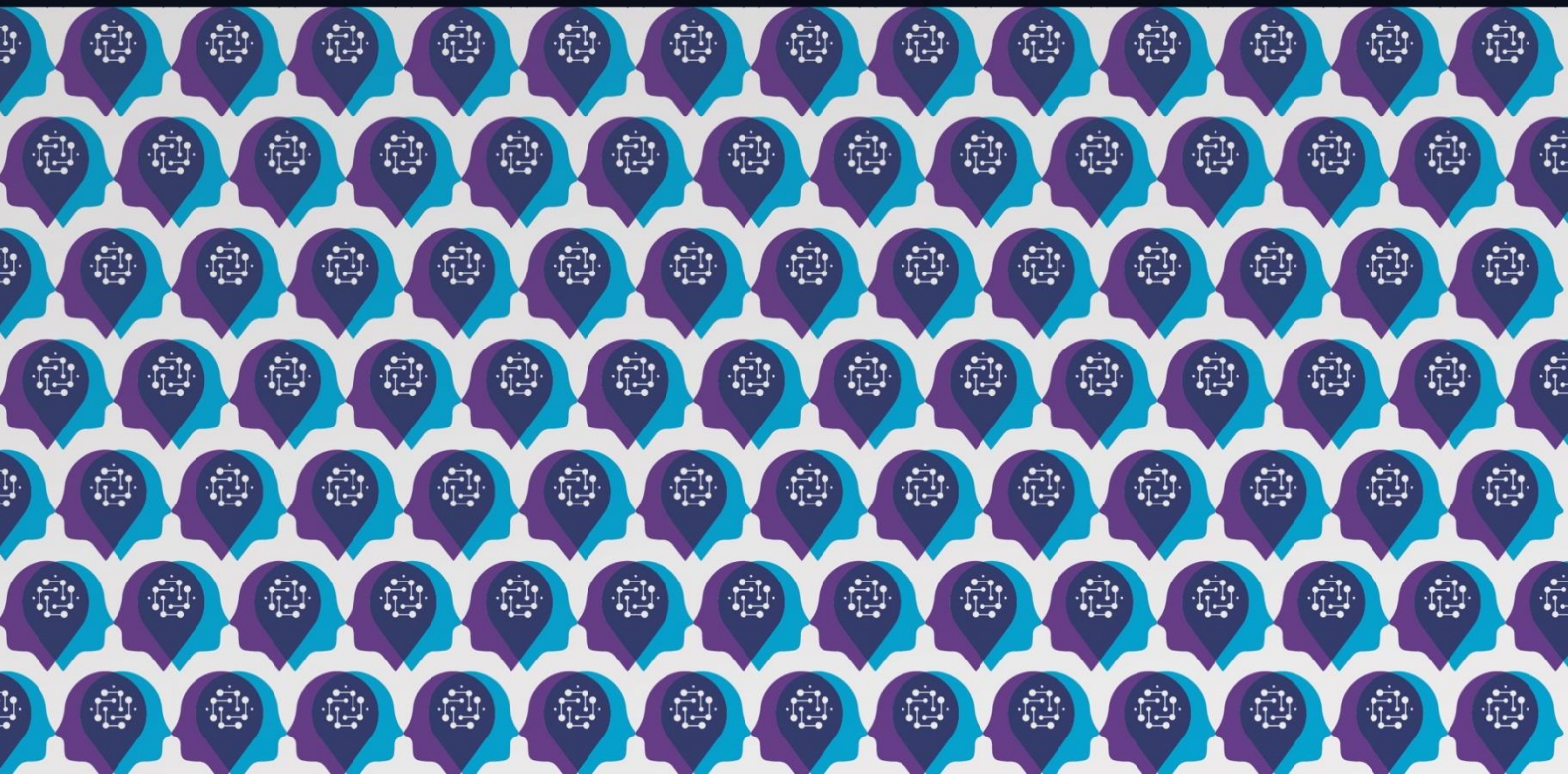
1.

# AI4Debunk

# D1.1 DATA MANAGEMENT PLAN

JUNE 2024

Funded by
the European Union



Grant Agreement No.: 101135757
Call: HORIZON-CL4-2023-HUMAN-01-CNECT
Topic: HORIZON-CL4-2023-HUMAN-01-05
Type of action: HORIZON Innovation Actions

# D1.1 DELIVERABLE TITLE PROJECT HANDBOOK, QUALITY ASSURANCE PLAN AND DATA MANAGEMENT PLAN

| | |
|---|---|
| **Project Acronym** | AI4Debunk |
| **Project Number** | 101135757 |
| **Project Full Title** | Participative Assistive AI-powered Tools for Supporting Trustworthy Online Activity of Citizens and Debunking Disinformation |
| **Work package** | WP 1 |
| **Task** | Task 1.1 |
| **Due date** | 30/06/2024 |
| **Submission date** | 30/06/2024 |
| **Deliverable lead** | UL |
| **Version** | 0.4 |
| **Authors** | Zane Zeibote (UL) |
| **Contributors** | None |
| **Reviewers** | Alessia D'Andrea, Arianna D'Ulizia (CNR) |
| **Abstract** | The Data Management Plan (DMP) of the AI4Debunk project serves as an evolving document which is periodically updated throughout the project implementation, detailing the procedure for data collection, consent procedure, storage, reuse, protection, retention and destruction of data, and confirmation that the usage and sharing of data comply with national and EU legislation. The DMP is created, confirmed, updated, and regularly monitored by the IPR working group (IPR WG) which is approved and supervised by the Innovation Management Team (IMT). The DMP contributes to the D1.1 Project Handbook, Quality Assurance Plan and Data Management Plan and D1.2 Self-assessment plan. |
| **Keywords** | Data, plan, management, fair, use, re-use, open access, ethics. |

## DOCUMENT DISSEMINATION LEVEL

| Dissemination level | |
|---|---|
| **X** | PU – Public |
| | SEN – Sensitive |

## DOCUMENT REVISION HISTORY

| Version | Date | Status | List of contributor(s) |
|---|---|---|---|
| 0.1 | 06/03/2024 | Draft version | Zane Zeibote (UL) |
| 0.2 | 07/06/2024 | Final draft version | Zane Zeibote (UL) |
| 0.3 | 29/06/2024 | Revised final draft version | Zane Zeibote (UL) |
| 0.4 | 30/06/2024 | Final version | Zane Zeibote (UL) |

## STATEMENT ON MAINSTREAMING GENDER

The AI4Debunk consortium is committed to including gender and intersectionality as a transversal aspect in the project's activities. In line with EU guidelines and objectives, all partners – including the authors of this deliverable – recognise the importance of advancing gender analysis and sex-disaggregated data collection in the development of scientific research. Therefore, we commit to paying particular attention to including, monitoring, and periodically evaluating the participation of different genders in all activities developed within the project, including workshops, webinars, and events but also surveys, interviews, and research, in general. While applying a non-binary approach to data collection and promoting the participation of all genders in the activities, the partners will periodically reflect and inform about the limitations of their approach. Through an iterative learning process, they commit to plan and implement strategies that maximise the inclusion of more and more intersectional perspectives in their activities.

## DISCLAIMER

This project has received funding from the European Union's Horizon Innovation Actions under Grant Agreement No 101135757. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

## COPYRIGHT NOTICE



How to cite this report: AI4Debunk (2024). Data Management Plan. *Link from website when deliverable is public*

The AI4Debunk consortium is the following:

| Participant number | Participant organization name | Short name | Country |
|---|---|---|---|
| 1 | LATVIJAS UNIVERSITATE | UL | LV |
| 2 | FREE MEDIA BULGARIA | EURACTIV | BE |
| 3 | PILOT4DEV | P4D | BE |
| 4 | INTERNEWS UKRAINE | IUA | UA |
| 5 | CONSIGLIO NAZIONALE DELLE RICERCHE | CNR-IRPPS | IT |
| 6 | UNIVERSITA DEGLI STUDI DI FIRENZE | MICC/UNIFI | IT |
| 6.1 | CONSORZIO NAZIONALE INTERUNIVERSITARIO PER LE TELECOMUNICAZIONI | CNIT | IT |
| 7 | BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION | BSC | ES |
| 8 | DOTSOFT OLOKLIROMENES EFARMOGES DIADIKTIOY KAI VASEON DEDOMENON AE | DOTSOFT | EL |
| 9 | UNIVERSITE DE MONS | UMONS | BE |
| 10 | UNIVERSITY OF GALWAY | UoG | IE |
| 11 | STICHTING HOGESCHOOL UTRECHT | HU | NL |
| 12 | STICHTING INNOVATIVE POWER | IP | NL |
| 13 | F6S NETWORK IRELAND LIMITED | F6S | IE |

# TABLE OF CONTENTS

# LIST OF TERMS AND ABBREVIATIONS

| | |
|---|---|
| AF | Application form |
| AI | Artificial Intelligence |
| CA | Consortium Agreement |
| CWG | Communication Working Group |
| EC | Ethics Committee |
| DMP | Data Management Plan |
| DoA | Description of Action |
| DX.X | Deliverable X.X |
| GA | General Assembly |
| GDPR | General Data Protection Regulations |
| IT | Information Technology |
| IMT | Innovation Management Team |
| IPR | Intellectual Property Rights |
| KPI | Key performance indicator |
| MX.X | Milestone X.X |
| MST | Management Support Team |
| PC | Project Coordinator |
| PH | Project Handbook |
| PM | Project Manager |
| QAP | Quality Assessment Plan |
| SAP | Self-Assessment Plan |
| SC | Steering Committee |
| SO | Specific objective |
| TX.X | Task X.X |
| WG | Working group |
| WP | Work Package |
| WPL | Work Package Leader |

# EXECUTIVE SUMMARY

The objective of this report is to set main principles and procedures for the AI4Debunk project Data Management Plan (DMP) with respect to data and other research outputs, allocation of resources, data security, ethics and other issues to achieve and fulfil all the contractual obligations that the consortium has acquired with the European Commission. The DMP is structured according to the Horizon Europe DMP template and follows it requirements. This is the initial version of the AI4Debunk DMP and there will be two additional versions developed during the project lifetime, which will take into account achievements during the project lifetime and arising issues following these developments.

The AI4Debunk partners will strive to ensure that most or all the research data created by AI4Debunk becomes available as open data as soon as possible. Restrictions to access some datasets are applied in the case if collected data belongs to a third party which has denied permission for sharing on account of confidentiality and proprietary issues. The consortium members will manage the digital research data generated in the project in line with the FAIR principles according to requirements set out by the Grant Agreement.

This Project Data Management Plan is approved by the AI4Debunk General Assembly on its meeting on 27 June 2024.

# INTRODUCTION

The Data Management Plan (DMP) of the AI4Debunk project serves as an evolving document which is periodically updated throughout the project implementation, detailing the procedure for data collection, consent procedure, storage, reuse, protection, retention and destruction of data, and confirmation that the usage and sharing of data comply with national and EU legislation. The DMP is created, confirmed, updated, and regularly monitored by the IPR working group (IPR WG), which is approved and supervised by the Innovation Management Team (IMT). The DMP contributes to the D1.1 Project Handbook, Quality assurance plan and Data management plan and D1.2 Self-assessment plan.

The potential GDPR issues are tackled by the Ethics Guidelines and monitored by the Ethics Committee accordingly. The main principle in the protection of personal data is personal data minimization – it means that data may be processed ONLY if it is adequate, relevant, and limited to what is necessary for the project.

This DMP will provide details on the research data collected and generated within the AI4Debunk project. It will explain how various data is handled, organized, licensed, and made available to the public, and how it will be preserved after the project is completed.

# 1 DATA SUMMARY

Previous EU projects with similar objectives as the AI4Debunk provide a very good basis on which the project can rely for further developments. Particularly, the data model (schema) will be started similarly as the Edmo dataset, and a part of the AI4Debunk first dataset will come directly from Edmo, given its relevance. The re-use of relevant multimedia content from several reputable fact-checking websites will also be considered. In addition, the responsible partners will make available their own datasets to the consortium and/or will use data provided by other partners.

## 1.1 USE AND RE-USE OF DATA

During the AI4Debunk implementation, different datasets provided to the research community by other consortia like Vera.ai and AI4Media, and other datasets available via public repositories like Kaggle will be gathered to build our systems: for training and testing machine learning models, building the knowledge graphs, testing the different interfaces, etc. Data containing disparate characteristics will be selected to improve the generalisation capability of the implemented systems.

Knowledge graphs will be built for information storage and to provide contextual information to the rest of the AI4Debunk system. Machine learning models will be built for two main reasons:

1. for estimating a disinformation score,
2. for extracting information from data to feed in the knowledge graphs.

For this, data of different types and modalities related to disinformation are required: news articles, related images, news in audio and video formats, information and commentary about news, etc. Thus, the first step will be the analysis of relevant data sources and collection of required data.

To train and test machine learning models, data of different modalities will be collected: mainly text, but also audio and video data, images. This data will be re-used from existing datasets, as well as crawled from social media platforms, discussion forum websites and other sources relevant to tackle the AI4Debunk use cases. The responsible partners will ensure that they have the legal basis and consent to collect and use data for research and academic purposes related to the project activities.

## 1.2 TYPES AND FORMATS OF DATA THE PROJECT GENERATES OR RE-USES

In the project, we will consider experimental and observational data – image and video data, text and audio data, numerical datasets (formats: txt, xlsx, doc, pdf, mp3, m4a, m3u, vtt, etc.) – used for quantitative and qualitative analyses, training and evaluation of machine learning models, building the knowledge graphs. If applicable, in combination with, and provenance of, existing data. Research/experimental outputs will be accessible to user groups defined in the DMP. We will also consider releasing reports from collaborative efforts, such as benchmarking and adaptation of the knowledge graphs.

Data of case studies will be collected by Euractiv Bulgaria and Internews Ukraine – from their platforms. This data will be annotated by journalists regarding their level of disinformation, and the analysis will be provided. Given the amount of data required for the different systems, CNR will complement these datasets with data from other available sources (e.g., the Edmo dataset, Skepticalscience.com, Science.feedback.org).

At the current stage of the project, the datasets used in the project will contain the following data fields (as defined in the GA): the textual statement of the claim, the author, the source, the date of publication of the claim, the audio data, video data and/or images related to the claim, the topic, the keywords, the language, the fact-checking analysis, the rating scale. Therefore, any personal data which we might encounter (e.g., the name of the author of the claim, names and surnames of people cited in the fake statements, etc.) will only be collected and processed where there is a valid legal basis for doing so, or informed consent is obtained. Possible limitations on copyright, privacy and confidentiality are those defined by the individual sources (e.g., fact-checking websites) that have made this data publicly available. The data contained in the datasets of disinformation cases will be made available to the entire consortium for the implementation of the project activities (e.g., for the development and testing of the generative AI detection modules, applications and user interfaces, etc.).

The collaborative platform that we aim to develop as well as the other applications will accumulate and store data, such as texts, images, links to the news articles, etc. They will use similar data formats as described above. In the long-term, the data than can be shared publicly will also be deposited on Zenodo or CLARIN (see Section 2.1), otherwise it will be archived on the responsible partner's storage servers.

Concerning potential anonymization or pseudonymization, it is important to preserve the names of public figures (names of politicians, activists, actors, etc.) in the datasets – for building the knowledge graphs and for training the AI models. It will be ensured that all the data used for the implementation of AI4Debunk grants the right of use for research and academic purpose, therefore re-use of open-source datasets is favoured in the project.

The datasets created for communication, dissemination and exploitation purposes are outlined in the project Application form (WP15, WP16, WP17) and will be used for communicating and disseminating the project activities and results. The plans for dissemination and communication, and exploitation developed by the project will foresee the involvement of organizations representing key stakeholder clusters along the disinformation process chain. The data generated from event registration, event participation or project website will not be shared outside of the project. Webinars and other video recordings may be used by the consortium members for dissemination purposes or to increase engagement in the project. Personal data collected will not be shared with third parties.

## 1.3 OBJECTIVES OF THE PURPOSE OF DATA GENERATION OR RE-USE AND ITS RELATION TO THE PROJECT

From the technology development perspective, the collected and generated data, in general, will be used to develop and test all the AI4Debunk systems: to update the knowledge graphs, to increase the performance of the AI models, and to improve the user experience.

Although there are available datasets which can be re-used for this purpose, they do not cover sufficiently many specific topics that we aim to tackle, namely, the war in Ukraine and climate change. The existing datasets also must be normalized format-wise, since the data come from different sources with heterogeneous data formats. Similarly, the input data to the collaborative platform and other applications will be used to improve those systems. There will be four human-centred online applications developed: a collaborative platform, a web plugin, a smartphone app, and an AR/VR interface with the ultimate goal to provide a European standard debunking API. This will require the input of text data, speech data, image and video data, as well as real-time information.

From the research perspective, the sociological assessment of the resilience mechanisms will include the organization of reports, in depth analysis and focus groups, as well as the organization of surveys and meetings, such as local groups, focus group discussions. It will include analysis of community engagement, and will create a multi-stakeholder dialogue, including the private sector, to achieve the project objectives and achieve the shift in policies and practices to deter disinformation. Gender equality and other engagement processes are also essential for operationalizing policies, programs, and interventions in disinformation mechanisms.

The event registration and participation datasets will be used by the consortium to increase participant

engagement. It will also be used to generate aggregated or anonymized statistics and reports about the participants of interviews and/or events. The personal data that will be collected about participants is forename, surname, and email. The project will organise activities where it may be useful to collect also images and audio recordings of participants for dissemination actions. Upon consent by registrants, the dataset may be used to communicate other events or initiatives organised by AI4Debunk or other projects that pursue similar objectives. The project website and the social media data generated by the project will be used by consortium members to analyse visitors and their engagement, and to improve the engagement strategy. The dataset will also be used to generate statistics and reports about the visitors and participants, which will be aggregated or anonymized data that will not compromise any personal details of the participants nor any other confidential information. Such information will not be shared outside of the project. Webinars and video recordings may be used by the consortium members for dissemination purposes or to increase external engagement in the project.

## 1.4 THE EXPECTED SIZE OF THE DATA INTENDED TO GENERATE OR RE-USE

Overall, the project will handle a significant volume of data, such as several thousands of case studies on fake news related to the war in Ukraine and climate change. These case studies will contain multimedia contents, such as videos, images, and audio recordings. The DMP will be adjusted accordingly to comply with personal data and intellectual property protection regulations.

The current size of the dataset we have collected so far is around 2GB.

## 1.5 THE ORIGIN/PROVENANCE OF THE DATA, EITHER GENERATED OR RE-USED

The data will mainly originate from online platforms and databases. The project will use data from various reputable online sources, including social media platforms, news websites, and fact-checking websites. These sources are essential for gathering real-time and historical data relevant to misinformation and disinformation.

For the collection of data contained in the public and open-source datasets, data from existing highly reputable fact-checking websites will be used. They may contain personal data related to the authors of the claims, people cited in the fake statements, etc. Possible limitations on copyright, privacy, confidentiality are those defined by each source (fact-checking websites).

As mentioned previously, a dataset of fake statements and related multimedia contents (videos, images, audios, etc.) will be created. Euractiv Bulgaria and IUA will gather data from their own sources, such as case studies of fake news. CNR will extract fake statements and related multimedia content from a number of reputable fact-checking websites. They will start from existing datasets available online on Russian propaganda and climate change (e.g., CLIMATE-FEVER[1] consisting of 1,535 real-world claims regarding climate-change) that will be opportunely modified/integrated according to the necessary information to be extracted for the knowledge graph construction. Further data will be gathered from reputable fact-checking websites both specific for War in Ukraine (e.g., EUvsDisinfo, EdMo, VoxUkraine) and for climate change (e.g. Verificat, Eufactcheck, factcheck-org).

Other partners will use existing datasets for training neural networks for deepfake detection in images and videos, i.e., datasets of third parties like FaceForensics++[2]. FaceForensics++ comes with certain limitations of use[3] that the AI4Debunk partners will be complying with.

[1] CLIMATE-FEVER Dataset | Papers With Code
[2] FaceForensics++ Dataset | Papers With Code
[3] https://docs.google.com/forms/d/e/1FAIpQLSdRRR3L5zAv6tQ_CKxmK4W96tAab_pfBu2EKAgQbeDVhmXagg/viewform

Regarding surveys and interviews, data collected through surveys and interviews conducted by project partners will include insights from stakeholders and the public, which are vital for understanding the societal impact of misinformation and refining AI tools accordingly.

Data for the communication and dissemination activities will originate from event registration, participation, webinars, and other video recordings.

## 1.6 USABILITY OF DATA OUTSIDE THE PROJECT

The datasets created or modified within the project and used to build and test all the AI4Debunk systems (AI models, knowledge graphs, user interfaces) are planned to be distributed as open data for unlimited purposes outside the consortium (to the wider community at international level) if the licenses of the source data allow it (regarding the re-used resources). The new versions of re-used dataset which were modified in any way for the purpose of any of our objectives would allow communities outside the consortium to benefit from the improvement brought to them (following the “share alike” principle). Permissive licenses will be selected to allow for the re-use of the data to the widest community possible (academia, industry, etc.).

The project data can be very useful to other academic and research institutions. The data generated from the project, particularly that related to case studies and analytics of fake news, could be of significant interest to the research community, including social and communication sciences, digital humanities, language technology communities, etc.

For public sector and policy makers – to understand the spread and impact of misinformation on society. This information can help in formulating policies and regulations to combat misinformation effectively, enhancing public discourse and national security.

Technology companies, especially those developing AI tools for content moderation, could use the AI4Debunk data (not containing personal data, except if it is included in the media content) to refine their algorithms. This can help improve the accuracy and efficiency of tools used across social media platforms and other online forums to identify and mitigate harmful content.

For educational organizations and NGOs, data from the project can be useful for educational programs aimed at improving media literacy. NGOs focusing on digital rights and freedoms might use the data to advocate for more robust protections against online misinformation.

For media organizations and journalists – to improve their fact-checking processes and reporting accuracy. This would be particularly valuable in strengthening the credibility of media in an era of pervasive misinformation.

Finally, data gathered for communication and dissemination purposes will be used by the consortium to analyse and increase engagement.

# 2 FAIR DATA

This DMP follows EU guidelines and establishes data management procedures according to the general FAIR data principles: findability, accessibility, interoperability, reusability.

The consortium partners will strive to ensure that most or all the research data created by AI4Debunk becomes available as open data as soon as possible. Restrictions to access some datasets are applied in the case if collected data belongs to a third party which has denied permission for sharing on account of confidentiality and proprietary issues. The consortium members will manage the digital research data generated in the project in line with the FAIR principles according to requirements set out by the Grant Agreement.

## 2.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

### 2.1.1 IDENTIFICATION OF DATA BY A PERSISTENT IDENTIFIER

Project partners will primarily use the Zenodo repository for EU-funded research outputs, which will automatically create and register a Digital Object Identifier (DOI) for a record, i.e., a dataset once it is published. The DOI is a globally unique persistent identifier which ensures that the record can be uniquely cited, which is important for reproducibility and attribution of credit. Zenodo registers DOIs with DataCite.

Partners will also share other research outputs that will already have a DOI (e.g., journal publications) on Zenodo. In this case, the DOI will be provided during the upload of the object to Zenodo, so that there is no new DOI created for the same AI4Debunk content.

In addition, for certain AI4Debunk datasets, partners will also consider the CLARIN.eu digital research infrastructure, which is a distributed language resource repository for technology developers, digital humanities and digital social sciences. Thus, the open language resources created by AI4Debunk would be discoverable by a more targeted research community in parallel to the general-purpose Zenodo repository. Similarly to Zenodo, CLARIN data repositories register persistent dataset identifiers which can be used as data citations that are indexed, for instance, by Google Scholar.

If necessary, partners will consider the use of other unique identifiers, such as standard metadata frameworks and ORCID researcher identifiers, as well as other trusted repositories. Suitable publication platforms such as Open Research Europe will be also considered.

### 2.1.2 USE AND CREATION OF THE RICH METADATA

The Zenodo and CLARIN repositories allow for rich metadata to facilitate discovery of the AI4Debunk datasets. The CLARIN metadata records ensure provision of even more detailed language data specific metadata. These metadata categories are used to find and filter digital language resources via the pan European CLARIN Virtual Language Observatory[4] which is widely known and used by the community.

[4] https://vlo.clarin.eu

## 2.2 MAKING DATA ACCESSIBLE

### 2.2.1 DEPOSITING THE DATA IN A TRUSTED REPOSITORY

Each partner is responsible to manage the data they process, store and share. In case if personal data is collected for the purpose of the project activities, it will be secured in a private storage facility of the respective partners, only accessible by the project team.

Information related to events participation datasets will be stored on the project's internal repository to enable access to this data by the consortium members. Information will be stored according to each event organised in the project. Information will be stored as it is received by the project team. This data will not be deposited in any public repositories.

Statistics of the project website and social media profiles will remain stored on the respective analytics platforms to allow for a more thorough and varied analysis. If relevant, data may be exported to the project's documentation repository to enable access by consortium members. Nevertheless, this data will not be deposited in any public repositories.

Webinars and other video recordings will be stored in the project's internal repository (MS Teams, integrated with SharePoint) and/or on public platforms, such as YouTube, managed by H20-People. Consortium members have access to all the recordings saved in the internal repository. Webinars will be made publicly accessible through the project website and social media platforms.

The long-term preservation strategy of AI4Debunk will ensure that both the software tools and the primary qualitative and quantitative datasets produced within the project can be discovered, understood, accessed, and used after the project has been completed, for at least ten years. By the end of the project, these datasets will be deposited in the Zenodo repository by the responsible consortium partners, which ensures sustainable archiving of the final research data.

### 2.2.2 ARRANGEMENTS WITH THE IDENTIFIED REPOSITORY WHERE THE DATA WILL BE DEPOSITED

Zenodo is a general-purpose open-access repository widely used for publishing deliverables and datasets of EU research projects. Zenodo exposes the data to OpenAIRE5, a network of open access repositories to support the EC publication policies. No special arrangements are necessary to deposit datasets in Zenodo.

CLARIN is a specialised open-access European research infrastructure and repository for language resources and technology. Datasets can be deposited in CLARIN also with restricted-access licences, e.g. for research and academic use only, requiring authentication via eduGAIN. No special arrangements are necessary to deposit datasets in CLARIN, however, UL is a member of the CLARIN-LV consortium and can get assistance if necessary.

MS SharePoint is a cloud-based storage service that allows to store, share and collaborate on files online. UL has created an AI4Debunk SharePoint account, integrated with MS Teams and used internally by the consortium for storing and sharing work-in-progress data. However, it will not be used for the long-term preservation and sharing of the final datasets.

### 2.2.3 ASSIGNMENT AND IDENTIFICATION OF THE DATA IN THE REPOSITORY

Zenodo automatically registers a Digital Object Identifier (DOI) for a data record once it is published. The DOI is a globally unique persistent identifier which ensures that the record can be uniquely cited, which is important for reproducibility and attribution of credit. DOIs registered by Zenodo are resolved by DataCite.

CLARIN also automatically registers Persistent Identifiers (PID) for data records upon their publication. CLARIN repositories typically use PIDs that are maintained and resolved by HandleNet.

### 2.2.4 ENSURING OR RESTRICTING THE AVAILABILITY OF THE DATA

To contribute to the open sciences practices, the following approaches will be used:

1. Open access to the datasets will be granted via the mentioned trusted repositories (at the latest, at the time of publication).
2. Publications will be licensed under CC BY (or equivalent); CC BY-NC/ND (or equivalent) is allowed for long-text formats.
3. Provision of physical or digital access to data or other results needed for validation of conclusions in scientific publications and/or patent applications.

In general, full open access will be provided to all project deliverables, including reports on the process, technology benchmark and process characterization. All these deliverables will be fully accessible to enable a broader impact of the project's results. In the case of restricted deliverables deemed to contain documented background, deliverables containing foreground with undetermined IPR/commercial potential, and deliverables used by industrial partners within commercial projects, products, and services, the consortium will follow specific provisions for access to restricted data for verification and research purposes by registering them as sensitive and their access to be limited under the conditions of the Grant Agreement.

Most AI4Debunk datasets (catalogue data, newly created resources) be published under an open license, since they will not contain data that would require personal data protection or other security measures. For the cases when partners generate privative data containing personal information, security measures will be adopted to comply with the General Data Protection Regulation (GDPR, EU Regulation 2016/679).

Final decisions regarding the publication and sharing of data, including licensing and timing will be made after the data sources have been thoroughly selected and evaluated as an integral part of the AI4debunk project.

Webinars and video recordings will be stored on public platforms, such as project website and social media channels. Personal data will not be shared via metadata. Video recordings will be created and managed by the consortium partners organising webinars and training.

The Open Science practices to be adopted in the project are described in detail in the Application form (1.2.6).

### 2.2.5 THE USE OF AN EMBARGO OR PROTECTION OF THE INTELLECTUAL PROPERTY

The embargo with respect to scientific publications will depend on regulation and practices of a particular publisher. There will be no embargo placed in terms of publishing scientific articles or relevant research outputs by the consortium partners.

In the case of intellectual property, according to the Consortium Agreement, the project has created the Intellectual Property Rights Working Group (IPR WG) appointed and steered by IMT, which shall ensure the execution of the IPR as stipulated by the Grant Agreement and activities foreseen by the Project application. The IPR WG will look over the intellectual property issues and make sure that research data and results are published as soon as possible.

### 2.2.6 THE ACCESSIBILITY OF THE DATA THROUGH FREE STANDARDIZED ACCESS PROTOCOL

The data will be accessible through free and standardized access and interoperability protocols as provided by Zenodo, CLARIN and other trusted repositories. Personal data will not be shared in the public domain and will not be publicly accessible.

### 2.2.7 RESTRICTIONS ON THE USE OF THE DATA AND ACCESS TO IT DURING AND AFTER THE END OF THE PROJECT

This will be ensured as much as possible that all the data created is available at least to the scientific community (for reproducibility and other research purpose) if not in the public domain, but primarily we will strive for open-data and open-source material if possible. With respect to the open materials, the distribution of the data will be made in an ethical way through appropriate and trusted data repositories.

In the case of restricted-access data, we will use the access control mechanisms provided by the trusted repositories (e.g. for proving academic affiliation and for accepting the terms of use).

The access to data processed via the AI4Debunk software tools and interfaces will be made available to the wider community (at the international level) through personal accounts with a secure connection and password. To discourage any misuse of the data, the registration of users will be ensured. More details on what personal data will be collected to register a user will be specified in the interim version of DMP.

### 2.2.8 ASCERTAINING THE IDENTITY OF THE PERSON ACCESSING THE DATA

Each partner manages the access rights to its internal repositories and will only share data with duly identified consortium members. The individual who makes a request to access data will use an email that matches a record the consortium already has stored (in an internal repository). A response will be given, or access granted, to the same email address.

In the case of trusted repositories, the identity and academic affiliation will be verified via a global identity federation, like eduGAIN (Education Global Authentication Infrastructure).

### 2.2.9 A NEED FOR A DATA ACCESS COMMITTEE

There is no need for a data access committee. If there will be a need to evaluate or approve access requests to personal or sensitive data, then it will be undertaken by the AI4Debunk Ethics Committee.

### 2.2.10 ENSURING THE OPEN ACCESS TO THE METADATA AND LICENCING UNDER A PUBLIC DOMAIN DEDICATION CC0

The metadata will be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement. It will contain information to enable the user to access the data (if eligible). The only case where access will be limited: if re-using a dataset requires compliance with a restrictive license.

### 2.2.11 THE AVAILABILITY AND FINDABILITY OF THE DATA

Project data will be stored for the duration of the project and at least for 5 years after that by all partners. The datasets and their accompanying metadata used for developing and improving the AI4Debunk systems and distributed for the purpose of reproducibility, will be deposited for long-term archiving in trusted repositories like Zenodo and CLARIN which will ensure that these datasets will remain available and discoverable after the project ends as long as these platforms will be still providing their services.

### 2.2.12 THE REQUIREMENTS WITH RESPECT TO DOCUMENTATION OR REFERENCE ABOUT ANY SOFTWARE TO ACCESS OR READ THE DATA

The documentation or reference about any software necessary to access or read the datasets will be included. In

general, no specific software will be needed to access or read the datasets, since open standards and data formats will be used. However, specific software (e.g. scripts) will be provided as open source if needed for such purpose. The software, AI models, and other systems and their documentation will be distributed for reproducibility purposes through source code versioning platforms like GitHub, ensuring long-term accessibility and versioning of the resources.

## 2.3 MAKING DATA INTEROPERABLE

### 2.3.1 DATA AND METADATA VOCABULARIES, STANDARDS, FORMATS OR METHODOLOGIES FOR ENSURING THE DATA INTEROPERABILITY TO ALLOW DATA EXCHANGE AND RE-USE WITHIN AND ACROSS DISCIPLINES

Zenodo uses the Dublin Core metadata schema to describe datasets. This is a widely adopted standard that provides a simple and flexible way to describe resources using a set of metadata elements such as title, creator, subject, description, publisher, and date. For enhanced interoperability, Zenodo supports Schema.org metadata, which is used by search engines to improve the discoverability of datasets on the web. Zenodo adheres to the OpenAIRE guidelines for metadata, which facilitate the integration and interoperability of research outputs across European repositories and ensure compliance with Open Access policies. Zenodo employs the DataCite metadata schema for DOI assignment, which includes elements like creator, title, publisher, publication year, resource type, and description.

CLARIN uses CMDI (Component Metadata Infrastructure), which is a flexible framework for designing and using metadata schemas. CMDI allows for the creation of domain-specific metadata profiles while maintaining interoperability using standardized components. CLARIN also uses ISOcat and RELcat: registries for linguistic data categories (ISOcat) and relation types (RELcat). They provide standardized vocabularies for describing linguistic data, ensuring consistency and interoperability across different datasets and tools.

As for the content (encoding) of the datasets, commonly used data formats will be exploited, like CSV, JSON, XML, TXT, various open image, video, and audio formats, Unicode, etc. Each dataset and its data format will be described separately upon releasing and depositing in a trusted repository.

### 2.3.2 MAPPINGS TO MORE COMMONLY USED ONTOLOGIES

If standard vocabularies or ontologies cannot be used for some reason, mappings to more common used ontologies will be provided.

### 2.3.3 QUALIFIED REFERENCES[5] TO OTHER DATA

The concrete use of qualified references to other data will be considered for each dataset separately and will be reported in the next versions of DMP. In general, qualified references are an essential requirement for knowledge graphs. They will ensure that the data within the knowledge graph is reliable, traceable, and interoperable, which is critical for creating a robust, useful and re-usable knowledge graph.

[5] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

## 2.4 INCREASE DATA RE-USE

### 2.4.1 THE PROVISION OF DOCUMENTATION NEEDED TO VALIDATE DATA ANALYSIS AND FACILITATE DATA RE-USE

Reproducible analyses methods (software code mainly) will be provided along with all the documentation necessary for reproducibility and re-use through readme and documentation files on GitHub, Zenodo, Hugging Face, and other repositories. This includes the processing of data at all stages.

### 2.4.2 AVAILABILITY OF THE DATA IN THE PUBLIC DOMAIN TO PERMIT THE WIDEST RE-USE POSSIBLE

Whenever possible, datasets created within the project will be made available in the public domain (unless in contradiction with the IPR or GDPR restrictions). Such datasets will be released using standard licenses, like Creative Commons and Open Data Commons. For datasets that will not be possible to release with a public domain license, we will consider the use of more restrictive academic licenses (e.g., the CLARIN ACA and RES licenses[6]) to make such datasets available at least to the research community. In general, AI4Debunk does not plan releasing datasets containing personal data (except limited personal data of public figures), data protected with copyrights, information of a limited use, and any information for which quotation is not allowed by the author/informant.

Collected datasets will be made available only if the reproduction and sharing are allowed by expressed permission of the right holders or by applicable copyright exceptions. Restrictions also apply where data anonymization is not possible, or data availability would jeopardize the project's main aim.

Specific details regarding the future use of the data will be detailed in the sustainability task. Whenever possible, the project consortium intends to utilize open licensing models for research data to encourage the broadest possible re-use.

### 2.4.3 USABILITY OF THE DATA PRODUCED BY THIRD PARTIES, IN PARTICULAR AFTER THE END OF THE PROJECT

All open research data will be accessible for re-use without any embargo, ensuring that the data is freely available and open for re-use immediately upon publication. Final decisions regarding the publication and sharing of data, including licensing and timing, can only be made after the data sources have been thoroughly selected and evaluated as an integral part of the AI4Debunk project.

Personal data, however, will not be usable by third parties. Also, data gathered from communication and dissemination activities will not be shared.

### 2.4.4 THE DOCUMENTATION OF THE PROVENANCE OF THE DATA

The provenance of each dataset (and its subsets if relevant) will be documented. This will include recording the origins of the data, how it was created or collected, and any transformations it underwent. Metadata standards like Dublin Core and DataCite, supported by both Zenodo and CLARIN, will be used.

[6] https://www.clarin.eu/content/licenses-and-clarin-categories

### 2.4.5 RELEVANCE OF DATA QUALITY ASSURANCE PROCESSES

The quality of all project outputs will be assured following the Quality Assessment Plan of the AI4Debunk project.

The task leader will be responsible for ensuring consents are provided when personal data is to be collected and will implement rigorous data validation checks, verifying participant details, and maintaining secure data transmission.

# 3 OTHER RESEARCH OUTPUTS

## 3.1 THE MANAGEMENT OF OTHER RESEARCH OUTPUTS THAT MAY BE GENERATED OR REUSED THROUGHOUT THEIR PROJECTS

Other resources produced by AI4Debunk – AI models and software tools – will be open-source and will be based on open-source resources. The software and AI models developed by the consortium will be properly documented and available via source code and model versioning and distribution repositories like GitHub and Hugging Face.

## 3.2 THE MANAGEMENT, SHARE OR REUSE OF THE RESEARCH OUTPUTS IN LINE WITH THE FAIR PRINCIPLES

Regarding software and AI models, the use of platforms like GitHub and Hugging Face will ensure that these resources are also managed and shared in line with the FAIR principles.

# 4 ALLOCATION OF RESOURCES

## 4.1 COSTS FOR MAKING DATA OR OTHER RESEARCH OUTPUTS FAIR IN THE PROJECT

Currently, this is estimated that making data and other research outputs FAIR in the AI4Debunk project will have no additional costs, since the consortium partners will leverage free services like Zenodo, CLARIN, Hugging Face and GitHub for storing, versioning, sharing and archiving data and other resources.

Nevertheless, partners should be ready to spend 5-10% of time and resources needed for the researcher in charge for processing, storage, sharing and long-term preservation of data.

## 4.2 RESPONSIBILITY FOR THE DATA MANAGEMENT IN THE PROJECT

The coordinator will have the overall responsibility for data management, and it will be responsible for updating the data management plan, with partner contributions. All partners will be responsible for organising data backup, storage, archiving, and depositing their produced data within the repositories.

## 4.3 THE DATA PRESERVATION TERM AFTER THE END OF PROJECT

Data will be preserved for at least 5 years after the end of the project. Using platforms like Zenodo, GitHub, and Hugging Face to share data and software resources ensures long-term preservation as long as there remains a free version of these platforms.

# 5 DATA SECURITY

None of the project data will ever be left inadvertently available. The partner's data security policy should be in place to address the management of security and the security controls. The access to data processed via the software interfaces will be made available to the wider community (at international level) through a personal account with a secure password. To encourage any misuse of the database the registration of users will be ensured. If the knowledge graph aims to enhance security measures or improve public safety, the collection of such data may be deemed necessary to achieve these goals. National security interests may indeed supersede certain EU regulations in specific contexts, particularly if there's a demonstrable need to mitigate security threats or protect public safety.

## 5.1 PROVISIONS FOR DATA SECURITY

The project will use pre-trained large language models (LLM), focusing on open-source models like Llama3-8B, Mistral-7B, etc. The training and use of LLMs locally, avoiding cloud services, will be favoured. However, if the training or use of very large open-source LLMs (10B+ parameters) will be required, for which computing and storage requirements exceed the resources internally available to the project consortium, the use of cloud computing providers will be considered. In such case, the supercomputers like the LUCIA and LUMI located in EU and supported by EU funding will be preferred.

This is possible that the proprietary GPT-4 model will be employed to evaluate some of research methods, although never sending any personal data to their servers.

SharePoint is adopted for internal data sharing among the consortium members, managed by the project coordinator, where all partners have access only after logging in with username and password. Regular backup of the data will be performed to ensure data recovery. In the case if personal data is collected for the purpose of implementing the project activities, such data will be securely stored by the partners, accessible only to the project consortium. Some partners (e.g. F6S) have additional storage solutions with approved and documented processes and procedures, among them data classification, cryptographic controls, access control, removable media, remote access, and possibly encryption.

## 5.2 SAFETY OF THE DATA STORING IN TRUSTED REPOSITORIES FOR LONG TERM PRESERVATION AND CURATION

The data will be safely stored in trusted repositories for long term preservation and curation. Specifically, public datasets will be deposited in the Zenodo open-access repository.

# 6 ETHICS

The project will follow the EU guidelines on fundamental rights specified in the Charter of Fundamental Rights of the European Union and identify four ethical principles that AI system should respect: respect for human autonomy, prevention of harm, fairness, and explicability.

## 6.1 RELEVANCE TO ETHICS DELIVERABLES AND ETHICS CHAPTER IN THE DESCRIPTION OF THE ACTION

An independent Ethics Committee for the project AI4Debunk is created to address the various ethical, social, and technical challenges posed by the proposal and, in particular, using algorithms and technologies such as eye tracking, face recognition, voice analysis using the ML and multimodal AI modules. The Ethics Committee will also ensure an ethical use of the future AI4Debunk tools to be developed and will make sure that all data is used within the project according to the best ethical guidance principles. This committee will play a crucial role in ensuring responsible and beneficial use of these systems throughout the implementation of the AI4Debunk project.

The Ethics Committee is to advise the project coordinator, the partners, as well as the program management to make sure that the highest ethical principles are used throughout the project. It will provide Ethics guidelines and will alert on possible issues, uncertainties and misuses. The Ethics Committee will provide guidance for both the technologies developed and the tools to be tested within AI4Debunk. It will cover areas like safety, transparency, privacy, use of data, and possible future uses of the tools.

The Ethics Committee involves at least 5 experts proposed by the partners and selected by the project's coordinator. One independent ethics advisor from an independent legal and ethics department from UL will be appointed as Main Ethics Advisor and will chair the Ethics Committee and follow the implementation of the project, advice partners and undertake responsibility for preparing guidelines and other deliverables.

The Ethics Committee will be independent from commercial and political interests to maintain its credibility and impartiality. In the deliverable on guidance, it will provide mechanisms for accountability and transparency in decision-making processes. They also will be independent from the partners' organizations. Ethics will be taken into consideration in the way data is collected, stored and regarding who can visualise and use it. The project will work to ensure that management of personal data is compliant with GDPR and other applicable legal frameworks related to personal data protection.

## 6.2 INFORMED CONSENT FOR DATA SHARING AND LONG-TERM PRESERVATION IN QUESTIONNAIRES DEALING WITH PERSONAL DATA

The AI4Debunk project partners will ensure that the informed consent will be obtained when it will be necessary. Where permission to share personal data has not been granted, these cases will be anonymised before the dataset is shared with the consortium. The questions which should be asked to persons, target groups, as well as the process for receiving informed consent are described in the AI4Debunk Ethics Guidelines. These Guidelines also include questions on the informed consent in a case if the collected data must be stored for a long-term preservation.

Some of the collected data will be shared with other AI4Debunk consortium members if it will be necessary for the achievement of project's objectives; it will be done in compliance with the GDPR. The media information related to project's activities will be shared in the social media, project website and project repository. Other personal data collected will not be shared and will not be available to third parties.

By proactively addressing personal data issues, the AI4debunk partners under the supervision of the Ethics Committee will minimize and prevent potential negative consequences of applied methodologies and techniques

and ensure that their use aligns with ethical and societal values.

# 7 OTHER ISSUES

## 7.1 OTHER NATIONAL/FUNDER/SECTORAL/DEPARTMENTAL PROCEDURES FOR DATA MANAGEMENT

Data management will be compliant with the European laws regarding data security and the protection of privacy (e.g., GDPR). The data protection issues are described in the AI4Debunk Ethics Guidelines and will be monitored by the Ethics Committee.

In addition to the data management procedures described in this DMP, other data management procedures might be used and described in a follow-up version of the DMP.

# 8 CONCLUSION

This DMP will follow EU guidelines and establishes data management procedures according to the general FAIR data principles: findability, accessibility, interoperability, reusability.

The project will make a use of the previous EU projects with similar objectives as the AI4Debunk which provide a very good basis for further developments. Particularly, the data model (schema) will be started similarly as the Edmo dataset, and a part of the AI4Debunk first dataset will come directly from Edmo, given its relevance. The re-use of relevant multimedia content from several reputable fact-checking websites will also be considered. In addition, the responsible partners will make available their own datasets to the consortium and/or will use data provided by other partners.

The consortium partners will strive to ensure that most or all the research data created by AI4Debunk becomes available as open data as soon as possible. Restrictions to access some datasets are applied in the case if collected data belongs to a third party which has denied permission for sharing on account of confidentiality and proprietary issues. The consortium members will manage the digital research data generated in the project in line with the FAIR principles according to requirements set out by the Grant Agreement.

None of the project data will ever be left inadvertently available. The partner's data security policy should be in place to address the management of security and the security controls. The access to data processed via the software interfaces will be made available to the wider community (at international level) through a personal account with a secure password. To encourage any misuse of the database the registration of users will be ensured. If the knowledge graph aims to enhance security measures or improve public safety, the collection of such data may be deemed necessary to achieve these goals. National security interests may indeed supersede certain EU regulations in specific contexts, particularly if there's a demonstrable need to mitigate security threats or protect public safety.

The project will follow the EU guidelines on fundamental rights specified in the Charter of Fundamental Rights of the European Union and identify four ethical principles that AI system should respect: respect for human autonomy, prevention of harm, fairness, and explicability.

This is the initial version of the AI4Debunk DMP, which will be further complemented and updated by project partners in M22 and M44.