DisAI.eu

D5.2 Data Management Plan – Version 2

Project Title	Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies
Contract Nº.	101079164
Type of Action	HORIZON-CSA
Topic	Disinformation Combating
Project start date	1st Dec 2022
Duration	36 months

Funded by
the European Union

Deliverable title	Data Management Plan – Version 2
Deliverable number	D5.2
Deliverable version	2.1
Contractual date of delivery	30 Apr 2024
Actual date of delivery	30 Apr 2024, updated 1 July 2024
Nature of deliverable	Report (R)
Dissemination level	Public (PU)
Work Package	WP5
Task(s)	T5.3
Partner responsible	KInIT
Author(s)	Marián Šimko (KInIT), Matúš Pikuliak (KInIT), Michal Gregor (KInIT)

Abstract	This is the second version of the data management plan (DMP) outlining how the research data are collected or generated and how it will be handled during a project, and after it is completed.
Keywords	Data Management Plan, Desinformation Combating, Multilingual Claim Retrieval Dataset

History of Changes

Version	Date	Description/Note
1.0	31 May 2023	Version 1 of this deliverable submitted by the contractual date of delivery
2.0	30 April 2024	Version 2 of this deliverable submitted by the contractual date of delivery: ● Update note on the utilised MultiClaim dataset provided (end of section 3.1)
2.1	1 July 2024	Incorporated feedback based on the the mid-term review: ● Added section 3.2 *MultiClaim Dataset's metadata* ● Added section 5 Broader Utility and Applications of the MultiClaim Dataset Beyond the Project ● Summary of Changes replaced by History of Changes at the beginning of the document

Table of Contents

1 Introduction

This document is the second version of the data management plan (DMP) for the DisAI project. The DMP is a living document: it will be updated continuously. DMP includes: (1) What data will be collected/generated; (2) What standards will be used; (3) What types and format of data will be created; (4) How will metadata be generated; (5) How will the data be documented; (6) What data will be exploited, shared, made open; (7) How and by whom will data be curated and preserved.

2 General Principles of Data Management

All the data will be collected, processed, and stored respecting the GDPR (General Data Protection Regulation)[1] requirements. It will be securely stored and anonymized where possible under the supervision of each consortium member's Data Protection Officer (coordinated by the KInIT's Data Protection Officer).

In this project, we foresee utilisation of the following types of data:

1. Research Data
    a. The research project data – data that will be utilised within the exploratory research project (WP2). These are the primary research data in the project.
    b. Other scientific data – data that will be utilised within other work packages. These include, e.g.:
        - Datasets used as a part of replication studies (WP1),
        - Datasets used as a part of shared task (WP1)
        - Submissions of papers and associated research artefacts (annotations, models, ...) for workshops and/or summer school (WP1)
2. Shareable Information Artefacts
    a. Webinar recordings - video recordings of four planned webinar activities focused on research topics related to capacity building focus areas of the project (WP1), event. webinars on research management (WP3)
    b. Training workshop materials - materials used for training workshops (both research and research management; WP1, WP2)
    c. Project website and other propagation outputs (WP4)
3. Other internal project documentation – other documentation created within the project as internal documents or documents used as deliverables of the project.
4. Personal data for the purpose of communication (e.g. emails for newsletters) or as a necessary part of task realisation (e.g., personal data of summer school participants to arrange accommodation)

[1] Consolidated text: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)

3 Research Data

In this part, we describe the research data utilised across the project using the Horizon Europe Data Management Plan Template.

3.1 MultiClaim Dataset

Dataset summary	**Responsible partner: KInIT** **Re-use of existing data: No, this is a new dataset introduced by us.** **Type/format: Collection of texts, documented in the dataset repository** **Purpose: A claim matching dataset for fact-checking. We have collected texts from social media platforms and appropriate fact-checks that fact-check the content from these posts. The purpose is to develop claim matching methods that are able to recommend appropriate fact-checks to arbitrary input texts.** **Expected size: 200k fact-checks, 30k social media posts, 30k pairs.** **Data origin: We have collected the dataset from social media and fact-checking platforms.** **Data utility: Data will be usable by all three teams within the DisAI research component, as well as other researchers that are concerned with fact-checking or information retrieval.**
Findability	Is data discoverable: Yes, the dataset is available at Zenodo. Naming conventions: No Metadata creation: Metadata are documented at Zenodo. Search keywords: Keywords are provided at Zenodo. Findable metadata: Metadata are documented at Zenodo.
Accessibility	**Repository:** Trusted repository: Yes, the dataset is available at Zenodo. Have you explored appropriate arrangements with the identified repository where your data will be deposited. Yes Data identifier: Yes, Zenodo provides data identifiers **Data:** Data openly accessible: The data will be available upon request. The dataset will not be fully open because of ethical concerns (see the ethical section in the accompanying paper for further discussion) How data will be accessible: The dataset is available at Zenodo. Restrictions on use: Yes, the dataset will be available only for research purposes. Identity assertion: The identification will be provided by Zenodo. Data access committee: Yes. **Metadata:** Openly available and licenced: Yes

	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? Indefinitely. Methods/software tools to access data: The code to work with the data will be published in a separate code repository on GitHub soon.
Interoperability	Data and metadata vocabularies: The data are stored in a common CSV format. Mappings to commonly used vocabularies: NA Qualified references: No
Reusability	Documentation: The dataset is documented at Zenodo. The code to work with the data will be published in a separate code repository on GitHub soon. Licence: The licence will be limited to research only purposes, and it will not be able to share the data further (see the ethical section in the accompanying paper for further discussion). Usable by third parties after end of the project: The data will be available indefinitely. Data provenance: NA Data quality assurance process: Data quality is described in the accompanying paper.
Other research outputs	Are there any other research outputs that may be generated or re-used throughout the project? No Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles? No
Allocation of resources	Costs: Estimate the costs for making your data FAIR. Describe how you intend to cover these costs. For ethical reasons, this dataset will not be FAIR. Data management: Clearly identify responsibilities for data management in your project. NA Long term preservation: Describe costs and potential value of long term preservation. Zenodo will host the data in the long term.
Security	Security measures: The data is available at Zenodo and they provide their own security measures. Otherwise, the data is stored in our internal disk spaces not accessible from outside the organisation. Repositories policies and procedures: Yes, the data is available at Zenodo
Ethical aspects	Possible ethical and legal aspects preventing sharing: See the ethical section in the accompanying paper for further discussion. Is informed consent for data sharing and long term preservation given: NA
Other issues	Refer to other national/funder/sectorial/departmental procedures for data management that you may be using (if any). NA

Update note on the utilised MultiClaim dataset: Given the focus of its research questions, CBFA 2 “Multimodal Natural Language Processing” requires multimodal (vision-language) data. While no new dataset was collected to meet these requirements and all research activities are still based on the MultiClaim dataset, additional visual content from already included posts and fact-checks is being downloaded. This additional content is only used internally and not distributed to other parties with the original dataset.

3.2 MultiClaim Dataset’s Metadata

To make the document more self-contained and to aid researchers in understanding how to effectively utilise the MultiClaim dataset, we are going to provide a brief description of the dataset’s metadata (as also defined at Zenodo) in this section.

General Information

- Dataset Title: MultiClaim: Multilingual Previously Fact-Checked Claim Retrieval
- Version: 1.0
- Publication Date: 2023-05-12
- Publisher: Zenodo
- DOI: 10.5281/zenodo.7737983

Dataset Description

- Purpose: Training and testing models for disinformation combating
- Contents:
    - 206,000 claims fact-checked by professional fact-checkers.
    - 28,000 social media posts, each associated with at least one claim.
- Key Features:
    - Multilingual dataset with claims and posts in various languages and their English translations.
    - Mapping between fact-checks and social media posts.
    - Detailed information about fact-checks (claim, instances, title).
    - Social media post data (instances, OCR transcripts, verdicts, user-written text).

Access and Usage

- Access Status: Restricted
- Terms and Conditions:
    - Agree to specific terms and conditions.
    - Request access using an official email address from a university, faculty, or research institution.
    - Use the dataset strictly for research purposes.
    - Do not attempt to identify or contact the authors of the social media posts.

    - Cite associated papers when using the dataset in any publication, project, or tool.
- Potential Applications (examples):
    - Developing information retrieval models that assign appropriate claims to social media posts.
    - Training and evaluating models for disinformation detection and fact-checking;
    - Conducting research on multilingual approaches to combating disinformation;
    - Analysing the spread and characteristics of disinformation across different languages and platforms;

By leveraging this dataset, researchers can contribute to the development of more effective tools and strategies for identifying and countering disinformation across multiple languages and social media platforms.

Creators and Affiliations

- Creators: Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Šimko, Juraj Podroužek, Mária Bieliková
- Affiliations: Kempelen Institute of Intelligent Technologies

Related Resources

- Paper: https://aclanthology.org/2023.emnlp-main.1027/
- Preprint: https://arxiv.org/abs/2305.07991
- GitHub Repository: https://github.com/kinit-sk/multiclaim

Hosting and Community

- Hosted on Zenodo.
- Part of the EU Open Research Repository (Pilot) community.
- Indexed in OpenAIRE.

4 Other Relevant Data and Considerations

4.1 Shareable Information Artefacts

4.1.1 Webinar recordings

Webinar recordings will be shared at the Youtube platform, using KInIT's Youtube channel. We will provide appropriate descriptions. As a part of the public video platform Youtube, the findability, accessibility and reusability will be guaranteed implicitly. Interoperability is supported either implicitly or by using 3rd party tools.

4.1.2 Project web site and other propagation outputs

Project web site is created using the web content management system Wordpress. It is findable and available on the world wide web using conventional search engines or via direct URL, https://disai.eu. Relevant information artefacts will be primarily of text modality, hence the interoperability and reusability will be assured.

4.2 Other internal project documentation

The internal project documentation is stored in a cloud storage that is part of Google Workspace. It is allocated in the shared drive *EXT DisAI (KInIT, DFKI, CERTH, UCPH)* with access for the project coordinator and all partners. Project coordinator has the administrative rights and is responsible for data storage and access management. Accesses are granted based on the email groups aliases, which are managed by the project coordinator. Continuous data backups are created inline with the KInIT internal directive.

None of this documentation is intended to be shared publicly. Final public versions of documents will be shared using relevant means for that particular document (e.g. sharing on the web site, submitting to the portal, ...)

4.3 Personal data

When processing personal data, we will ensure compliance with data protection regulations including Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, repealing Directive 95/46/EC (General Data Protection Regulation).

All data (including personal data) will be securely stored. In case of publishing any datasets for the wider scientific community (e.g., for the planned shared task), we will publish only anonymised data (e.g., we will publish only content labels without any identification of the annotators giving the label). We will also comply with other legal requirements, e.g., we will not publish content that might be protected by the copyrights or if publication of such content would breach the terms and conditions of the source where the content was collected (e.g., in case of Twitter data, it is possible to publish only individual tweet IDs, not the content of the tweets itself; similarly, for news articles, the dataset might contain only the URLs instead of the full texts).

5 Broader Utility and Applications of the MultiClaim Dataset Beyond the Project

While the primary purpose of collecting and developing the MultiClaim dataset was to support the research activities within the DisAI project, especially the work on claim matching and fact-checking, this dataset has significant potential for broader utility and applications across other research domains.

The dataset represents a valuable resource for researchers working on a range of problems related to online disinformation, natural language processing, and social computing. Some potential areas where this dataset could be leveraged include (inexhaustive):

- **Misinformation detection and analysis:** Beyond just claim matching, the dataset can support development of models and techniques for identifying and characterising misinformation narratives across social media.
- **Integrating fact-checking signals into content ranking and recommendation:** The dataset can be used to develop techniques for integrating fact-checking signals and credibility indicators into content ranking and recommendation algorithms. For example, researchers can explore ways to down-rank or provide context for content that has been associated with false claims, or to surface fact-checks and debunking information proactively to users who have been exposed to misinformation. This can help to create healthier information ecosystems and mitigate negative impacts of misinformation, which are now often being harmfully amplified by recommendation algorithms.
- **Cross-platform information diffusion:** By connecting fact-checks to social media posts, the dataset should in principle be able to support studies of how claims and narratives spread. Note, however, that its level of usefulness for this purpose would need to be evaluated using an exploratory analysis targeting this particular aspect.
- **Linguistic analysis of false claims:** The textual data can enable research into the linguistic signatures and stylistic patterns of false or misleading claims.
- **Temporal evolution of narratives:** The longitudinal nature of the data allows tracking the emergence and evolution of false narratives over time. Again, the actual usefulness for this purpose would need to be evaluated experimentally.
- **Sociological and psychological aspects:** If the data could be paired to additional signals such as engagement metrics, it might also be useful for research into the sociological and psychological factors driving the spread of misinformation.

We encourage external researchers to utilise the data in novel and impactful ways that go beyond the core focus areas of DisAI. We believe that enabling such extended applications can significantly increase the overall scientific value and real-world impact of the dataset.