw-database.github.io/MAFW/ Expected size: 35 GB Data utility: It is useful in the context of T6.6 for evaluating video-based facial expression recognition models. In general, it can be used for evaluating models with spatial-temporal feature learning.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is discoverable in the MAFW website and GitHub: https://mafw-database.github.io/MAFW/ Search keywords: MAFW dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: No, the data are accessible for academic purposes upon contacting the team responsible for developing the dataset. How it will be accessible: Users must contact the team responsible for developing the dataset after completing appropriate documentation including the terms and conditions for the dataset's use. Methods/software tools to access data: N/A Repository: The data are hosted at the MAFW website: https://mafw-database.github.io/MAFW/ Restrictions on access: The team responsible for the dataset must be contacted to allow access to the dataset.
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: User must sign a license agreement: https://github.com/MAFW-database/MAFW/blob/main/academics/mafw-academics-final.pdf Availability for re-use: Yes Usable by third parties after end of project: Yes, 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 QMUL's servers. QMUL 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

5.4.22 Dynamic Facial Expression in-the-Wild (DFEW) video dataset

DMP component	AI4Media_Data_141_WP6_VIDEO_DFEW_v1 Partner: QMUL
Data Summary	Purpose: Dynamic Facial Expression in-the-Wild (DFEW) is a large-scale facial expression database with 16,372 very challenging video clips taken from movies. Clips in the DFEW database are of various challenging interferences, such as extreme illumination, occlusions, and capricious pose changes. Based on the crowdsourcing annotations, we hired 12 expert annotators, and each clip has been independently labeled ten times by them. DFEW database has enormous diversities, large quantities, and rich annotations, including: 16372 number of very challenging video clips from movies, a 7-dimensional expression distribution vector for each video clip, single-labeled expression annotation for classic seven discrete emotions, baseline classifier outputs based on single-labeled annotation. Type/format: mp4, png Re-use of existing data: We are reusing an existing dataset Data origin: https://dfew-dataset.github.io/index.html Expected size: 25.79 GB Data utility: It is useful in the context of T6.6 for evaluating video-based facial expression recognition models. In general, it can be used for evaluating models with spatial-temporal feature learning.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data is discoverable in the DFEW website: https://dfew-dataset.github.io/index.html Search keywords: DEFW dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: No, the data are accessible for academic purposes upon contacting the team responsible for developing the dataset. How it will be accessible: The data can be downloaded from the DFEW project website, after contacting the team responsible for the dataset and being given the password: https://dfew-dataset.github.io/download.html Methods/software tools to access data: N/A Repository: The data are hosted at the DEFW website: https://dfew-

	dataset.github.io/download.html Restrictions on access: The team responsible for the dataset must be contacted to allow access to the dataset.
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: Each video is individually annotated by ten annotators under professional guidance and assigned to one of the seven basic expressions (i.e., happiness, sadness, neutral, anger, surprise, disgust, and fear). Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: DFEW database is available for non-commercial research purposes only. User must sign a license agreement: https://dfew-dataset.github.io/download.html Availability for re-use: Yes Usable by third parties after end of project: Yes, 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 QMUL's servers. QMUL 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

5.4.23 FERV39k video dataset

DMP component	AI4Media_Data_142_WP6_VIDEO_FERV39k_v1 Partner: QMUL
Data Summary	Purpose: The FERV39k dataset is a large-scale multi-scene dataset designed for facial expression recognition in the wild. It contains 38,935 video clips labelled with 7 classic expressions across 22 fine-grained scenes in 4 isolated scenarios. The dataset uncovers several new challenges: 1) difficulty and confusion of 7 basic expression classes; 2) discrepancy across 4 scenarios; 3) unsatisfactory cross-scenario performance; 4) long-tail distribution of expressions and duration. The scenarios and scenes were designed following four reasons: 1) Plenty of video sources and samples; 2) Expandability of 22 fine-grained scenes; 3) Large variations and limited overlapping; 4) Distinct associations with scene context. Type/format: mp4

	Re-use of existing data: We are reusing an existing dataset Data origin: https://wangyanckxx.github.io/Proj_CVPR2022_FERV39k.html Expected size: 13.45 GB Data utility: It is useful in the context of T6.6 for evaluating video-based facial expression recognition models. In general, it can be used for evaluating models with spatial-temporal feature learning.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data are discoverable in the FERV39k website: https://wangyanckxx.github.io/Proj_CVPR2022_FERV39k.html Search keywords: FERV39k dataset Versioning: N/A Metadata creation: N/A
Making data openly accessible	Data openly accessible: No, the data are accessible for academic purposes upon contacting the team responsible for developing the dataset. How it will be accessible: The data can be downloaded from Baidu or Google Drivers, after contacting the team responsible for the dataset and being given the password and link: https://github.com/wangyanckxx/FERV39k Methods/software tools to access data: N/A Repository: The data are hosted at the Baidu or Google Drivers. Restrictions on access: The team responsible for the dataset must be contacted to allow access to the dataset: https://github.com/wangyanckxx/FERV39k
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 annotations of FERV39k are copyright and published under the Creative Commons Attribution-NonCommercial 4.0 International License. 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 QMUL's servers. QMUL 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

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), WP2 (AI Media Observatory), WP7 (contributions to AIoD), WP8 (use cases), WP9 (AIDA), WP10 (open calls) and WP11 (dissemination). **24 non-research datasets** have been identified in the final DMP (12 datasets had been identified in the previous version).

In the following, we present the DMP plan for each of these datasets using the template presented in Section 4 (see Table 1). The Table below briefly summarizes the 24 datasets presented in this section and offers a glance at the structure of the section and its subsections. New datasets that were not included in the initial DMP are indicated with yellow.

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_143_WP1_Text_AI4MediaConsortiumContactInfo_v1	WP1	AI4Media consortium contact info dataset	6.1.1
Data collected in WP2 (European AI vision, policy and common research)			6.2
AI4Media_Data_144_WP2_TEXT-IMAGE_ObservatoryData_v1	WP2	AI Media Observatory Dataset	6.2.1
AI4Media_Data_145_WP2_Text_MediaResponsetoDataCrawls_v1	WP2	Curated dataset of resources on how the media sector responds to content crawling for AI model training	6.2.2
Data collected in WP7 (Integration with AI-on-Demand platform)			6.3
AI4Media_Data_146_WP7_Text_AICafe_v1	WP7	AI-Cafe mailing list members and AI-Cafe participants	6.3.1
AI4Media_Data_147_WP7_Text_AI-Assets_v1	WP7	Candidate AI assets dataset	6.3.2
Data collected in WP8 (Use cases & demonstrators in media, society and politics)			6.4
AI4Media_Data_148_WP8_UserData-TrulyMedia-UC1-ATC_v1	WP8	User data from Truly Media for Use case 1	6.4.1
Data collected in WP9 (Doctoral Academy and exchange programme)			6.5
AI4Media_Data_149_WP9_Text_AIDACourseOfferings_v1	WP9	AIDA course offerings dataset	6.5.1
AI4Media_Data_150_WP9_Text_AIDAStudents_v1	WP9	AIDA students dataset	6.5.2
AI4Media_Data_151_WP9_Text_AIDAMailingList_v1	WP9	AIDA mailing list dataset	6.5.3
AI4Media_Data_152_WP9_Text_AIDAAIEducationalResources_v1	WP9	AIDA AI educational resources dataset	6.5.4
AI4Media_Data_153_WP9_Text_AIDALecturers_v1	WP9	AIDA lecturers dataset	6.5.5
AI4Media_Data_154_WP9_Text_AIExcellenceLectures_v1	WP9	AIDA AI Excellence Lecture Series dataset	6.5.6
AI4Media_Data_155_WP9_Text_AIDACurators_v1	WP9	AIDA curators dataset	6.5.7

AI4Media_Data_156_WP9_TEXT_AIDAWebsiteGoogleAnalytics_v1	WP9	AIDA website analytics dataset	6.5.8
AI4Media_Data_157_WP9_Text_JuniorFellowsExchange_v1	WP9	AI4Media Junior Fellows exchange program dataset	6.5.9
Data collected in WP10 (Community Outreach and Growth)			6.6
AI4Media_Data_158_WP10_Competitive_call_application_datasets_v1	WP10	Competitive call application datasets	6.6.1
AI4Media_Data_159_WP10_sub-granted_projects_dataset_v1	WP10	Sub-granted projects datasets	6.6.2
AI4Media_Data_160_WP10_sub-granted_projects_dataset_v1	WP10	External experts and evaluators datasets	6.6.3
AI4Media_Data_161_WP10_ParticipantsCompetitiveCallsEvents _v1	WP10, WP11	Participants of competitive call related events datasets	6.6.4
Data collected in WP11 (Communication, dissemination, exploitation and sustainability)			6.7
AI4Media_Data_162_WP1_Text_AI4MediaAssociateMembersContactInfo_v1	WP11	AI4Media associate members contact info dataset	6.7.1
AI4Media_Data_163_WP11_Text_NewsletterSubscribers _v1	WP11	AI4Media newsletter subscribers dataset	6.7.2
AI4Media_Data_164_WP11_Text_WebsiteMessages_v1	WP11	AI4Media website messages dataset	6.7.3
AI4Media_Data_165_WP11_TEXT_AI4MediaWebsiteGoogleAnalytics_v1	WP9	AI4Media website analytics dataset	6.7.4
AI4Media_Data_166_WP11_Text_RegistrantstoEvents_v1	WP11	Dataset of registrants for AI4Media events	6.7.5

6.1 Datasets collected in the context of WP1

6.1.1 AI4Media consortium contact info dataset

DMP component	**AI4Media_Data_143_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	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.

metadata	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 WP2

6.2.1 AI Media Observatory Dataset

DMP component	AI4Media_Data_144_WP2_TEXT-IMAGE_ObservatoryExpertsRepo_v1 Partner: LOBA
Data Summary	Purpose: The data was collected in order to build an AI Media expert repository for the AI Media Observatory. The AI Media Expert is an expert directory where visitors of the Observatory can easily search, find and contact a relevant expert within the field of media and AI - whether it be a technical, legal or social expert. The directory aims to help civil society and media professionals who are seeking knowledge on a specific topic on AI in media. Type/format: The data is in textual format, with some elements such as photos being in image format (e.g., JPEG, PNG). Re-use of existing data: No Data origin: The data originates from applications submitted by experts via a form on the Media AI Observatory website. Expected size: The expected size per expert application is relatively small, with the resume limited to 50 words max and the description limited to 150 characters. Depending on the number of expert submissions, the total size is estimated to be a few MBs. Data utility: The visible results of the datasets (meaning the webpage where the experts are featured) aims to help civil society and media professionals who are seeking knowledge on a specific topic on AI in media. However, the dataset is held internally and is not shared, but the information is freely accessible on the page: https://www.ai4media.eu/find-an-ai-media-expert/
Making data findable, incl. provisions for metadata	Is data discoverable: The data is featured in a publicly accessible webpage on the AI Media Observatory. There are no barrier to access. Search keywords: You can find the Expert Repository page on the Observatory by typing in a search engine: AI Media Expert AI4Media. Versioning: N/A Metadata creation: Metadata includes the expert's name, position, and keywords related to their expertise. This metadata is created and managed within the WordPress system used by the website.
Making data openly accessible	Data openly accessible: Not applicable, the data are visible online but the dataset will not be shared for GDPR reasons. How it will be accessible: Online on the AI Media Observatory:

	https://www.ai4media.eu/find-an-ai-media-expert/ Methods/software tools to access data: The data can be accessed using standard web browsers to access to the AI Media Observatory. No special software tools are needed beyond those used for website access. Repository: The data is stored in the WordPress system of the AI Media Observatory website. Restrictions on access: Access will be granted to data subjects who make a specific request to access their personal data. Otherwise, request will not be granted, the data are already visible online in the Observatory.
Making data interoperable	Interoperability: Anyone interested can have a look at the webpage and experts featured in the repository. 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: NA. The data cannot be re-used otherwise another consent is needed from the individuals featured in the expert repository. Usable by third parties after end of project: The data will remain visible on the observatory after the end of the project. Unless expert no longer want to be part of the repository, their data will be removed. Re-use timeframe: The data will remain available during the period the website is maintained, for up to 5 years after the completion of the project. Data quality assurance process: Data are updated as soon as a new form is received or when someone wishes to update or delete their information. They can reach out the observatory through the online contact form or through our email address: observatory@ai4media.eu.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: The costs for long-term preservation include maintaining the website and server hosting.
Data security	Security measures: Security measures include controlled access to the backend of the website, user authentication, firewalls, encryption, and regular back-ups.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The data contain personal data. Is informed consent for data sharing and long-term preservation given: The data will not be shared. Participants are free to update or ask the deletion of their information at any time without any justification. The purpose for the Observatory is to be a long lasting AI4Media milestone. However, if the Observatory is no longer sustained, the website will no longer be supported after (31 August 2029) and the data will be deleted.
Other Issues	N/A

6.2.2 Curated dataset of resources on how the media sector responds to content crawling for AI model training

DMP component	AI4Media_Data_145_WP2_Text_MediaResponsetoDataCrawls_v1 Partner: NISV, KUL
Data Summary	Purpose: A curated repository of third-party resources (e.g. news articles, blogs) that monitors how media organisations are responding to content scraping for AI model training. Type/format: An online spreadsheet with a description of third-party links, organised into topical categories (anti-scraping statement, legal action, licensing deals, new techniques & methods, opinion pieces and analysis - general analysis of the topic). Re-use of existing data: No Data origin: The majority of resources were curated by AI4Media partners. Additionally, an open invitation via an online survey allows crowdsourced contributions from anyone interested in the topic. Expected size: 40-80 entries. Data utility: A curated repository of third-party resources (e.g. news articles, blogs) that monitors how media organisations are responding to content scraping for AI model training – will be used for the AI Media Observatory.
Making data findable, incl. provisions for metadata	Is data discoverable: No Search keywords: No Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: Yes How it will be accessible: Publicly accessible link to the Google spreadsheet with the data: https://docs.google.com/spreadsheets/d/1UZF-H-SfzpeaZ7OjEmbJFippDESyRiIfX3l5l_US7uE/edit Methods/software tools to access data: Through a web browser Repository: Google drive Restrictions on access: No
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: Anyone interested in this dataset is free to download it, perform analysis based on it or combine it with other datasets. Usable by third parties after end of project: NISV and KUL are committed to maintain the repository online for as long as deemed relevant. If at a certain point after the end of the project a decision will be made to stop maintaining the repository, a final

	version of it will be uploaded to the AI4Media Zenodo account. Re-use timeframe: N/A Data quality assurance process: Publicly crowdsourced submissions are reviewed by NISV and KUL teams.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: N/A
Data security	Security measures: N/A
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	No

6.3 Datasets collected in the context of WP7

6.3.1 AI-Cafe mailing list members and AI-Cafe participants

DMP component	AI4Media_Data_146_WP7_Text_AICafe_v1 Partner: GAR
Data Summary	Purpose: GAR organises and moderates regular AI-Cafe sessions since 2019 in the AI4EU Cafe and then continued the AI-Cafes for the AI4Media Project. GAR uses the tool Gotowebinar for the Cafes. In 2021, all former AI4EU Cafe participants were asked if they want to receive future invitations from AI4Media for the upcoming AI-Cafes. Only 10% answered in written form, they would like to receive further email invitations. These participants get regular invitations by email to the Cafe and they are part of a confidential email list. Type/format: Text Re-use of existing data: No Data origin: The email list includes emails from participants who registered to the AI-Café. Expected size: ~400 emails Data utility: It is useful for the AI- Cafe organisation in the AI4Media project.
Making data findable, incl. provisions for metadata	Is data discoverable: No Search keywords: No Versioning: No Metadata creation: No
Making data openly accessible	Data openly accessible: No. The mailing list contains personal information. 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: N/A Data and metadata vocabularies: We use the the following metadata fields: First name, Last name, Email Use of standard vocabularies: N/A Mappings to commonly used vocabularies: N/A
Increase data re-use	Licence: N/A Availability for re-use: In the case the AI-Cafe continues after the AI4Media project ends, then again each list participant will receive an email, if they want to receive further invitations to future Cafes. 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 list is stored in GAR internal servers and only selected GAR employees have access to it.
Ethical aspects	Possible ethical and legal aspects preventing sharing: It contains personal information. Is informed consent for data sharing and long term preservation given: N/A
Other Issues	N/A

6.3.2 Candidate AI assets dataset

DMP component	**AI4Media_Data_147_WP7_Text_AI-Assets_v1 Partner: FHG-IAIS**
Data Summary	Purpose: This dataset is a list of project results from AI4Media consortium members, including contact information (names, affiliation, emails). The collected data is required to contact project partners to publish their project results on the AI-on-Demand platform. Type/format: Excel Re-use of existing data: No Data origin: Data provided by AI4Media partners in work package reports, deliverables and on the Resources Library on the AI4Media website. Expected size: < 1MB Data utility: This data is necessary to keep track of the publication of candidate AI Assets on the AioD platform.
Making data findable, incl. provisions for metadata	Is data discoverable: No, data is not discoverable from outside. The data is stored on the FhG internal storage and is only discoverable for selected employees working in the project. Search keywords: N/A

	Versioning: No Metadata creation: N/A
Making data openly accessible	Data openly accessible: No. How it will be accessible: Restricted access. The data is accessible only by selected employees. Methods/software tools to access data: Microsoft Excel. 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 FhG internal storage and is only discoverable for selected employees in the project. Access to the storage is protected by FhG user access management. 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: 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

6.4 Datasets collected in the context of WP8

6.4.1 User data from Truly Media for Use case 1

DMP component	AI4Media_Data_148_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

6.5 Datasets collected in the context of WP9

6.5.1 AIDA course offerings dataset

DMP component	AI4Media_Data_149_WP9_Text_AIDACourseOfferings_v1 Partner: AUTH
Data Summary	Purpose: This dataset contains information about course offerings from collaborating AIDA full and international members, for the purpose of advertising them to AIDA 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 the details in web applications or provide them in shared excel files and are responsible for maintenance updates. All data is 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 KBs Data utility: The data is useful to AIDA registered students to apply for AIDA course offerings.
Making data findable, incl.	Is data discoverable: Yes, the data will be discoverable in the web, though search engines.

provisions for metadata	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 is 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 is stored at an internal AUTH server in an encrypted format.
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 AIDA program.
Other Issues	No

6.5.2 AIDA students dataset

DMP component	**AI4Media_Data_150_WP9_Text_AIDAStudents_v1 Partner: AUTH**
Data Summary	Purpose: This dataset contains minimal personal identity information about AIDA PhD students (full name, e-mail, affiliation, supervisor id, gender) and details about their

	progress in the form of courses attended, grades, ECTS collected within the AIDA 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 AIDA certificates). Type/format: text Re-use of existing data: No. Data origin: Students enter their personal details through a secure registration login process. Their supervisors are notified and validate this information. Details about the procedure are provided future i-aida.org website updates. Expected size: Few KBs Data utility: The data is useful for offering AIDA 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 the students themselves, registered lecturers that offer the respective courses that the students attended, and AIDA 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	Costs for making data FAIR: N/A

resources	Costs for long-term preservation: N/A
Data security	Security measures: The data is stored at an internal AUTH server in an encrypted form. Proper data security measures are implemented.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The data will not be shared since it contains some personal information about AIDA 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 AIDA program.
Other Issues	No

6.5.3 AIDA mailing list dataset

DMP component	AI4Media_Data_151_WP9_Text_AIDAMailingList_v1 Partner: AUTH
Data Summary	Purpose: This dataset contains business contact information from students and lecturers and interested researches in AIDA 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 KBs Data utility: This mailing list will advertise AIDA activities to the general public.
Making data findable, incl. provisions for metadata	Is data discoverable: No personal data is retrievable except by the 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. Proper data security measures are implemented.
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

6.5.4 AIDA AI educational resources dataset

DMP component	**AI4Media_Data_152_WP9_Text_AIDAAIEducationalResources_v1 Partner: AUTH**
Data Summary	Purpose: This dataset contains information about resources provided by collaborating AIDA full and international members, for the purpose of advertising them to all AI enthusiasts. Such information includes non-personal data (resource title, resource description), and minimal personal data about the contributors (Full Name, contributor affiliation). 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 enter the details in the AIDA website and are responsible for maintenance updates. AUTH moderates all data to resolve inconsistencies. At any time, lectures may alter or remove their personal data and resources 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 resources, having full access to their own content at all times. Expected size: Few KBs Data utility: The data will be useful to everyone to discover high-quality AI educational materials.
Making data findable, incl. provisions for metadata	Is data discoverable: Yes, the data will be discoverable on the web, through search engines. Search keywords: Through web metadata.

	Versioning: Yes, through scheduled website backups. Metadata creation: Resource title, Resource author, Resource Description and Resource type
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/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. 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 is stored at an internal AUTH server in an encrypted format. Proper data security measures are implemented.
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 the AIDA program.
Other Issues	N/A

6.5.5 AIDA lecturers dataset

DMP component	**AI4Media_Data_153_WP9_Text_AIDALecturers_v1 Partner: AUTH**
Data Summary	Purpose: This dataset will contain minimal personal identity information about AIDA Lecturers (full name, e-mail, affiliation). The collected data are necessary for the execution of WP9 Tasks T9.1 and T9.2, for book-keeping and administrative purposes. Type/format: text Re-use of existing data: No

	Data origin: Lecturers will enter their personal details through a secure registration login process. Details about the procedure are analytically provided on the i-aida.org website. Expected size: Few KBs Data utility: The data will be useful for the AIDA website administrators (for book-keeping and administrative purposes)
Making data findable, incl. provisions for metadata	Is data discoverable: No, except by the AIDA website administrators. Search keywords: N/A Versioning: Yes, through scheduled website backups. Metadata creation: No
Making data openly accessible	Data openly accessible: No, because it includes the personal information of lecturers. 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 lecturer data will only be accessible by the lecturers themselves 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 through 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 is stored at an internal AUTH server in an encrypted form. Proper data security measures are implemented.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The data will not be shared since it contains some personal information about AIDA lectures. 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 the AIDA program.
Other Issues	No

6.5.6 AIDA AI Excellence Lecture Series dataset

DMP component	AI4Media_Data_154_WP9_Text_AIExcellenceLectures_v1 Partner: AUTH
Data Summary	Purpose: This dataset contains information about AIDA's AI Excellence Lecture Series provided by collaborating AIDA full and international members, for the purpose of advertising them to AI enthusiasts. Such information includes non-personal data (AI Excellence Lecture title, AI Excellence Lecture Abstract), and minimal personal data about the contributors (lecturer's full name and short bio). The collected data are necessary for the execution of WP9 Task T9.3. Type/format: text Re-use of existing data: No Data origin: Lecturers enter the details in the AIDA web application and are responsible for maintenance updates. All data is moderated by AUTH to resolve inconsistencies. At any time, lecturers may alter or remove their personal data and resources that appear on i-aida.org website, using the web applications or contacting the website moderators (AUTH). Expected size: Few KBs Data utility: The data will be useful to everyone to discover high-quality lecture materials.
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: Excellence Lecture title, Excellence Lecture Abstract, and lecturer full names and short bio.
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/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 is 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. Proper data security measures are implemented.
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 AIDA program.
Other Issues	No

6.5.7 AIDA curators dataset

DMP component	**AI4Media_Data_155_WP9_Text_AIDACurators_v1 Partner: AUTH**
Data Summary	Purpose: This dataset contans minimal personal identity information about AIDA Curators (full name, e-mail, affiliation, specific expertize, AI Educational Taxonomy). The collected data are necessary for the execution of WP9 Tasks T9.1 and T9.2, for book-keeping and administrative purposes. Type/format: text Re-use of existing data: No Data origin: Curators provide their personal details to system administrators. Details about the procedure are provided in the i-aida.org website. Expected size: Few KBs Data utility: The data will be useful for offering and managing AIDA Resources.
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 the personal information of AIDA curators. 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 curator data will only be accessible by the curators themselves and AIDA 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 is 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 format. Proper data security measures are implemented.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The data will not be shared since it contains some personal information about AIDA curators. Is informed consent for data sharing and long term preservation given: Yes. A form will be requested to be signed for every curator involved in AIDA program.
Other Issues	No

6.5.8 AIDA website analytics dataset

DMP component	AI4Media_Data_156_WP9_TEXT_AIDAWebsiteGoogleAnalytics_v1 Partner: AUTH
Data Summary	Purpose: The dataset comprises data and statistics collected from Google Analytics related to the AIDA website. This includes metrics such as visitor numbers, page views, session duration, bounce rates, traffic sources, and user demographics. The purpose of collecting this data is to analyse the performance and reach of the AIDA website to understand user engagement, and to inform strategies for improving communication and dissemination activities. In addition to Google Analytics, additional statistics related to students and course attendance is tracked through WordPress. Type/format: The data is in numerical and categorical format, typically exported as CSV or JSON files. Re-use of existing data: No Data origin: The data originates from Google Analytics tracking implemented on the AIDA website and WordPress feature. Expected size: The expected size of the dataset is relatively small, estimated to be a few megabytes, depending on the volume of web traffic and the duration of data collection. Data utility: This data is useful for monitoring and improving the website's performance and assess the impact of the content on the community.
Making data findable, incl. provisions for	Is data discoverable: The data is identifiable and locatable through standard identification mechanisms.

metadata	Search keywords: Search keywords such as "AIDA website analytics," "visitor statistics," and "web traffic data". Versioning: N/A Metadata creation: Metadata includes details such as the date range of data collection, types of metrics collected, and any filtering or processing applied. This is created following standard web analytics metadata practices.
Making data openly accessible	Data openly accessible: No, the Google Analytics data will not be made openly available due to privacy and security concerns. How it will be accessible: The data is mainly accessed by AUTH for monitoring purposes. Methods/software tools to access data: The data can be accessed using Google Analytics and WordPress dashboards and reports, which are available through standard web browsers. Repository: The data will be stored within the Google Analytics platform and WordPress and accessible through secure login. Restrictions on access: Access will be restricted to authorized project members to ensure data privacy and security.
Making data interoperable	Interoperability: The data is interoperable within the Google Analytics platform, allowing data exchange and re-use within the project team. 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 is assured through regular checks and validation processes, ensuring accuracy and reliability.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: Long-term preservation costs include maintaining access to the Google Analytics account.
Data security	Security measures: Security measures include controlled access to the Google Analytics account, user authentication, firewalls, encryption, and regular backups.
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

6.5.9 AI4Media Junior Fellows exchange program dataset

DMP	AI4Media_Data_157_WP9_Text_JuniorFellowsExchange_v1

component	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

6.6 Datasets collected in the context of WP10

6.6.1 Competitive call application datasets

DMP component	AI4Media_Data_158_WP10_Competitive_call_application_datasets_v1 Partner: F6S
Data Summary	Purpose: The competitive call applications dataset includes data collected during the open call application process. It contains: • 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 two open call procedures. Data origin: The data was collected from applicants who submitted a proposal to AI4Media - Open Call #1 or Open Call #2 via the F6S portal. Data was collected and consolidated by the leader of the evaluation process. Communication channels with applicants were defined in the "Guidelines for Applicants"; relevant messages were collected by the participating consortium members. All the information related to competitive calls was exported to relevant repositories to enable the long-term storage and access to data by consortium members and auditors. Once the competitive calls closed, all data was exported from the F6S portal to the project repository. Expected size: Several GB Data utility: Data collected was 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 was used to create the sub-granted project dataset. The datasetwas also used to generate statistics and reports about the AI4Media project as requested by the Grant Agreement, which was then aggregated into anonymised data that do 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 is stored in a relevant project platform. It will be discoverable only by authenticated 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 platform supports versioning. Metadata creation: A part of the data, such as applicant and evaluation data, isorganised as structured data. Metadata used includes: • How data was created; • Time and date of creation or modification; • Source of data; • Who created the data.
Making data openly accessible	Data openly accessible: The raw data is not 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 is public. How it will be accessible: The collected raw data is stored in arelevant project platform, only accessible by those partners with direct management of the open calls and respective information. Data including anonymised or aggregated data from the open call participation has been made public on the project website and public deliverables. Methods/software tools to access data: Download files via web browser. Repository: The collected data is stored ona relevant project platform. A dataset including anonymised or aggregated data from the open call participation has been made public on the project website. Restrictions on access: Given the confidential nature of the information, access is restricted to those required for managing the data.
Making data interoperable	Interoperability: N/A. Data and metadata vocabularies: Data is compiled using standard spreadsheet formats (Open Office, Excel, Google Sheets, etc.), standard text editors (Open Office, Word, Google format) and 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 have been kept 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 forms have been 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 provided 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

6.6.2 Sub-granted projects dataset

DMP component	AI4Media_Data_159_WP10_sub-granted_projects_dataset_v1 Partner: F6S
Data Summary	Purpose: In each AI4Media competitive open call (Open Call #1 and Open Call #2), several projects were selected and invited 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 was 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 includes (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 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. The dataset is created during the implementation process as information becomes available. Expected size: Several GB Data utility: The data was 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 was also 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 a relevant project platform. It will be discoverable only by autenticated 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 platform supports versioning. Metadata creation: A part of the data, such as applicant and evaluation data, will be organised as structured data. Metadata used includes: • How data was created; • Time and date of creation or modification; • Source of data; • Who created 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 is public. How it will be accessible: The collected raw data is stored in a relevant project platform, only accessible to those partners responsible with direct management of the open calls and respective information. Data including anonymised or aggregated info from the open call participation has been made public on the project website and project deliverables. Methods/software tools to access data: Download files via web browser. Repository: The collected data is stored on a relevant project platform. A dataset including anonymised or aggregated data from the open call participation has been made public on the project website. Restrictions on access: Given the confidential nature of the information, access is restricted to only those required for managing the data.
Making data interoperable	Interoperability: N/A Data and metadata vocabularies: Data is compiled using standard spreadsheet formats (Open Office, Excel, Google Sheets, etc.), standard text editors (Open Office, Word, Google format) and 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 have been kept 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 forms have been 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 provided 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

6.6.3 External experts and evaluators datasets

DMP component	**AI4Media_Data_160_WP10_external_experts_evaluators_datasets_v1 Partner: F6S**
Data Summary	Purpose: As defined in the Guidelines for Applicants for the WP10 open calls, external evaluators were involved in multiple evaluation tasks during the AI4Media programme. This dataset is refers to 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 includes (but is not limited to):

	contracts, payments request, proof of payments, evaluation reports and other. Type/format: Several types of data: documents, emails, texts etc. Re-use of existing data: No, this is an original dataset created in the context of WP10 as a result of the open call procedure. Data origin: External experts' availability was 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 was 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.
Making data findable, incl. provisions for metadata	Is data discoverable: The data is stored on a relevant project platform. It is discoverable only by authenticated 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 platform supports versioning. Metadata creation: A part of the data, such as applicant and evaluation data, will be organised as structured data. Metadata used includes: • How data was created; • Time and date of creation or modification; • Source of data; • Who created data.
Making data openly accessible	Data openly accessible: The raw data is not openly accessible due to AI4Media's commitments to its applicants, sub-grantees and evaluators in relation to personal information and business private information. No data related to evaluators is public. 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: N/A Data and metadata vocabularies: Data is compiled using standard spreadsheet formats (Open Office, Excel, Google Sheets, etc.), standard text editors (Open Office, Word, Google format) and 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 have been kept 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 formshave been 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 provided 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

6.6.4 Participants of competitive call related events datasets

This dataset was collected in the context of both WP10 and WP11.

DMP component	AI4Media_Data_161_WP10_ParticipantsCompetitiveCallsEvents _v1 Partner: F6S
Data Summary	Purpose: During the organisation and delivery of the two competitive calls (AI4Media - Open Call #1 and Open Call #2) several information-related events were organised. The data included in this dataset is related to the participants of these events. Data collected - varying from event to event - included for example name, organisation, e-mail, country, industry. Type/format: Several types of data: documents, emails, texts etc.

	Re-use of existing data: No, this is an original dataset created in the context of WP10 as a result of the open call procedure. Data origin: Participants' information was collected through event pages on the F6S portal or the registration functionalities of other event platforms. Expected size: Several GB Data utility: Data collected was used to identify potential applicants of the competitive calls or those with general interest in the competitive calls results.
Making data findable, incl. provisions for metadata	Is data discoverable: The data will be stored in a relevant project platform. It will be discoverable only by authenticated 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 platform supports versioning. Metadata creation: N/A
Making data openly accessible	Data openly accessible: The raw data is not be openly accessible due to AI4Media's commitments in the framework of the competitive calls and respect to GDPR. Only anonymised or aggregated data will be made public and openly accessible. How it will be accessible: Participation statistics without information about individuals will be included in AI4Media reports. Methods/software tools to access data: N/A Repository: N/A Restrictions on access: Access will be limited to those required for managing the events.
Making data interoperable	Interoperability: N/A. Data and metadata vocabularies: Data is compiled using standard spreadsheet formats (Open Office, Excel, Google Sheets, etc.), standard text editors (Open Office, Word, Google format) and 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 raw data is not openly accessible due to AI4Media's commitments in the framework of the competitive calls and respect to GDPR. Only anonymised or aggregated data will be made public and openly accessible. 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 as it will not be reused.
Allocation of resources	Costs for making data FAIR: N/A 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 have been kept 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. AI4Media participants (data owners) will be able to exercise their right to be forgotten.
Ethical aspects	Possible ethical and legal aspects preventing sharing: The dataset includes personal and business information and therefore can not be shared. Only anonymised or aggregated data will be made public. Is informed consent for data sharing and long term preservation given: The event participants 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

6.7 Datasets collected in the context of WP11

6.7.1 AI4Media associate members contact info dataset

DMP component	AI4Media_Data_162_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 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.

	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 and LOBA servers and access is restricted only to authenticated users. CERTH and LOBA fully comply 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. 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

6.7.2 AI4Media newsletter subscribers dataset

DMP component	**AI4Media_Data_163_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

6.7.3 AI4Media website messages dataset

DMP component	AI4Media_Data_164_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 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 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

6.7.4 AI4Media website analytics dataset

DMP component	AI4Media_Data_165_WP11_TEXT_AI4MediaWebsiteGoogleAnalytics_v1 Partner: LOBA
Data Summary	Purpose: The dataset comprises data and statistics collected from Google Analytics related to the AI4Media website. This includes metrics such as visitor numbers, page views, session duration, bounce rates, traffic sources, and user demographics. The purpose of collecting this data is to analyse the performance and reach of the AI4Media website, and particular web pages for example the AI Media Observatory, to understand user engagement, and to inform strategies for improving communication and dissemination activities. This aligns with the project objectives of enhancing visibility and impact of AI4Media's research and outputs. Type/format: The data is in numerical and categorical format, typically exported as CSV or JSON files. Re-use of existing data: No Data origin: The data originates from Google Analytics tracking implemented on the

	AI4Media website. Expected size: The expected size of the dataset is relatively small, estimated to be a few megabytes, depending on the volume of web traffic and the duration of data collection. Data utility: This data is useful for monitoring and improving the website's performance and assess the impact of our content on the community, including the use and benefits of the AI Media Observatory.
Making data findable, incl. provisions for metadata	Is data discoverable: The data identifiable and locatable through standard identification mechanisms. Search keywords: Search keywords such as "AI4Media website analytics," "visitor statistics," and "web traffic data" Versioning: N/A Metadata creation: Metadata includes details such as the date range of data collection, types of metrics collected, and any filtering or processing applied. This is created following standard web analytics metadata practices.
Making data openly accessible	Data openly accessible: No, the Google Analytics data will not be made openly available due to privacy and security concerns. How it will be accessible: The data is mainly accessed by LOBA for monitoring purposes, but access can be granted to any other AI4Media partner with interest in the information. Methods/software tools to access data: The data can be accessed using Google Analytics dashboards and reports, which are available through standard web browsers. Repository: The data will be stored within the Google Analytics platform and accessible through secure login. Restrictions on access: Access will be restricted to authorized project members to ensure data privacy and security.
Making data interoperable	Interoperability: The data is interoperable within the Google Analytics platform, allowing data exchange and re-use within the project team. 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 is assured through regular checks and validation processes, ensuring accuracy and reliability.
Allocation of resources	Costs for making data FAIR: N/A Costs for long-term preservation: Long-term preservation costs include maintaining

	access to the Google Analytics account.
Data security	Security measures: Security measures include controlled access to the Google Analytics account, user authentication, firewalls, encryption, and regular backups.
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

6.7.5 Dataset of registrants for AI4Media events

DMP component	AI4Media_Data_166_WP11_Text_RegistrantstoEvents_v1 Partner: LOBA, CERTH
Data Summary	Purpose: This dataset contains contact information (including name and email and messages) about individuals who register to attend various AI4Media events. The collected data is necessary for managing AI4Media events. Type/format: The dataset is structured as a relational database or stored in a tabular format like CSV or Excel spreadsheets. Each record represents a registrant and includes fields such as name, email address, affiliation, role, event registered for, registration date, dietary restrictions and any specific preferences or interests related to the event topics. Re-use of existing data: No. Data origin: Data is primarily collected directly from registrants through online registration forms Expected size: < 1MB Data utility: This data is used to facilitate event organisation, communication with registrants, personalisation of the event experience, and post-event follow-ups. It also aids in statistical analysis of participant demographics and interests to improve future event planning and targeting.
Making data findable, incl. provisions for metadata	Is data discoverable: No, data is not discoverable from outside. The data is stored on registration systems used for managing event registrations. The data is only accessed by the event organisers. Search keywords: N/A Versioning: No. 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 event organisers. Methods/software tools to access data: Registration systems used for managing event registrations e.g. eventbrite, google forms, pretix, etc.

	Repository: Yes, on the registration systems. Restrictions on access: Yes, only event organisers.
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: Security practices include data encryption, access controls via authentication and authorisation, and regular security audits.
Ethical aspects	Possible ethical and legal aspects preventing sharing: Dataset includes personal data (name and email, etc.) of registrants and will thus not be shared. Is informed consent for data sharing and long-term preservation given: N/A (Consent is given by the registrants to only use their personal data for event purposes)
Other Issues	No

7. Conclusions

This final version of the DMP identifies all the datasets managed by the AI4Media consortium during the project lifetime, organized per work package. The initial version of the DMP discussed 70 research datasets (19 created within the project by project partners and 51 third-party research datasets used by project partners) as well as 12 non-research datasets. In the present document, we present **142 research datasets in total**, including 81 pre-existing ones and **61 newly-created within the project** in the context of WP2, WP3, WP4, WP5, WP6, and WP8. Most of the datasets are openly available, to the benefit of the broader scientific community. In addition, we describe how we manage and protect non-research data, including **24 non-research datasets** collected in the context of WP1, WP2, WP7, 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 were employed for the design and development of AI methodologies, algorithms, and tools in the context of WP3, WP4, WP5 and WP6, aiming to support the seven use cases; **questionnaire data** including questionnaires collected from project partners, associate members and end-users aiming to identify use case requirements, provide guidance to technical partners for the development of the AI tools but also assess the effectiveness and impact of the developed tools and demonstrators; **user activity data and software analytics** automatically collected during the use cases evalution, with the purpose of evaluating and improving the various AI tools integrated in the demonstrators; **personal data of members of the research, academic, student and business communities** participating in AI4Media educational, outreach and dissemination activities, applying for funding through the AI4Media open calls, or contributing to AI Media Observatory.

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

- **Findable**: The datasets that are made publicly available are uploaded in open repositories like Zenodo or GitHub, thus making this data both easily discoverable and identifiable from the outside. Datasets that are only used internally by project partners are stored on partners' 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 try to make publicly available the research datasets created by consortium members in the context of WPs 2, 3, 4, 5, 6, and 8, wherever possible. The data is shared on open repositories like Zenodo and GitHub or institutional (open) repositories of partners. More specifically, **34 out of the 61 datasets created within AI4Media are already openly shared** with the community and are listed in the project website[15]. The remaining 27 datasets are not openly shared due to various ethical and legal aspects that prevent/prohibit sharing (e.g. some social media data cannot be shared due to social media platforms' terms of use, use case evaluation data cannot be shared because they usually include personal data of end users, datasets of news articles

[15] https://www.ai4media.eu/open-datasets/

cannot be shared because the copyright is owned by the relevant media outlets, datasets collected from media companies' archives cannot be shared due to copyright etc.).

Many of the data used in the project for research and development of new AI methods and tools (in WPs 3, 4, 5, 6) is already **open data**, made openly available by research organizations or media companies. Since it is already open, as a general policy, we do not re-share them. Sharing some of this data is 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 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 are also **privately owned datasets**. Such data has been provided to the project for research purposes and will be only used internally by project partners. Effort is 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 is fully anonymized before being shared. The collected data (whether public, private, or personal) is analysed and analysis results are published as part of public project deliverables or publications that are available in open repositories like Zenodo.

- **Interoperable**: Effort is dedicated to making the data interoperable, wherever possible. The data and metadata vocabularies adopted for each data source have been also discussed.
- **Reusable**: Effort is made to increase the re-use of the data that we plan to make open, through clarifying licenses. Licensing is 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 created within AI4Media are either openly shared (by uploading them in open repositories) or shared internally among specific partners (usually stored on the institutional servers of AI4Media partners). Datasets to be openly shared, are mostly deposited in certified repositories like Zenodo and GitHub 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' institutional 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 provides 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 have been provided to users whose personal data were collected and processed in the context of the AI4Media use cases or other project activities.

The DMP has been a **living document** collaboratively updated throughout the lifetime 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