



ALFIE

ASSESSMENT OF LEARNING TECHNOLOGIES AND FRAMEWORKS
FOR INTELLIGENT AND ETHICAL AI

D1.4 Data Management Plan

Deliverable No:	D1.4
Deliverable Name:	Data Management Plan
Work Package:	WP1
Task:	T1.4
Dissemination Level:	PU
Deliverable Type:	DMP
Lead Organization:	EHU
Submission month:	M6
Date:	31 March 2025



Funded by
the European Union

Funded by the European Union under Grant Agreement 101177912. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.



Project description

Acronym	ALFIE
Title	Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI
Coordinator	CATALINK
Reference	101177912
Type	Research and Innovation Actions (RIA)
Programme	Horizon Europe (HORIZON)
Topic	HORIZON-CL2-2024-TRANSFORMATIONS-01-06 Beyond the horizon: A human-friendly deployment of artificial intelligence and related technologies
Start	01.10.2024
Duration	36 months
Website	alfie-project.eu
Consortium	Catalink Limited (CTL), Cyprus, Coordinator University of Brighton (UoB), United Kingdom Centre for Research & Technology Hellas (CERTH), Greece Distributed Analytics Solutions LTD (DAS), United Kingdom Edge Hill University (EHU), United Kingdom Kempelen Institute of Intelligent Technologies (KINIT), Slovakia Universitat Autònoma de Barcelona (UAB), Spain Robert Bosch Espana SLU (Bosch), Spain Eindhoven University of Technology (TU/e), Netherlands Diadikasia Business Consulting (DBC), Greece

Document history

Version	Date	Beneficiary	Description
0.1	19.03.2025	EHU	Structure proposal and initial draft.
0.2	24.03.2025	CERTH, CTL	Reviewed the document and proposed amendments.
0.3	25.03.2025	EHU	Resolution of review comments.
1.0	28.03.2025	CTL	Final version.

Authors

Partner	Name(s)
EHU	Yannis Korkontzelos, Marcello Trovati, Nonso Nnamoko

Contributors

Partner	Contribution type	Name
All	Filled in Data Management Plan Questionnaire	
CERTH	Review	Stamatis Samaras
CTL	Review	Christos Avgerinos

Glossary

Acronym	Definition
AI	Artificial intelligence
AutoDW	AutoML Data Warehouse
DMP	Data Management Plan
DPIA	Data Protection Impact Assessments
EC	European Commission
ETD-Hub	EthiTech Dialogue Hub
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
GA	Grant Agreement
GDPR	General Data Protection Regulation
RRI	Responsible Research and Innovation

Executive Summary

This deliverable presents the first, early version of ALFIE's Data Management Plan (DMP). It details the data management strategy for the ALFIE project, which aims to bridge AI accessibility, availability and usage needs and associated legal and ethics requirements expressed by citizens, policymakers, AI professionals, and legal entities (ETD-Hub) with an advanced implementation platform that seamlessly translates these deliberative outcomes into actionable AI tools and solutions (AutoML).

The project will generate a variety of datasets, which will be primarily stored in ALFIE's AutoML Data Warehouse (AutoDW), the projects central persistence facility. The project will also produce several other research outcomes, such as AutoDW and ETD-Hub software. We are committed to making our data Findable, Accessible, Interoperable and Reusable (FAIR). We will adopt European Commission (EC) guidelines for data management and comply with the General Data Protection Regulation (GDPR). We are aiming to make the majority of ALFIE's datasets freely available during ALFIE and beyond the end of the project, in fully anonymised and publicly accessible formats.

This deliverable discusses the importance of Data Management Plans and FAIR principles. It introduces guidelines about data collection, generation, processing, handling, protection, sharing and preservation in ALFIE. It explains how the Data Management Plan Template for EU projects has been used to collect data management plan information for all ALFIE partners. In succession, the deliverable describes in detail all datasets and other research outcomes that ALFIE partners plan to develop, create, generate and/or use for the purposes of ALFIE. It presents information about how each of these items will comply to the FAIR principles, the GDPR and ALFIE's Grant Agreement (GA).

The DMP will be updated frequently through the duration of the project. A second and final version of this deliverable will be prepared and submitted in M34.

Contents

Executive Summary	4
1 Introduction	7
1.1 Overview	7
1.2 Relation to other tasks and deliverables	7
1.3 Structure of the deliverable	7
2 Data Management Plans.....	8
2.1 Overview	8
2.2 What is a Data Management Plan and why it is useful.....	8
2.3 Data Findability, Accessibility, Interoperability, and Reusability (FAIR).....	9
3 Guidelines	10
3.1 Overview	10
3.2 Data Handling	10
3.3 Data Collection, Generation, Processing	10
3.4 Data Protection	10
3.5 Data Sharing	11
3.6 Data Curation/Preservation.....	11
4 ALFIE Datasets and Other Research Outcomes	12
4.1 Overview	12
5 FAIR.....	17
5.1 Overview	17
5.2 Data and Metadata Findability	17
5.3 Data and Metadata availability.....	18
5.4 Data interoperability	20
5.5 Data Re-use.....	21
5.6 Ethics	23
6 Conclusions.....	25
References.....	26
Appendix A: Data Management Plan Template for EU Grants	27
Appendix B: Data Management Plan Questionnaire.....	32
Appendix C: Data Management Plan Questionnaire with Partner Responses	37



Tables

Table 1. Summary information for all ALFIE datasets and research outcomes (up to M5).... 12

Table 2. Size, origin and purpose of all ALFIE datasets and research outcomes (up to M5) 13

Table 3. FAIR data for all ALFIE datasets and research outcomes (up to M5)..... 17

Table 4. Data and Metadata availability for all ALFIE datasets and research outcomes (up to M5) 19

Table 5. Data interoperability for all ALFIE datasets and research outcomes (up to M5)..... 20

Table 6. Data re-use information for all ALFIE datasets and research outcomes (up to M5) 21

Table 7. Ethics concerns for all ALFIE datasets and research outcomes (up to M5) 23

1 Introduction

1.1 Overview

The scope of this deliverable is to describe in detail how data and other research outcomes will be managed during and after the ALFIE project. It concerns data or datasets that will be created or generated during the project and also pre-existing data or datasets that will be reused and explored during the project for some of its research and analysis purposes. The deliverable will also provide details about other research outcomes, such as software, persistence mechanisms and platforms, that will be designed, developed and employed during ALFIE.

The deliverable provides a detailed list of all data, datasets and other research outcomes that ALFIE partners currently plan to design, create, generate or reuse. For every item in the list, the deliverable presents information about its origin, its size, its purpose and its relation to ALFIE tasks and deliverables, the FAIR principles (Wilkinson et al., 2016), as well as additional ethics considerations. These data management dimensions follow the Data Management Plan for EU Grants template (Appendix A). Due to the large number of project partners and data, dataset and research output items, the template was converted into a Data Management Questionnaire (Appendix B), that was communicated to all ALFIE partners and has been filled in once for every item. Responses of all ALFIE partners to the Data Management Questionnaire about data, datasets and other research outcomes that relate to their contribution in ALFIE have been processed collectively to form the basis for the DMP presented in this deliverable.

This early version of the DMP assumes that there will be further additions, modifications, clarifications to be made. This DMP will be used in ALFIE as a “living-document” aiming to reflect how data and research outcomes are being managed at any point during the project. A more complete, final version of the DMP will be submitted in M34.

1.2 Relation to other tasks and deliverables

This deliverable is related to the majority of ALFIE tasks and deliverables, as data management is a cross-cutting concern. The DMP, in its current early form, is directly related to the implementation of the ETD-Hub and the design and development of Auto-DW, ALFIE's data warehouse, i.e., the facility that will store and make available datasets for training AI models using AutoML.

1.3 Structure of the deliverable

Section 2 introduces DMPs, discusses their importance and the FAIR principles. Section 3 discusses guidelines to be followed for data handling, collection, generation, processing, sharing curation and preservation. Sections 4 and 5 present the plans of ALFIE project partners for every data, dataset or other related research outcome. Section 4 reviews general characteristics for each item, Section 5 concerns FAIR aspects and ethics. Section 6 concludes this deliverable.

2 Data Management Plans

2.1 Overview

This deliverable presents the DMP for the ALFIE project, detailing how research data will be handled to ensure compliance with ethical, legal, and regulatory frameworks, particularly those set by the European Union (EU). The EU places a strong emphasis on data protection, privacy, and responsible data management, as outlined in the General Data Protection Regulation (GDPR) ([Regulation \(EU\) 2016/679](#)), which serves as a cornerstone for ensuring that personal and sensitive data are processed lawfully, fairly, and transparently.

The DMP outlines the following key aspects:

- **Data Collection & Processing:** Identifying the types of data generated, ensuring alignment with GDPR principles, and implementing necessary safeguards for data protection.
- **Storage & Security:** Establishing mechanisms to secure data integrity and confidentiality while adhering to EU guidelines on responsible data governance.
- **FAIR Principles:** Ensuring that data remains Findable, Accessible, Interoperable, and Reusable to maximise scientific impact while respecting privacy and security regulations.
- **Data Sharing & Preservation:** Addressing how research data will be shared, whether publicly or under controlled access, in compliance with GDPR and EU open science initiatives.

This DMP is a living document that will be updated throughout the project to reflect evolving best practices, regulatory requirements, and ethical considerations. By following the EU's strong commitment to data protection and responsible AI development, the ALFIE project ensures that its data management practices uphold the highest standards of transparency, security, and accountability.

2.2 What is a Data Management Plan and why it is useful

A DMP is a structured document that outlines how data is collected, stored, processed, shared, and preserved throughout a research project. It ensures that data handling practices align with legal, ethical, and scientific standards, promoting transparency, security, and reusability. The EU places a strong emphasis on responsible data governance, particularly through the GDPR, which mandates strict requirements for data protection, privacy, and security.

The usefulness of a DMP lies in its ability to:

- **Ensure Compliance:** By aligning with EU regulations such as GDPR, the DMP helps safeguard personal and sensitive data, ensuring lawful and ethical research practices.
- **Enhance Data Security:** A well-structured plan includes provisions for data encryption, access controls, and anonymisation, reducing the risk of unauthorized access and data breaches (European Data Protection Board, 2020).
- **Promote Open Science:** In line with the EU's Open Science policy, the DMP ensures that research outputs are FAIR, while respecting privacy constraints.

- **Facilitate Reproducibility and Long-term Preservation:** Good data management supports research integrity by making data reusable for future studies, increasing the impact and credibility of scientific work.
- **Support Ethical AI Development:** Given ALFIE's focus on intelligent and ethical AI, the DMP plays a crucial role in defining responsible data practices that align with EU values on human-centric AI and digital ethics.

The European Commission's Horizon Europe framework explicitly requires funded projects to develop a DMP as part of their commitment to Responsible Research and Innovation (RRI). By adhering to these guidelines, the ALFIE project ensures that its data management approach aligns with the EU's broader mission of safeguarding data privacy while fostering scientific progress.

2.3 Data Findability, Accessibility, Interoperability, and Reusability (FAIR)

The FAIR principles provide a structured framework for ensuring that research data is managed effectively and can be used beyond its initial collection. These principles, widely endorsed by the EU's Open Science policy, align with the GDPR by promoting responsible data handling while ensuring privacy and security.

- **Findability:** Findable data ensures that datasets are assigned unique and persistent identifiers, such as DOIs (Digital Object Identifiers) or URNs (Uniform Resource Names), making them discoverable in public repositories. Metadata should be richly described, indexed in widely used data registries, and follow standardised naming conventions.
- **Accessibility:** Accessibility ensures that data can be retrieved and used by authorised stakeholders. This requires clear data access policies, adherence to GDPR compliance for personal data protection, and sustainable long-term storage solutions. Where restrictions apply (e.g., due to ethical concerns), access conditions should be explicitly defined.
- **Interoperability:** Interoperability ensures that datasets can be integrated with other data sources and used across different platforms. To achieve this, data should be formatted using open, widely accepted standards (e.g., XML, JSON, RDF) and include standardised vocabularies to facilitate cross-disciplinary compatibility.
- **Reusability:** Reusability ensures that data can be reliably reused in future research. This requires well-documented datasets, clear licensing (e.g., Creative Commons, Open Data Commons), and provenance tracking to maintain data integrity. Ethical considerations, particularly GDPR compliance, must be embedded to protect sensitive data while ensuring usability.

By adhering to FAIR principles, the ALFIE project ensures that research outputs are maximised, ethically managed, and contribute to a wider ecosystem of transparent and responsible AI-driven education technologies.

3 Guidelines

3.1 Overview

The scope of this deliverable is to outline the ALFIE project's approach to data management, ensuring compliance with GDPR and ethical standards. It details procedures for data handling across its lifecycle, from collection and processing to storage, protection, and sharing. The project will collect, generate, and process relevant research data while implementing robust data protection measures, including encryption, access controls, and Data Protection Impact Assessments (DPIA) where required. Data sharing will be carefully managed, with only approved datasets made available to external stakeholders while maintaining privacy safeguards. These measures ensure secure and ethical data usage throughout and beyond the project's duration.

3.2 Data Handling

Research data in the ALFIE project will be managed securely throughout its lifecycle, ensuring compliance with GDPR. During the project delivery, data will be stored in partner institution's repositories and/or in ALFIE's Data Warehouse (AutoDW) with controlled access, employing encryption and pseudonymisation where necessary. After project completion, anonymised datasets will be archived in trusted repositories, while personal or sensitive data will be either deleted or de-identified in accordance with GDPR principles. Datasets related to the ALFIE's use cases, will be owned by the respective use case partners, who may choose to make them partially or wholly available for particular purposes, e.g. for research purposes. These versions will also be anonymised with personal or sensitive data either deleted or de-identified in accordance with GDPR principles.

3.3 Data Collection, Generation, Processing

The project will collect, generate, and process various datasets related to ethical AI-driven learning technologies, including user interaction data, survey responses (with informed consent), and educational content metadata. Specific to the ALFIE project, the EthiTech Dialogue Hub (ETD-Hub) will allow the public to join the conversation about the future of AI. This dynamic forum is intended to bring together diverse perspectives from citizens, policymakers, experts, and legal professionals to discuss the ethical and legal implications of Artificial Intelligence (AI). Likewise, AutoML which is our automated AI model evaluation platform will provide a streamlined way to assess AI performance, ensuring that it aligns with ethical standards and societal values. Processing methods such as data anonymisation, machine learning analysis, and NLP techniques will adhere to GDPR principles of minimisation, lawful processing, and purpose limitation, ensuring that only necessary and relevant data are utilised.

3.4 Data Protection

Compliance with GDPR and EU cybersecurity frameworks will be maintained through encryption, access controls, and regular security audits. Ethical considerations will be addressed through explicit consent and Data Protection Impact Assessments (DPIA), which evaluate the potential impact of data processing on individuals' rights. All ALFIE project partners will commit to conducting a DPIA before processing personal data, particularly when



there is a high risk to individuals' rights and freedoms. Additionally, institutional ethics board approvals will be obtained to ensure research integrity and data protection.

3.5 Data Sharing

Only anonymised research data, processed datasets, and trained AI models will be shared with external stakeholders such as academic collaborators, policymakers, and industry partners. Any data-sharing activities will follow strict legal and ethical guidelines, ensuring compliance with GDPR through formal agreements that define access, purpose, and usage rights.

3.6 Data Curation/Preservation

To ensure long-term accessibility and compliance with FAIR principles, ALFIE will store anonymised datasets in open repositories, such as Zenodo or OpenAIRE, maintaining structured metadata for discoverability. Version control mechanisms will be implemented to track modifications, preserving data integrity and aligning with EU Open Science policies.

4 ALFIE Datasets and Other Research Outcomes

4.1 Overview

This section summarises the main characteristics of the datasets and other research outcomes that ALFIE partners plan to develop or use. The content is based on the responses of all ALFIE partners to the data summary questions of the Data Management Plan for EU Grants (Appendix A).

Table 1 presents the ID, type and title of all ALFIE datasets and other research outcomes, as well as the partners that are mainly involved in their development, creation or exploration. The AutoML Data Warehouse (1-AutoDW) is a storage facility for all ALFIE datasets, while ETD-Hub software (14-ETDSOFT) is software that implements the ETH-hub platform. All other items concern datasets.

Table 1. Summary information for all ALFIE datasets and research outcomes (up to M5)

ID	Type	Title	Partner
1-AutoDW	Data storage facility	Auto Data Warehouse	CERTH
	Description: AutoDW will provide connectors to open access data repositories (e.g., Hugging Face) and will allow users to upload their own data.		
2-IRIS	Dataset	IRIS Drivers' behaviour profile dataset	CTL
	Description: Drowsiness and stress level information for IRIS application users. Data is collected from mobile phones' front cameras and heart-rate wearable sensors (e.g. smartwatches). Camera frames are annotated with 3 events: yawning, closed eyes, distracted or none of the above.		
3-WebAcc	Dataset	COVID-19 Web Accessibility Dataset	UAB
	Description: Evaluates the web accessibility compliance of 199 official COVID-19 information websites. These websites belong to government agencies, public health institutions, and international organizations that provided pandemic-related information. The analysis identifies accessibility barriers affecting users with disabilities, ensuring alignment with WCAG 2.1 standards.		
4-BestPrac	Dataset	Best Practices and Guidelines for Human-Centric and Trustworthy AI Development	DBC
	Description: Information gathered from the ETD-Hub for the development of best practices and guidelines in human-centric and trustworthy artificial intelligence (AI). It contains analyses, recommendations, and best practices based on ethical principles and EU values for AI development.		
5-BOSCH	Dataset	BOSCH Dataset	BOSCH
	Description: The dataset contains compliance alerts fetched from paid Internet Services. The task associated to this dataset is to automatically check and predict if these alerts comply with Bosch Compliance Regulations.		
6-AIEthics	Dataset	T2.2 AI Ethics Frameworks Evaluation Dataset	KInIT
	Description: Dataset containing structured analysis of AI ethics frameworks, including manual evaluations and automated annotations. The dataset maps frameworks against Trustworthy AI principles and moral sensitivity dimensions.		
7-AIoD	Dataset	AI on Demand (AIoD) metadata catalogue	KInIT
	Description: Dataset of metadata about different AI Assets aggregated from different sources (HuggingFace, Zenodo, OpenML, AIDA). It's available through a REST API:		

	https://api.aiod.eu/docs .		
8-T2.2Sur	Dataset	T2.2 Surveys & interviews responses	KInIT
	Description: A dataset of responses from stakeholders who participated as respondents in either our quantitative (survey) or qualitative (interview) research.		
9-T2.2Ann	Dataset	T2.2 AI Ethics Frameworks Evaluation Dataset	KInIT
	Description: Dataset containing structured analysis of AI ethics frameworks, including manual evaluations and automated annotations. The dataset maps frameworks against Trustworthy AI principles and moral sensitivity dimensions.		
10-ThreadSum	Dataset	Thread Summaries	DAS
	Description: AI summaries of threads, without user information.		
11-PostMeta	Dataset	Post Metadata (e.g. Upvotes/downvotes metadata)	DAS
	Description: Post metadata including upvote/downvote etc.		
12-PostText	Dataset	Post Text data	DAS
	Description: Contents of/summaries of posts on ETD-Hub.		
13-Diversity	Dataset	Stakeholder Diversity Data	DAS
	Description: Stakeholder Diversity Data will be collected during the onboarding process. Shortly after its completion, the data will be discarded.		
14-ETDSoft	Software	ETD-Hub software	DAS
	Description: ETD-Hub platform web software.		

IRIS Drivers' behaviour profile dataset (2-IRIS) and the BOSCH Dataset (5-BOSCH) are associated with ALFIE use cases. Proprietary rights of these datasets belong to the respective ALFIE Business partners. Consequently, although the datasets will follow the FAIR principles and GDPR, they will not be made freely available, unless the respective ALFIE Business partners decide so. For completeness, it should be noted that COVID-19 Web Accessibility Dataset (3-WebAcc) is also associated to an ALFIE use case. However, it is a pre-existing, freely available dataset.

Table 2 illustrates some more information about ALFIE datasets and research outcomes, in particular whether they are pre-existing or will be developed in ALFIE, their size, their origin and their purpose of development, creation or involvement in ALFIE.

Table 2. Size, origin and purpose of all ALFIE datasets and research outcomes (up to M5)

ID	Pre-existing	Size	Origin	Purpose
1-AutoDW	No	No specific limitation	To be developed by ALFIE partners	ALFIE's fundamental component
2-IRIS	Partially	637MB per 30min session	Privately collected and annotated sessions.	ALFIE's use case

3-WebAcc	Yes	717KB	Generated by the Universidad de Las Américas (Ecuador) and Universitat d'Alacant (Spain)	Research around a major ALFIE concern, web accessibility
4-BestPrac	Partially	TBC	To be collected from ETD-Hub participants	Links ETD-Hub to AutoML
5-BOSCH	Partially	30M entries, 30GB	Privately collected and annotated daily.	ALFIE's use case
6-AIEthics	No	TBC	To be generated by ALFIE partners	To systematically evaluate the effectiveness of existing AI ethics frameworks, identify gaps in current implementations, and develop evidence-based recommendations for improving ethical AI principles and their practical application.
7-AIoD	Yes	TBC	Available in HuggingFace, Zenodo, OpenML, AIDA, etc.	To be integrated in the AutoML platform. In the current phase of the project, it will be used to recommend existing AI models and datasets to the users based on their query.
8-T2.2Sur	No	TBC	To be collected by ALFIE partners	Necessary for task T2.2. Data will be used for enhancing methodology of the meta-analysis regarding the effectiveness of ethical frameworks and will be integrated into D2.2.
9-T2.2Ann	No	TBC	To be generated by ALFIE partners	Research around a major ALFIE concern, AI Ethics
10-ThreadSum	No	<1GB	To be generated by ALFIE partners	Links ETD-Hub to AutoML
11-PostMeta	No	<1GB	To be generated by ALFIE partners	Links ETD-Hub to AutoML
12-PostText	No	<1GB	To be generated by ALFIE partners	Links ETD-Hub to AutoML
13-Diversity	No	<100MB	To be generated by ALFIE partners	To ensure that stakeholders are representative enough.
14-ETDSOft	No	Not applicable	To be developed by ALFIE partners	ALFIE's fundamental component

All ALFIE datasets and research outcomes are strongly linked and necessary for carrying out tasks and functions of the ALFIE project.

AutoML Data Warehouse (1-AutoDW) provides a data hub that is necessary in ALFIE since it will provide trained AI models and ML solutions. While certain datasets can be found on the web, the need to upload and pre-process data within ALFIE is important for the objectives of the project.

IRIS Drivers' behaviour profile dataset (2-IRIS) will be a very large and diverse dataset about car drivers. It is the basis of the CTL use case and necessary for developing ethical models for populating driver profiles. There exist several alternative datasets, primarily focused on identifying distraction when driving, not only drowsiness. There are also some datasets for drowsiness, but the data are not sufficient for the purpose.

The COVID-19 Web Accessibility Dataset (3-WebAcc) promotes the use of AI for fair, ethical and inclusive digital environments. By analysing how websites perform in accessibility compliance, AI solutions can be developed to automatically detect and correct barriers (e.g., low contrast, missing alt text, poor keyboard navigation). There are several alternative datasets focused on web accessibility, each with different areas of emphasis. Some datasets primarily assess general accessibility compliance, while others focus on specific aspects such as colour contrast, screen reader compatibility, or keyboard navigation. The COVID-19 Web Accessibility Dataset has been selected as superior of all alternatives in terms of scope, insights and focus. It covers 199 official websites from multiple countries and international organizations. It allows for AI-based evaluation of accessibility issues, which aligns with ALFIE's focus on ethical, unbiased AI solutions. Unlike general accessibility datasets, it focuses on government and institutional websites, where accessibility compliance is critical.

The purpose of collecting and processing the Best Practices and Guidelines for Human-Centric and Trustworthy AI Development (4-BestPrac) data aligns with the goals of ALFIE, which focuses on developing ethical and trustworthy AI systems that uphold EU values and human rights. The best practices and guidelines developed will contribute to achieving these goals. Existing alternatives include data sources that represent initiatives and guidelines for ethical AI, e.g., AI Act of the EU. However, the ETD-Hub is preferable as it will be continuously updated and specifically designed to support the development of ethical and trustworthy AI, in line with the EU's values and the needs of the ALFIE project.

The BOSCH Dataset (5-BOSCH) supports ALFIE's objectives by providing a real-world use case for compliance monitoring and automated classification, which aligns with the project's goals of developing automated learning and intelligence systems.

The T2.2 AI Ethics Frameworks Evaluation Dataset (6-AIEthics) is useful for enhancing policies and suggesting best practices focused on promoting ethical AI.

The AI on Demand (AloD) metadata catalogue (7-AloD) will recommend existing models to ALFIE users. In broad terms, it contributes to the democratization of AI by making existing reusable AI assets easier to find.

The T2.2 Surveys & interviews responses (8-T2.2Sur) will assist in developing ethical, trustworthy and compliant AI models tailored on people needs, and will contribute to enhancing policies and suggesting best practices focused on promoting ethical AI.

Post Text data (12-PostText), Post Metadata (11-PostMeta) and Thread Summaries (10-ThreadSum) will be generated by the interaction of stakeholders in the ETD-Hub. The three datasets are crucial for the purposes of ALFIE, as link the ETD-Hub to the AutoML part of the project, i.e., they represent stakeholder opinion and will inform the development of responsible AI.

The Stakeholder Diversity Data (13-Diversity) will aid in confirming that stakeholders participating in ETD-Hub will be diverse enough and capable of representing all necessary



perspectives. Since the analysis of ETH-Hub data will be ALFIE's main source of information for developing responsible AI methods, democratising AI and enhancing related policies, making sure it is a well-rounded and objective source is of absolute necessity and importance.

Finally, ETD-Hub software (14-ETDSOFT) is instrumental in delivering ETD-Hub as a proof of concept.

It should be noted that Table 2 depicts several cases where ALFIE partners have not yet concluded/decided about a particular aspect of a dataset. These cases are marked as "to be confirmed (TBC)".

5 FAIR

5.1 Overview

This session contains FAIR information about the data, datasets and other research outcomes that will be generated, created, developed and used during ALFIE. FAIR aspects that are covered in the section are compatible with EC Data Management Plan template for EU grants (Appendix A) and have been compiled from the Data Management Questionnaires (Appendix B) that ALFIE partners have filled in (Appendix C). The scope of this section is to present this information briefly and collectively, and to facilitate a good overview of all data-related activities in ALFIE.

5.2 Data and Metadata Findability

Table 3 demonstrates how the 14 items in Table 1 are findable. Some of these aspects need to be discussed further and confirmed by ALFIE partners.

Table 3. FAIR data for all ALFIE datasets and research outcomes (up to M5)

ID	Persistent IDs	Metadata	Adoption of Metadata Standards	Keywords	Metadata for harvesting and indexing
1-AutoDW	No. Data will be accessible in ALFIE, only.	Currently: creators name, type of modalities present in the dataset, associated tasks (e.g. emotion recognition), licenses, number of instances	Yes	Yes	Yes (within ALFIE)
2-IRIS	Pre-existing	No	No	TBC	No
3-WebAcc	Yes	Descriptive, procedural, and technical details to enhance discoverability, reproducibility, and usability. Procedural Metadata (Steps to Reproduce), outlines the method for dataset generation: 1. Select the entire sample; 2. Access each website; 3. Test with the WAVE tool; 4. Record the data and analyse.		Yes. Keywords: "World Wide Web", "Accessibility Issue", "COVID-19"	Pre-existing
4-BestPrac	N/A	Dataset title, description, the ethical principles followed.	Yes, international metadata standards for AI datasets, e.g.,	N/A	Yes

			DCAT		
5-BOSCH	TBC	Standard metadata describing the compliance alert structure, classification categories, and data lineage would be appropriate.	Yes, standards for compliance	Yes, keywords for compliance monitoring, alert classification, regulatory compliance.	Yes (within ALFIE)
6-AIEthics	Yes. DOI through Zenodo.	Themes and components of moral sensitivity	There is no widely adopted standard specifically for AI ethics framework analysis	Yes	Yes
7-AIoD	Pre-existing	No	No	No	No
8-T2.2Sur	No	No	No	No	No
9-T2.2Ann	Yes. DOI through Zenodo.	Themes and components of moral sensitivity identified	There is no widely adopted standard specifically for AI ethics framework analysis	Yes	Yes
10-ThreadSum	Yes	Date, Timestamp, Theme, Themed ID, Post Title ID, Summary, Summary ID	General Best Practices	Theme	Yes
11-PostMeta	Yes	Row number, Date, Timestamp, Theme ID, Post Title ID, Comment ID, Username, Action (Upvote/Downvote)	General Best Practices	No	Yes
12-PostText	Yes	Row number, Date, Timestamp, Theme/Topic, Theme ID, Post Title, Post Title Id, Comment, Comment ID, Username, Attachments, Linked Post IDs	General Best Practices	Row number, Post Title, Theme, Comment, Username	Yes
13-Diversity	No	No	N/A	N/A	N/A
14-ETDSOft	N/A	N/A	N/A	N/A	N/A

5.3 Data and Metadata availability

This section presents data and metadata availability aspects for 14 ALFIE datasets and research outcomes in Table 1. Table 4 shows if and how each item and its associated metadata will be made available openly and a plan for their availability beyond the end of the project.

Table 4. Data and Metadata availability for all ALFIE datasets and research outcomes (up to M5)

ID	Data open availability	Data accessible through a free and standardized access protocol	Metadata openly available and licenced under public domain dedication (CC0)	Usage restrictions after ALFIE
1- AutoDW	No. This depends on the dataset	Yes	Yes	Most likely data will be removed. Specific custom datasets created for ALFIE use cases are owned by the use case partners.
2-IRIS	No, the dataset contains personal data.	Possibly. If we proceed with publishing them under research purposes, they will be open freely to any researcher.	Currently no. Might reconsider in the future.	If we proceed with publishing the dataset, we will set very strict restrictions for its usage (e.g. only for research purposes). The data will be available for 2 years after the end of the project.
3- WebAcc	Yes. Pre-existing	Yes. Pre-existing	Yes	Indefinitely (pre-existing dataset)
4- BestPrac	Yes	Yes, the data will be accessible via free and standardized access protocols such as APIs or repositories supporting open access data.	In clarification as explained in B.1 point.	If there are any usage restrictions, these will be specified in the dataset's terms of use and may relate to ensuring ethical and responsible use.
5- BOSCH	TBC. If yes, it should be anonymized.	If possible, yes, following ALFIE's recommendations.	Data is downloaded from an Internet paid service; terms of use to be reviewed.	N/A
6- AIEthics	Yes	Yes	Yes	Indefinitely
7-AIoD	Yes, through the AIoD platform	It already is through the AIoD platform.	No	No restrictions
8- T2.2Sur	The outputs of survey will be disseminated: anonymized partial data (aggregated or statistical data, specific quotes from questionnaires or interviews)	Yes, to the extent explained	No	GDPR restrictions

9-T2.2Ann	Yes	Yes	Yes	Indefinitely available
10-ThreadSum	Yes	Yes	Yes	No restrictions
11-PostMeta	No	No	Yes	12 Months
12-PostText	No, due to complex interactions with user content licence.	No, due to complex interactions with user content licence.	Yes	12 Months
13-Diversity	No. Personal, private, and no legal basis to do so	No, against GDPR	No	Never
14-ETDSoft	No, will be IP produced by DAS	No, against legitimate interests, against internal regulations	No	Platform may remain up but code never

Beyond the dimensions in Table 4, ALFIE partners have provided further details on the availability of datasets and other research outcomes, such as whether a data access committee should process requests for access to a dataset or research outcome before access is granted (Appendix C, Question D.6), if partners plan to make metadata available after the corresponding data has become unavailable (Appendix C, Question D.11) and if there are special software requirements to access or read the data (Appendix C, Question D.12).

5.4 Data interoperability

Interoperability refers to the compatibility of components to be combined and work together as a system. Datasets should also be interoperable, in the sense that they are accessible using an Application Programming Interface (API) or stored in a popular format, that is explored in code libraries in multiple programming languages. Table 5 reflects the main parts of data management questions about interoperability concerning all ALFIE datasets and research outcomes, while responses and more questions are available in Appendix C, section E.

Table 5. Data interoperability for all ALFIE datasets and research outcomes (up to M5)

ID	Processes to make data interoperable for data exchange/re-use within and across disciplines	Best practices to be followed
1-AutoDW	We will follow metadata standards such as the Dublin Core (DC), Data Documentation Initiative (DDI), or ISO	Open standards (e.g., JSON and XML for structured documents),

	19115. This depends on the use case data requirements.	metadata standards, API compatibility and the FAIR principles.
2-IRIS	Data will be stored in a Database or Data Warehouse (DW) in standard format and code to access them from there through appropriate APIs. This will follow the standard practices of saving data into a database to make it safe and compatible across disciplines.	TBC
3-WebAcc	Not applicable	
4-BestPrac	The dataset will follow common data formats. Metadata will follow the DCAT (Data Catalogue Vocabulary) and other relevant standards to ensure it can be reused across disciplines.	Yes, best practices endorsed by the open data community
5-BOSCH	The data is in CSV format, which is widely interoperable. Standard vocabularies for compliance classifications should be adopted.	Yes
6-AIEthics	N/A	
7-AIoD	It already is through the AIoD platform.	
8-T2.2Sur	N/A	
9-T2.2Ann	N/A	
10-ThreadSum	It inherently is – just plain text	No
11-PostMeta	It inherently is – just plain text	No
12-PostText	It inherently is – just plain text	No
13-Diversity	Data not to be shared or reused.	
14-ETDSOft	N/A - software output not a dataset	

5.5 Data Re-use

Table 6 summarises the main data management plan information that concern data re-use.

Table 6. Data re-use information for all ALFIE datasets and research outcomes (up to M5)

ID	Documentation types	Availability using standard licences	Availability after the end of ALFIE	Quality assurance processes
1-AutoDW	Readme files with installation and run instructions Online swagger page for the API, detailing endpoints.	Not applicable. AutoDW doesn't create any data, only associated metadata for every dataset	It depends on the owner of the data.	Functionalities for data validation and data cleaning, metadata documentation versioning capabilities and standardized data entry.

		saved.		
2-IRIS	Readme files and instructions will be shared within ALFIE.	Yes, using standard licences, not for free.	Yes, for 2 years after the end of the project.	Version control (e.g. Git) and automated quality control (i.e., to ensure that all the fields are filled in during data collection).
3-WebAcc	Yes (pre-existing)	Yes (pre-existing)	Yes (pre-existing)	N/A
4-BestPrac	A readme file, methodology description, codebooks, data cleaning procedures, and variable definitions	Yes. Free.	Yes, reusable by third parties, even after the project's completion.	Data validation and cleaning procedures to ensure the accuracy and consistency of the data, including checks for completeness, consistency, and adherence to predefined standards.
5-BOSCH	Documentation should include details about alert sources, classification criteria, and validation methods.	In clarification as explained in B.1 point.	In clarification as explained in B.1 point	Data is validated with few null values. The quality metric of particular importance is recall for the "Relevant" class, as it is important not to miss relevant cases that compliance officers need to investigate further.
6-AIEthics	Codebooks from manual analysis and prompts used for automated analysis.	Yes	Yes	Annotations will be cross-checked by multiple annotators.
7-AIoD	OpenAPI specification is available.	N/A, managed by the AIoD platform.	Yes, managed by the AIoD platform.	Managed by the AIoD platform.
8-T2.2Sur	No	No	No	N/A
9-T2.2Ann	Codebooks from manual analysis and prompts used for automated analysis.	Yes	Yes	Annotations will be cross-checked by multiple annotators.
10-ThreadSum	Readme on methods, format, field meanings and origin	Yes	Yes	Schema Conformity: Ensure that each record has the required fields (e.g., post ID, user ID, timestamp, content).
11-PostMeta		Yes	No	
12-PostText	Text data: Readme on methods format, field meanings, and origin	Yes	No	Metadata Checks: Validate timestamps, post-to-thread relationships. Logging and Version Control: Maintain logs of data processing steps and version control for changes to cleaning scripts or data schema. Anonymisation/De-identification: Remove or mask personally

				identifiable information (PII) and sensitive content. Policy Enforcement: Ensure compliance with relevant regulations (e.g., GDPR) during storage and processing. Regular Audits: Periodically sample posts to identify emerging issues like new spam patterns or unexpected data formats.
13-Diversity	None	No	No	N/A
14-ETDSOFT	N/A - software output not a dataset	N/A - software output not a dataset	N/A - software output not a dataset	N/A - software output not a dataset

Further than FAIR aspects, it is meaningful and important to investigate if making data and other research outputs FAIR would require any resources or financial costs. ALFIE partners have responded that there are only minimal costs involved that would be possible to cover through existing EC funding for ALFIE (Appendix C, G1, G2).

5.6 Ethics

The session contains ethics-related information about ALFIE datasets and other research outputs, as identified by ALFIE partners that are involved in creating, generating, processing and using them. Table 7 shows the most important of this information. ALFIE partners have identified ethics concerns that relate to personal sensitive information. Use case partners have reported more complex ethics issues, thus have not yet decided if the corresponding datasets will be made available and under which terms and conditions (see 2-IRIS and 5-BOSCH).

Table 7. Ethics concerns for all ALFIE datasets and research outcomes (up to M5)

ID	Issues that impact data sharing	Informed consent
1-AutoDW	No, ALFIE is a tool which prioritizes ethical AI development hence, the AutoDW cannot be utilized to create an unethical AI solution.	Yes, however only if those are present in the original data source.
2-IRIS	Yes, the data contains sensitive information that include physiological attributes of the participants (i.e. face, heart-rate).	Yes
3-WebAcc	None (pre-existing)	N/A
4-BestPrac	Potential ethical and legal issues could include privacy concerns if personal data is involved, or intellectual property considerations that may require a delay in	Yes, informed consent will be included in questionnaires to ensure that individuals are aware of how their data will be used and preserved in the long term.

	sharing certain datasets.	
5-BOSCH	Yes, the data contains information about entities (organizations, companies) and there are contractual considerations with the provider services that need clarification.	Not applicable as the data is about organizations rather than individuals.
6-AIEthics	We have to maintain existing licenses regarding the AI ethics frameworks.	N/A
7-AIoD	We're not aware of any.	We're not aware of such obligations.
8-T2.2Sur	Data will include personal information (GDPR applicable).	Yes
9-T2.2Ann	We have to maintain existing licenses regarding the AI ethics frameworks.	N/A
10-ThreadSum	Yes, including damaging legitimate interests – therefore data will not be shared.	N/A
11-PostMeta	Yes - may be linkable to other datasets to identify participants.	Yes
12-PostText	Yes – if stakeholders reveal excessive personal information/selfdox.	Yes
13-Diversity	Yes substantial – therefore data will not be shared.	N/A – data will not be shared or retained, but there will be a thorough disclosure and informed consent document anyway for use in the project.
14-ETDSOft	Yes, including damaging legitimate interests – therefore data will not be shared.	Yes, there will be a privacy policy and disclosures on the website

6 Conclusions

This deliverable presented the first, early version of ALFIE's Data Management Plan (DMP). It provided guidelines about collecting, generating, processing, handling, storing, sharing and preserving research data and other project outcomes during and after the end of the ALFIE project. ALFIE will adhere to all guidelines concerning the datasets it develops and uses. It will protect the privacy and rights of individuals according to the General Data Protection Regulation (GDPR). It will also adopt the principles of open science. The guidelines emphasised the importance of AutoDW, ALFIE's internal shared repository, for data security and sharing between ALFIE partners.

The deliverable also introduces a list of data, datasets and other research outcomes that will be generated, created, developed or used for the purposes of the project by diverse ALFIE partners. For each of these items, a wealth of information is presented, following the EC Data Management Plan template for EU Grants. The process of collecting this information is discussed in detail.

The DMP will be used as a living document in ALFIE. It will be updated frequently to reflect how partners are generating, creating, developing, using, storing, sharing research data and other research outcomes at any point in time. Due to the early nature of this deliverable, the DMP is not final and it is open to additions, amendments and clarifications based on how partners will manipulate data and other research outcomes in the duration of the project. The second version of this deliverable, due in M34, will present the ALFIE's final DMP.



References

- [1] European Data Protection Board. (2020, October). Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. Available at:
https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprtection_by_design_and_by_default_v2.0_en.pdf
- [2] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. <https://doi.org/10.1038/sdata.2016.18>

Appendix A: Data Management Plan Template for EU Grants

Overview

This appendix presents a copy of the Data Management Plan Template for EU Grants. Due to the large number of datasets and related research outputs, the template has not been adapted directly for the present deliverable. However, all areas of concern in the Data Management Plan Template for EU Grants, below, have been considered and converted into a Data Management Plan Questionnaire (Appendix B), that was filled in for each dataset and related research output by the project partner that leads the dataset generation process, employs an existing dataset or develops a specific research output. Appendix C presents a copy of the Data Management Plan Questionnaire that contains all combined raw responses of all partners for every dataset that will be generated as part of ALFIE or will be explored for the purposes and every related research output. Partner responses are the starting point for the content of this deliverable, as expressed in the main body of the report.

DATA MANAGEMENT PLAN

(To be filled in and uploaded as deliverable in the Portal Grant Management System, at the due date foreseen in the system (and regularly updated).)

 *The template is recommended but not mandatory. If you do not use it, please make however sure that you comply with the research data management requirements under Article 17 of the Grant Agreement.)*

PROJECT	
Project number:	[project number]
Project acronym:	[acronym]
Project name:	[project title]

DATA MANAGEMENT PLAN	
Date:	[dd/mm/yyyy]
Version:	[DMP version]

1. Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

What types and formats of data will the project generate or re-use?

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

What is the expected size of the data that you intend to generate or re-use?

What is the origin/provenance of the data, either generated or re-used?

To whom might your data be useful ('data utility'), outside your project?

2. FAIR data

2.1 Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

2.2 Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

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

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

2.3 Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

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? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references¹ to other data (e.g. other data from your project, or datasets from previous research)?

2.4 Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

¹ 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/>)



Will the data produced in the project be useable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

3. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

4. Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Who will be responsible for data management in your project?

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

5. Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Will the data be safely stored in trusted repositories for long term preservation and curation?

6. Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

7. Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

HISTORY OF CHANGES		
VERSION	PUBLICATION DATE	CHANGE
1.0	05.05.2021	Initial version (new MFF).
1.1	01.04.2022	Reformatted to align with other deliverables templates.

Appendix B: Data Management Plan Questionnaire

Overview

This Appendix presents a Data Management Plan Questionnaire that was prepared out of the Data Management Plan Template for EU Grants, in appendix A. The questionnaire sees the areas of concern independently for every dataset that will be generated or used in ALFIE and every related research product that will be produced. Each partner filled in a copy of the questionnaire for every dataset of research output that they will be involved with.

Data Management Plan Questions

Please fill in a copy of this document, for each dataset/data type that you plan to create/process/use². Please also fill in a copy for every other output you expect, digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

DATA SUMMARY

A.1. Dataset/Data Title:

A.2. Brief description of the dataset/data:

A.3. Is this dataset/data pre-existing?

A.4. What is the size of this dataset/data?

A.5. What is the origin/provenance of the data, either generated or re-used?

A.6. What is the purpose for collecting/processing/using this data:

A.7. How does the above purpose (A.4) relate to the objectives/aims of ALFIE?

A.8. Have you considered alternative datasets/data types for this purpose?

A.9. Why have you decided not to use the above alternatives?

² If you would like to provide the same answer for different datasets, Instead on copying over, please type "Same as in Data set [Data Title], for simplicity.



A.10. To whom may this dataset/data or the data processing outcome be useful, beyond ALFIE?

FAIR DATA

B.1. Is there/Do you plan to add a persistent identifier so that others can easily find the data/dataset?

B.2. What metadata/metadata types do you plan to create to allow discovery?

B.3. Do you follow any disciplinary or general metadata standards for this?

B.4. Will there be keywords in the metadata to optimise search and potential reuse?

B.5. Will you provide metadata in such a way that it can be harvested and indexed?

DATA ACCESSIBILITY

C.1. Will the data be deposited in a trusted repository?

C.2. Will the data be deposited in ALFIE's data repository?

DATA AND METADATA AVAILABILITY

D.1. Will this dataset/data be made openly available?

D.2. If it cannot be shared (or will be shared under restricted access conditions), please explain why.

D.3. Please, mention if the reasons are legal and contractual reasons.

Please, mention if sharing this data is against your legitimate interests or other constraints as per the Grant Agreement.

If you would like to apply an embargo so that you have time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will be.

D.4. Will the data be accessible through a free and standardized access protocol?

D.5. If there are usage restrictions, during or after the end of the project?



- D.6. Would you like access requests to be processed by a data access committee before access is granted?
- D.7. Do you intend to make the metadata openly available and licenced under a public domain dedication (CC0), as per the Grant Agreement?
- D.8. If not (D.7.), please clarify why.
- D.9. Will the metadata contain information to enable the user to access the data?
- D.10. For how long are you happy for your data to remain available and findable after the end of the project?
- D.11. Can metadata remain available after the data is no longer available?
- D.12. Is there any software needed to access or read the data?
- D.13. If yes (D.12.), will you include documentation/reference to that software?
- D.14. If yes (D.12.), would you include the software in your data release?

DATA INTEROPERABILITY

- E.1. How will you make your dataset/data interoperable to allow data exchange and re-use within and across disciplines?
(Will you follow any vocabularies, standards, formats or methodologies for your data/metadata?)
- E.2. Will you follow community-endorsed interoperability best practices? Which ones?
- E.3. In case it is unavoidable that you use bespoke project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
- E.4. Will your data include links/qualified references to other data (in ALFIE or external)?



DATA RE-USE

- F.1. What types of documentation will you provide to facilitate data re-use and data analysis validation?
(e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?
- F.2. Will your data be made freely available in the public domain?
- F.3. Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?
- F.4. Will the data produced in the project be useable by third parties, in particular after the end of the project?
- F.5. Will the provenance of the data be thoroughly documented using the appropriate standards?
- F.6. Briefly describe relevant data quality assurance processes.

RESOURCES ALLOCATION

- G.1. What is the costs for making data or other research outputs FAIR? (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?
- G.2. How will they be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

ETHICS

- H.1. Are there, or could there be, any ethics or legal issues that can impact data sharing?
- H.2. Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?



OTHER

- I.1. Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management?
- I.2. If yes, which ones (please list and briefly describe them)?

Appendix C: Data Management Plan Questionnaire with Partner Responses

Overview

This Appendix presents a copy of the Data Management Plan Questionnaire that contains all combined raw responses of all partners for every dataset that will be generated as part of ALFIE or will be explored for the purposes and every related research output. Partner responses are the starting point for the content of this deliverable, as expressed in the main body of the report.

Data Management Plan Questions

Please fill in a copy of this document, for each dataset/data type that you plan to create/process/use³. Please also fill in a copy for every other output you expect, digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

DATA SUMMARY

A.1. Dataset/Data Title:

CERTH_AutoDW_Task_3.1: Not limited to one dataset. The AutoDW will provide connectors to open access data repositories (such as Hugging Face) and will allow users to upload their own data. Hence multiple datasets will be present.

CTL: IRIS Drivers' dataset behaviour profile

UAB: Dataset web accessibility COVID-19 website

Acosta-Vargas, Patricia; Salvador-Acosta, Belén; Calvopina, Manuel; Kyriakidis, Nikolaos; Ortiz-Prado, Esteban; Salvador-Ullauri, Luis (2022), "(Dataset) Web accessibility COVID-19 website", Mendeley Data, V1, doi: 10.17632/9pkxzpcsm7.1

DBC: Best Practices and Guidelines for Human-Centric and Trustworthy AI Development

BOSCH: BOSCH Dataset

KInIT_AIEF: T2.2 AI Ethics Frameworks Evaluation Dataset

³ If you would like to provide the same answer for different datasets, Instead on copying over, please type "Same as in Data set [Data Title], for simplicity.

KInIT_AIoD: AI on Demand (AIoD) metadata catalogue

KInIT_T2.2_Survey: T2.2 Surveys & interviews responses

KInIT_T2.2_Annotation: T2.2 AI Ethics Frameworks Evaluation Dataset

DAS_Thread_Summaries: Thread Summaries

DAS_Post Metadata: Post Metadata (e.g. Upvotes/downvotes metadata)

DAS_Post Text data: Post Text data

DAS_Stakeholder_Diversity: Stakeholder Diversity Data

DAS_ETD-Hub_Software: ETD-Hub software

A.2. Brief description of the dataset/data:

CERTH_AutoDW_Task_3.1: Not applicable for AutoDW since multiple datasets of different purposes can exist.

CTL: The dataset contains information about the drowsiness and stress level of IRIS application users. Data is collected from a mobile phone's front camera and a heart-rate wearable sensor (e.g. smartwatch). Each frame of the video is annotated with either of 3 events: yawning, closed eyes, distracted or none of the above.

UAB: This dataset evaluates the web accessibility compliance of 199 official COVID-19 information websites. These websites belong to government agencies, public health institutions, and international organizations that provided pandemic-related information. The analysis identifies accessibility barriers affecting users with disabilities, ensuring alignment with WCAG 2.1 standards.

DBC: This dataset includes data and information gathered from the ETD-Hub for the development of best practices and guidelines in human-centric and trustworthy artificial intelligence (AI). It contains recommendations, analyses, and best practices based on ethical principles and EU values for AI development.

BOSCH: The dataset contains compliance alerts fetched from paid Internet Services. These alerts must be checked to determine if they are relevant for Bosch Compliance Regulations.

KInIT_AIEF: Dataset containing structured analysis of AI ethics frameworks, including manual evaluations and automated annotations. The dataset maps frameworks against Trustworthy AI principles and moral sensitivity dimensions

KInIT_AIoD: Dataset of metadata about different AI Assets aggregated from different sources (HuggingFace, Zenodo, OpenML, AIDA). It's available through a REST API: <https://api.aiod.eu/docs>.



KInIT_T2.2_Survey: A dataset of responses from stakeholders who participated as respondents in either our quantitative (survey) or qualitative (interview) research.

KInIT_T2.2_Annotation: Dataset containing structured analysis of AI ethics frameworks, including manual evaluations and automated annotations. The dataset maps frameworks against Trustworthy AI principles and moral sensitivity dimensions

DAS_Thread_Summaries: AI summaries of threads, without user information

DAS_Post Metadata: Post metadata including upvote/downvote etc.

DAS_Post Text data: Contents of/summaries of posts on ETD-Hub

DAS_Stakeholder_Diversity: Stakeholder Diversity Data used to inform onboarding then destroyed

DAS_ETD-Hub_Software: ETD-Hub platform web software

A.3. Is this dataset/data pre-existing?

CERTH_AutoDW_Task_3.1: Yes, some of the dataset are pre-existing.

CTL: We have collected some samples in the past, but we are going to enhance it during the ALFIE's project lifetime.

UAB: Yes. It is an external dataset.

DBC: Yes, the data is primarily pre-existing and collected from the ETD-Hub, which hosts information and resources for ethical and trustworthy AI development.

BOSCH: Yes, there is existing data from 2023 and 2024, with new data being collected daily.

KInIT_AIEF: No. While the ethics frameworks being analyzed exist, the evaluation data and annotations will be newly created through this research.

KInIT_AIoD: Yes.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: No. While the ethics frameworks being analyzed exist, the evaluation data and annotations will be newly created through this research.

DAS_Thread_Summaries: No.

DAS_Post Metadata: No.

DAS_Post Text data: No

DAS_Stakeholder_Diversity: No

DAS_ETD-Hub_Software: No

A.4. What is the size of this dataset/data?

CERTH_AutoDW_Task_3.1: It depends on each dataset, at the moment there is no specific limitation.

CTL: Currently we are storing every frame so for a frame rate of 30fps, a 30-minute drive results to 54k frames. That accounts for almost 637 megabytes for each 30 minute session.

UAB: Size 717 KB

DBC: The size of the dataset depends on the number of documents and data collected through the ETD-Hub and contributions from stakeholders. The size may grow as more information and feedback are collected.

BOSCH: Currently 28k unique alerts. Each alert has 1-N classifications by different Compliance officers. In total, the dataset has 7 million entries, stored in a CSV file of approximately 30 GB.

KInIT_AIEF: Dozens of manually annotated AI ethics frameworks.

Hundreds of automatically annotated AI ethics frameworks.

KInIT_AIoD: Hundreds of thousands of records.

KInIT_T2.2_Survey: Dozens of records.

KInIT_T2.2_Annotation: Dozens of manually annotated AI ethics frameworks.

Hundreds of automatically annotated AI ethics frameworks.

DAS_Thread_Summaries: <1 million rows, 30 columns, <1GB

DAS_Post Metadata: <1 million rows, 30 columns, <1GB

DAS_Post Text data: <1 million rows, 30 columns, <1GB

DAS_Stakeholder_Diversity: <100k rows, 30 columns, <100MB

DAS_ETD-Hub_Software: TBC – not yet built

A.5. What is the origin/provenance of the data, either generated or re-used?

CERTH_AutoDW_Task_3.1: Mostly re-used but could be original in case a user decides to upload his/her specific data for the associated use case.

CTL: The dataset is owned by Catalink Limited. The data will be generated by capturing data from participants.

UAB: The dataset was created by Universidad de Las Américas (Ecuador) and Universitat d'Alacant (Spain). It focuses on the accessibility analysis of 199 official COVID-19 information websites from multiple countries and regions from various health ministries and government agencies worldwide.

DBC: The data originates from various organizations, research institutions, policymakers, and stakeholders participating in the ETD-Hub.

BOSCH: Alert text is external but downloaded to Bosch's system. Labelling is done by internal employees.

KInIT_AIEF: - manual expert annotation of selected AI ethics frameworks

- automated annotation using either partner developed tools of publicly available ones

KInIT_AIoD: Dataset of metadata about different AI Assets aggregated from different sources (HuggingFace, Zenodo, OpenML, AIDA). It's available through a REST API: <https://api.aiod.eu/docs>.

KInIT_T2.2_Survey: Data collected from surveys sent to relevant stakeholders (AI companies, AI experts) and recordings (transcripts) of interviews conducted with selected participants.

KInIT_T2.2_Annotation: - manual expert annotation of selected AI ethics frameworks

- automated annotation using either partner developed tools of publicly available ones

DAS_Thread_Summaries: Generated

DAS_Post Metadata: Generated

DAS_Post Text data: Generated

DAS_Stakeholder_Diversity: Generated

DAS_ETD-Hub_Software: Generated

A.6. What is the purpose for collecting/processing/using this data:

CERTH_AutoDW_Task_3.1: The AutoDW is the main data hub of ALFIE and is in charge of providing storage to save data and make them searchable from the AutoML and then use them for training of AI and ML solutions. Pre-processing may

happen in order to normalize values and clean data in case of missing values, or severely underbalanced class distribution.

CTL: To train our AI models for IRIS application. One is about drowsiness detection and the other about stress level estimation.

UAB: The dataset aims to evaluate website accessibility for people with disabilities by detecting and addressing issues such as missing alternative text, improper HTML structure, insufficient colour contrast, poor navigation compatibility. The goal is to ensure compliance with WCAG 2.1 standards, making websites more usable for individuals with various disabilities.

DBC: The purpose of collecting and processing this data is to create a set of best practices and guidelines for the continued development of human-centric and trustworthy AI. The data will help align practices with ethical values and human rights and adapt them to evolving requirements.

BOSCH: The purpose is to identify and classify compliance alerts that are relevant to Bosch Compliance Regulations, with the goal of potentially automating the initial screening process.

KInIT_AIEF: To systematically evaluate the effectiveness of existing AI ethics frameworks, identify gaps in current implementations, and develop evidence-based recommendations for improving ethical AI principles and their practical application.

KInIT_AIoD: To be integrated in the AutoML platform. In the current phase of the project, the data will be used to recommend existing AI models and datasets to the users based on their query.

KInIT_T2.2_Survey: Data will be used for completing the deliverable of task T2.2. Data will be used for enhancing methodology of the meta-analysis regarding the effectiveness of ethical frameworks as well as integrated into the final report D2.2.

KInIT_T2.2_Annotation: To systematically evaluate the effectiveness of existing AI ethics frameworks, identify gaps in current implementations, and develop evidence-based recommendations for improving ethical AI principles and their practical application.

DAS_Thread_Summaries: To summarise perceptions on a given debate in AI

DAS_Post Metadata: To capture perceptions on debates in AI

DAS_Post Text data: To capture perceptions on debates in AI

DAS_Stakeholder_Diversity: To ensure representativity of stakeholders

DAS_ETD-Hub_Software: To host discussions on perceptions on debates in AI

A.7. How does the above purpose (A.4) relate to the objectives/aims of ALFIE?

CERTH_AutoDW_Task_3.1: ALFIE needs a data hub since it is going to provide trained AI models and ML solutions. While certain datasets can be found on the web, the need to upload and pre-process data within ALFIE is important for the objectives of the project.

CTL: We want to make ethical models for creating a driver profile. For this we need a very diverse dataset which is hard to construct or requires a large amount of data. We also need a dataset that is in our desired setting e.g. driving a car.

UAB: The dataset promotes the use of AI for fair, ethical and inclusive digital environments. By analysing how websites perform in accessibility compliance, AI solutions can be developed to automatically detect and correct barriers (e.g., low contrast, missing alt text, poor keyboard navigation).

DBC: The purpose of collecting and processing this data aligns with the goals of the ALFIE project, which focuses on developing ethical and trustworthy AI systems that uphold EU values and human rights. The best practices and guidelines developed will contribute to achieving these goals.

BOSCH: This dataset supports ALFIE's objectives by providing a real-world use case for compliance monitoring and automated classification, which aligns with the project's goals of developing automated learning and intelligence systems.

KInIT_AIEF: O3: To enhance policies and suggest best practices focused on promoting ethical AI

O4: To develop ethical, trustworthy and compliant AI models tailored on people needs

KInIT_AIoD: O4-Existing models will be recommended to the users.

O5-Contributes to democratization of AI by making existing reusable AI Assets to be easier to find.

KInIT_T2.2_Survey: O3: To enhance policies and suggest best practices focused on promoting ethical AI

O4: To develop ethical, trustworthy and compliant AI models tailored on people needs

KInIT_T2.2_Annotation: O3: To enhance policies and suggest best practices focused on promoting ethical AI

O4: To develop ethical, trustworthy and compliant AI models tailored on people needs

DAS_Thread_Summaries: To represent stakeholder opinion to inform development of responsible AI

DAS_Post Metadata: To represent stakeholder opinion to inform development of responsible AI

DAS_Post Text data: To represent stakeholder opinion to inform development of responsible AI

DAS_Stakeholder_Diversity: To represent diverse perspectives

DAS_ETD-Hub_Software: To deliver ETD-Hub proof of concept

A.8. Have you considered alternative datasets/data types for this purpose?

CERTH_AutoDW_Task_3.1: Not applicable since AutoDW can save multiple data even for the same purpose.

CTL: We have identified different datasets that are primarily focused on identifying distraction when driving, not only drowsiness. There are also some datasets for drowsiness but the data are not very sufficient.

UAB: Yes, there are several existing datasets focused on web accessibility, each with different areas of emphasis. Some datasets primarily assess general accessibility compliance, while others focus on specific aspects such as colour contrast, screen reader compatibility, or keyboard navigation.

DBC: Yes, there are other data sources, such as national and international initiatives that provide guidelines for ethical AI, like the AI Act of the EU or other research platforms.

BOSCH: No, this is an ad-hoc dataset.

KInIT_AIEF: No.

KInIT_AIoD: Not yet.

KInIT_T2.2_Survey: Not yet.

KInIT_T2.2_Annotation: No.

DAS_Thread_Summaries: No

DAS_Post Metadata: No

DAS_Post Text data: No

DAS_Stakeholder_Diversity: No

DAS_ETD-Hub_Software: No

A.9. Why have you decided not to use the above alternatives?

CERTH_AutoDW_Task_3.1: Not applicable.

CTL: Most of the big datasets we found are focused on identifying distraction when driving e.g driver texting. We want to focus on drowsiness. Also these datasets have different camera angles which might not be beneficial for our case. Finally we couldn't locate a dataset that combines camera feed and physiological indicators such as heart rate.

UAB: This dataset was selected because it is the most complete in terms of:

- Scope: it covers 199 official websites from multiple countries and international organizations.
- AI-driven insights: it allows for AI-based evaluation of accessibility issues, which aligns with ALFIE's focus on ethical, unbiased AI solutions.
- Public-sector and institutional focus: unlike general accessibility datasets, this one focuses on government and institutional websites, where accessibility compliance is critical

DBC: We considered alternatives but chose the ETD-Hub because it provides a continuously updated dataset and is specifically designed to support the development of ethical and trustworthy AI, in line with the EU's values and the needs of the ALFIE project.

BOSCH: Not applicable as no alternatives were mentioned.

KInIT_AIEF: N/A

KInIT_AIoD: AIoD metadata catalogue provides a convenient and efficient way how to access information about AI assets from different sources through one endpoint/place.

KInIT_T2.2_Survey: We are not aware of any publicly available alternatives to this dataset.

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: No other dataset would be relevant

DAS_Post Metadata: No other dataset would be relevant

DAS_Post Text data: No other dataset would be relevant



DAS_Stakeholder_Diversity: No other dataset would be relevant

DAS_ETD-Hub_Software: No other platform would be relevant – ETD-Hub specified in plan

A.10. To whom may this dataset/data or the data processing outcome be useful, beyond ALFIE?

CERTH_AutoDW_Task_3.1: Not applicable. The AutoDW is ALFIE's main data hub and storage system and doesn't own the data. Datasets are licensed and usage across multiple different users within ALFIE is determined by the licenses.

CTL: ADAS companies that develop smart applications for autonomous driving. Car manufacturers that would like to install ADAS technologies on their vehicles, such as the drowsiness detection and stress level estimation. It could possibly be added to public transportation and on track drivers to monitor their state for safer driving.

UAB: Beyond ALFIE, this dataset could be useful for:

- Web developers and designers: to enhance accessibility features on websites by identifying common WCAG 2.1 compliance issues.
- AI and machine learning researchers: to support the development of AI models for automated accessibility testing.
- Government and policy officials: to assess and improve digital accessibility standards for public services.
- Nonprofit and advocacy groups: to promote awareness and best practices for inclusive digital environments.
- Technology companies and UX specialists: to refine AI-based accessibility tools and ensure more user-friendly digital experiences.

DBC: This dataset and the outcomes of the data processing will be useful to researchers, AI development organizations, policymakers, regulatory bodies, and any stakeholders interested in ensuring that AI systems developed are responsible and compliant with ethical values and human rights.

BOSCH: The data processing outcome could be useful to compliance departments in other organizations that need to monitor and classify compliance alerts from external sources. This is, other organizations facing a similar problem.

KInIT_AIEF: AI organisations, AI practitioners and experts, regulatory bodies, AI/tech ethicists, AI auditors

KInIT_AIoD: AI researchers and practitioners, innovators.



KInIT_T2.2_Survey: AI organisations, AI practitioners and experts, regulatory bodies, AI/tech ethicists, AI auditors

KInIT_T2.2_Annotation: AI organisations, AI practitioners and experts, regulatory bodies, AI/tech ethicists, AI auditors

DAS_Thread_Summaries: Policymakers seeking perceptions of AI, developers seeking perceptions of AI

DAS_Post Metadata: Policymakers seeking perceptions of AI, developers seeking perceptions of AI

DAS_Post Text data: Policymakers seeking perceptions of AI, developers seeking perceptions of AI

DAS_Stakeholder_Diversity: Not to be used outside of specific purpose in ALFIE

DAS_ETD-Hub_Software: Policymakers seeking perceptions of AI, developers seeking perceptions of AI

FAIR DATA

B.6. Is there/Do you plan to add a persistent identifier so that others can easily find the data/dataset?

CERTH_AutoDW_Task_3.1: No. Data saved in AutoDW will not be accessed outside of ALFIE.

CTL: No

UAB: The dataset has already been assigned a DOI: 10.17632/9pkxzpcsm7.1

DBC: Yes, a persistent identifier will be added to allow users to easily locate the dataset.

BOSCH: In clarification as alerts are downloaded from an Internet paid service. Contract has to be reviewed about if the data can be exposed (even anonymized)

KInIT_AIEF: Yes. DOI through Zenodo.

KInIT_AIoD: It's already there.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: Yes. DOI through Zenodo.



DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: N/A - software output not a dataset

B.7. What metadata/metadata types do you plan to create to allow discovery?

CERTH_AutoDW_Task_3.1: At the moment metadata contain: the creators name, the type of modalities present in the dataset, tasks associated (e.g. emotion recognition), licenses, size in terms of instances. We may add more in the future depending on the requirements (e.g. we may need to adhere to specific metadata standard).

CTL: None for the moment.

UAB: The existing database provides structured guidelines, ensuring that the metadata for this dataset includes descriptive, procedural, and technical details to enhance discoverability, reproducibility, and usability. Procedural Metadata (Steps to Reproduce), outlines the methodology used to generate the dataset:

1. Select the entire sample;
2. Access each website;
3. Test with the WAVE tool;
4. Record the data and analyse.

DBC: Metadata will be created that include the dataset title, description, the ethical principles followed.

BOSCH: Standard metadata describing the compliance alert structure, classification categories, and data lineage would be appropriate.

KInIT_AIEF: themes and components of moral sensitivity identified

KInIT_AIoD: There's no plan to add metadata yet.

KInIT_T2.2_Survey: There's no plan to add metadata.

KInIT_T2.2_Annotation: themes and components of moral sensitivity identified

DAS_Thread_Summaries: Date, Timestamp, Theme, Themed ID, Post Title ID, Summary, Summary ID

DAS_Post Metadata: Rownumber, Date, Timestamp, Theme ID, Post Title ID, Comment ID, Username, Action (Upvote/Downvote)



DAS_Post Text data: Rownumber, Date, Timestamp, Theme (theme/topics under which the post resides e.g. image segmentation, fair AI, AI ethnicity bias), Theme ID, Post Title, Post Title Id, Comment, Comment ID, Username, Attachments, Linked To (connected post IDs)

DAS_Stakeholder_Diversity: None

DAS_ETD-Hub_Software: None - software output not a dataset

B.8. Do you follow any disciplinary or general metadata standards for this?

CERTH_AutoDW_Task_3.1: Yes

CTL: No

UAB: It is an existing database; however, it aligns with the FAIR data principles, ensuring that it is Findable, Accessible, Interoperable, and Reusable.

DBC: Yes, the metadata will comply with international metadata standards for AI datasets, such as the DCAT (Data Catalogue Vocabulary) and other relevant guidelines in line with the project's needs.

BOSCH: No specific metadata standards are used, but standards for compliance data should be applied.

KInIT_AIEF: We acknowledge there isn't a widely-adopted standard specifically for AI ethics framework analysis, but we build on existing meta-analyses.

KInIT_AIoD: No.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: We acknowledge there isn't a widely-adopted standard specifically for AI ethics framework analysis, but we build on existing meta-analyses.

DAS_Thread_Summaries: General Best Practices

DAS_Post Metadata: General Best Practices

DAS_Post Text data: General Best Practices

DAS_Stakeholder_Diversity: N/A - No to B.2.

DAS_ETD-Hub_Software: N/A - software output not a dataset

B.9. Will there be keywords in the metadata to optimise search and potential reuse?

CERTH_AutoDW_Task_3.1: Yes

CTL: Have not considered it yet.

UAB: The current keywords include World Wide Web, Accessibility Issue, COVID-19

DBC: Yes, keywords will be used to improve search and potential reuse of the data, such as "ethical AI," "European values," "responsible AI development," etc.

BOSCH: Yes, keywords related to compliance monitoring, alert classification, and regulatory compliance would be appropriate to include.

KInIT_AIEF: Yes.

KInIT_AIoD: No.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Theme

DAS_Post Metadata: No.

DAS_Post Text data: Rownumber, Post Title, Theme , Comment, Username

DAS_Stakeholder_Diversity: N/A - No to B.2.

DAS_ETD-Hub_Software: N/A - software output not a dataset

B.10. Will you provide metadata in such a way that it can be harvested and indexed?

CERTH_AutoDW_Task_3.1: Yes

CTL: No

UAB: As anticipated, the dataset and its metadata are already available. Additionally, there is a detailed article that explains how the dataset was established, ensuring transparency and accessibility for further research and indexing. This article can be accessed here: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9566462/>

DBC: Yes, the metadata will be made available in such a way that it can be harvested and indexed by other platforms, enhancing the discoverability and recognition of the data.

BOSCH: Yes, metadata will be structured to facilitate harvesting and indexing within the ALFIE framework.

KInIT_AIEF: Yes.



KInIT_AIoD: No, the data is provided by a 3rd party (AI on Demand platform).

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes

DAS_Stakeholder_Diversity: N/A - No to B.2.

DAS_ETD-Hub_Software: N/A - software output not a dataset

DATA ACCESSIBILITY

C.3. Will the data be deposited in a trusted repository?

CERTH_AutoDW_Task_3.1: Yes. MongoDB is the backend software for building the AutoDW and storing data on it.

CTL: Initially in CTL server. Then they will be deposited in ALFIE's data warehouse.

UAB: Yes.

DBC: Yes

BOSCH: In clarification as explained in B.1 point. If yes, it should be anonymized.

KInIT_AIEF: Yes.

KInIT_AIoD: Yes (if AIoD platform is considered as trusted – I'm not sure about what that means).

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: No (Is in ALFIE repo)

DAS_Post Metadata: No (Is in ALFIE repo)

DAS_Post Text data: No (Is in ALFIE repo)

DAS_Stakeholder_Diversity: Yes briefly, then destroyed

DAS_ETD-Hub_Software: Yes.



C.4. Will the data be deposited in ALFIE's data repository?

CERTH_AutoDW_Task_3.1: Yes. Data will be stored on a dedicated server on CERTH premises.

CTL: Yes, under specific conditions.

UAB: Yes.

DBC: Yes.

BOSCH: In clarification as explained in B.1 point. If yes, it should be anonymized.

KInIT_AIEF: Yes.

KInIT_AIoD: No.

KInIT_T2.2_Survey: Probably yes

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: No.

DATA AND METADATA AVAILABILITY

D.15. Will this dataset/data be made openly available?

CERTH_AutoDW_Task_3.1: No. This also depends on the owner of the data (check also B1).

CTL: Not foreseen.

UAB: This dataset is already existing and publicly available. The same response applies to the remaining questions.

DBC: Yes, the dataset will be made openly available to the public, adhering to the principles of transparency and open data to foster further research and collaboration.

BOSCH: In clarification as explained in B.1 point. If yes, it should be anonymized.

KInIT_AIEF: Yes.

KInIT_AIoD: It already is through the AIoD platform.

KInIT_T2.2_Survey: No. Only the outputs of survey in form of papers, reports or policy recommendations

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: No.

DAS_Post Text data: No.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: No.

D.16. If it cannot be shared (or will be shared under restricted access conditions), please explain why.

CERTH_AutoDW_Task_3.1: The AutoDW is part of ALFIE's pipelines and can be accessed only by ALFIE users. Hence, access levels of its data depend on the restrictions imposed by the owner of the data (check also B1).

CTL: The dataset contains personal data (e.g. heart rate, face photographs) that might harm the participants' privacy.

DBC: In case of any data sharing restrictions, they would typically relate to legal, ethical, or privacy concerns, especially if personal data or sensitive information is involved. However, we aim to make the dataset openly available, with possible access restrictions depending on specific terms or requirements.

BOSCH: In clarification as explained in B.1 point.

KInIT_AIEF: N/A

KInIT_AIoD: N/A

KInIT_T2.2_Survey: It will include some personal data of respondents that we don't need to share (there is no purpose for that). However, anonymized partial data (aggregated or statistical data, specific quotes from questionnaires or interviews) will be shared as a part of D2.2 and other publicly available outputs (papers, reports, policy recommendations).

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: N/A - Yes to D.1.

DAS_Post Metadata: May be linkable to other data to allow identification of individuals.

DAS_Post Text data: Complex interactions with any user content licence meaning sharing the raw data is not feasible. May be linkable to other data to allow identification of individuals.

DAS_Stakeholder_Diversity: Personal, private, and no legal basis to do so

DAS_ETD-Hub_Software: Software will be IP produced by DAS

D.17. Please, mention if the reasons are legal and contractual reasons.

Please, mention if sharing this data is against your legitimate interests or other constraints as per the Grant Agreement.

If you would like to apply an embargo so that you have time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will be.

CERTH_AutoDW_Task_3.1: Not applicable.

CTL: IRIS dataset contains participants' faces and other personal attributes that the users might not want to share with others.

DBC: The data sharing will not be restricted unless there are intellectual property concerns, in which case an embargo may be applied for a limited period to secure IP rights, such as patents. This would be done to ensure the protection of proprietary information before making it publicly available.

BOSCH: Data is downloaded from an Internet paid service; terms of use have to be reviewed.

KInIT_AIEF: N/A

KInIT_AIoD: N/A

KInIT_T2.2_Survey: Explained above. There are no other reasons to not share all data we will collect.

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: N/A - Yes to D.1.

DAS_Post Metadata: Legal, may be linkable to other data to allow identification of individuals.

DAS_Post Text data: Legal, may be linkable to other data to allow identification of individuals.

DAS_Stakeholder_Diversity: Legal, against GDPR to do so

DAS_ETD-Hub_Software: Legal & Contractual - against legitimate interests, against internal regulations

D.18. Will the data be accessible through a free and standardized access protocol?

CERTH_AutoDW_Task_3.1: Yes.

CTL: Possibly. If we proceed with publishing them under research purposes they will be open freely to any researcher.

DBC: Yes, the data will be accessible via free and standardized access protocols such as APIs or repositories supporting open access data.

BOSCH: If possible, yes, following ALFIE's recommendations.

KInIT_AIEF: Yes.

KInIT_AIoD: It already is through the AIoD platform.

KInIT_T2.2_Survey: Will the data be accessible through a free and standardized access protocol?

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: No.

DAS_Post Text data: No, Complex interactions with any user content licence meaning sharing the raw data not feasible

DAS_Stakeholder_Diversity: No, against GDPR to do so

DAS_ETD-Hub_Software: No, against legitimate interests, against internal regulations

D.19. If there are usage restrictions, during or after the end of the project?

CERTH_AutoDW_Task_3.1: It depends on the owner of the data. Accessibility and usage of data will be determined by the license selected by the owner.

CTL: If we proceed with publishing the dataset, we will set very strict restrictions for its usage (e.g. only for research purposes).

DBC: If there are any usage restrictions, these will be specified in the dataset's terms of use and may relate to ensuring ethical and responsible use, in line with the project's objectives and the Grant Agreement.

BOSCH: In clarification as explained in B.1 point.

KInIT_AIEF: No.

KInIT_AIoD: We're not aware of any restrictions.

KInIT_T2.2_Survey: Those defined in GDPR.

KInIT_T2.2_Annotation: No.

DAS_Thread_Summaries: No.

DAS_Post Metadata: Yes - cannot be matched with other data to attempt to identify users

DAS_Post Text data: Yes - cannot be matched with other data to attempt to identify users, cannot be used for commercial purposes

DAS_Stakeholder_Diversity: During and after the project. Legal restriction, against GDPR to do so

DAS_ETD-Hub_Software: During and after the project. Legal & Contractual restrictions - against legitimate interests, against internal regulations

D.20. Would you like access requests to be processed by a data access committee before access is granted?

CERTH_AutoDW_Task_3.1: The access requests may go directly to the owner of the data and then the owner decides if access is given or not.

CTL: If we proceed with publishing the IRIS dataset then yes, we intend to form an internal data access committee that will check the application forms of each request.

DBC: No, we aim to make the dataset openly accessible without the need for an access committee, although access may be monitored for compliance with ethical use.

BOSCH: Yes.

KInIT_AIEF: No.

KInIT_AIoD: No.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: No.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: Access is not to be granted

DAS_ETD-Hub_Software: Access is not to be granted

D.21. Do you intend to make the metadata openly available and licenced under a public domain dedication (CC0), as per the Grant Agreement?

CERTH_AutoDW_Task_3.1: Yes

CTL: Currently no. Might reconsider in the future.

DBC: Yes, the metadata will be made openly available and licensed under a public domain dedication (CC0), in accordance with the Grant Agreement, to ensure maximum accessibility and reuse.

BOSCH: In clarification as explained in B.1 point.

KInIT_AIEF: Yes.

KInIT_AIoD: N/A, it's not ALFIE dataset.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: No.

D.22.If not (D.7.), please clarify why.

CTL: The data contain personal sensitive information and they cannot be shared freely without consent.

DBC: N/A

BOSCH: Data is downloaded from an Internet paid service; terms of use have to be reviewed.

KInIT_AIEF:

KInIT_AIoD: N/A

KInIT_T2.2_Survey:

KInIT_T2.2_Annotation:

DAS_Thread_Summaries: N/A - Yes to D.7.

DAS_Post Metadata: N/A

DAS_Post Text data: N/A

DAS_Stakeholder_Diversity: Will breach GDPR

DAS_ETD-Hub_Software: Will breach legitimate interest and internal regulations

D.23.Will the metadata contain information to enable the user to access the data?

CERTH_AutoDW_Task_3.1: Yes. Can be used to find the data within AutoDW, access depends on the owner.

CTL: No, not for the time being.

DBC: Yes, the metadata will include relevant information such as dataset access links, usage rights, and any other required details to ensure users can easily find and use the data.

BOSCH: N/A

KInIT_AIEF: Yes.

KInIT_AIoD: It already does contain them through the AIoD platform.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: Yes.



DAS_Thread_Summaries: No.

DAS_Post Metadata: No (Unless this means to request deletion of their data or download of their data, then yes)

DAS_Post Text data: No (Unless this means to request deletion of their data or download of their data, then yes)

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: No - software output not a dataset

D.24. For how long are you happy for your data to remain available and findable after the end of the project?

CERTH_AutoDW_Task_3.1: Most likely data will be removed from the AutoDW after the end of the project. Specific custom datasets created for the project use cases are owned by the participating partners and they will hold on to the datasets after the end of the project.

CTL: 2 years after the end of the project.

DBC: The data should remain available and findable indefinitely, ensuring long-term access and enabling future research and reuse.

BOSCH: In clarification as explained in B.1 point.

KInIT_AIEF: Indefinitely.

KInIT_AIoD: N/A, managed by the AIoD platform.

KInIT_T2.2_Survey: N/A

KInIT_T2.2_Annotation: Indefinitely.

DAS_Thread_Summaries:

DAS_Post Metadata: 12 Months

DAS_Post Text data: 12 Months

DAS_Stakeholder_Diversity: Never

DAS_ETD-Hub_Software: Platform may remain up but code never

D.25. Can metadata remain available after the data is no longer available?

CERTH_AutoDW_Task_3.1: No, will be deleted alongside the dataset they are associated with.

CTL: Yes. They might be used by 3rd parties to train their drowsiness and stress level estimation models and compare with their own models.

DBC: Yes, metadata can remain available after the data is no longer accessible, as it still provides valuable information on the dataset's contents and provenance.

BOSCH: Yes.

KInIT_AIEF: N/A

KInIT_AIoD: N/A, managed by the AIoD platform.

KInIT_T2.2_Survey: N/A

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: N/A - No to B.2.

DAS_ETD-Hub_Software: No.

D.26. Is there any software needed to access or read the data?

CERTH_AutoDW_Task_3.1: No.

CTL: Not decided yet.

DBC: No.

BOSCH: The data is in CSV format, which is readable by standard software tools.

KInIT_AIEF: No.

KInIT_AIoD: Not necessarily, it can be accessed through a standard REST API.

KInIT_T2.2_Survey: No

KInIT_T2.2_Annotation: No.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: Web Browser

D.27.If yes (D.12.), will you include documentation/reference to that software?

BOSCH: Standard CSV documentation would be sufficient.

KInIT_AIEF:

KInIT_AIoD: N/A

KInIT_T2.2_Survey: N/A

DAS_Thread_Summaries: Yes (it's just notebook)

DAS_Post Metadata: Yes (it's just notebook)

DAS_Post Text data: Yes (it's just notebook)

DAS_Stakeholder_Diversity: N/A - No to D.12.

DAS_ETD-Hub_Software: No.

D.28.If yes (D.12.), would you include the software in your data release?

BOSCH: Not applicable as CSV is a standard format.

KInIT_AIEF:

KInIT_AIoD: N/A

KInIT_T2.2_Survey: N/A

DAS_Thread_Summaries: No.

DAS_Post Metadata: No.

DAS_Post Text data: No.

DAS_Stakeholder_Diversity: N/A - No to D.12.

DAS_ETD-Hub_Software: No.

DATA INTEROPERABILITY

E.5. How will you make your dataset/data interoperable to allow data exchange and re-use within and across disciplines?

(Will you follow any vocabularies, standards, formats or methodologies for your data/metadata?)

CERTH_AutoDW_Task_3.1: We will follow metadata standards like Dublin Core (DC), Data Documentation Initiative (DDI), or ISO 19115. This depends on the data requirements for our use cases.

CTL: Data will be stored in a Database or Data Warehouse (DW) in standard format and code to access them from there through appropriate APIs. This will follow the standard practices of saving data into a database to make it safe and compatible across disciplines.

DBC: The dataset will follow common data formats. Metadata will follow the DCAT (Data Catalogue Vocabulary) and other relevant standards to ensure it can be re-used across disciplines.

BOSCH: The data is in CSV format, which is widely interoperable. Standard vocabularies for compliance classifications should be adopted.

KInIT_AIEF: N/A

KInIT_AIoD: It already is through the AIoD platform. It's accessible through a standard protocol (HTTP, REST API) and it's well described to be machine-readable (OpenAPI specification is available).

KInIT_T2.2_Survey: Personal data is not meant for public sharing, they include personal information, and for that reason they must be treated with utmost caution. The statistical data will be published and can be reused by other experts.

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: It inherently is – just plain text

DAS_Post Metadata: It inherently is – just plain text

DAS_Post Text data: It inherently is – just plain text

DAS_Stakeholder_Diversity: We won't - not to be shared or reused.

DAS_ETD-Hub_Software: N/A - software output not a dataset

E.6. Will you follow community-endorsed interoperability best practices? Which ones?

CERTH_AutoDW_Task_3.1: We are going to use open standards, e.g. JSON and XML for structured documents. Follow metadata standards, ensure API compatibility and follow the FAIR principles.

CTL: Not decided yet.

DBC: Yes, we will follow best practices endorsed by the open data community, such as the FAIR principles (Findable, Accessible, Interoperable, Reusable), the DCAT standard for metadata.

BOSCH: Yes

KInIT_AIEF: N/A

KInIT_AIoD: See E.1

KInIT_T2.2_Survey: No

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: No.

DAS_Post Metadata: No.

DAS_Post Text data: No.

DAS_Stakeholder_Diversity: N/A - not to be shared or reused.

DAS_ETD-Hub_Software: N/A - software output not a dataset

- E.7. In case it is unavoidable that you use bespoke project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

CERTH_AutoDW_Task_3.1: Yes.

CTL: No ontologies will be used at the first stage of IRIS development. In the future we might consider extending our dataset with semantics and ontology information.

DBC: If project-specific ontologies or vocabularies are necessary, we will provide mappings to more commonly used ones and openly publish the generated ontologies, allowing others to reuse, refine, or extend them.

BOSCH:

KInIT_AIEF: N/A

KInIT_AIoD: N/A, managed by the AIoD platform.

KInIT_T2.2_Survey: No

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: No.

DAS_Post Metadata: No.



DAS_Post Text data: No.

DAS_Stakeholder_Diversity: No - not to be shared or reused.

DAS_ETD-Hub_Software: N/A - software output not a dataset

E.8. Will your data include links/qualified references to other data (in ALFIE or external)?

CERTH_AutoDW_Task_3.1: Yes.

CTL: As far as currently we are concerned, No.

DBC: Yes, the dataset will include references and links to other related datasets, both within the ALFIE project and externally, to support further research and foster connections.

BOSCH: No

KInIT_AIEF: N/A

KInIT_AIoD: It already does include the links (e.g., to the original data source) through the AIoD platform.

KInIT_T2.2_Survey: No

KInIT_T2.2_Annotation:

DAS_Thread_Summaries: No.

DAS_Post Metadata: No.

DAS_Post Text data: No.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: N/A - software output not a dataset

DATA RE-USE

F.7. What types of documentation will you provide to facilitate data re-use and data analysis validation?

(e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

CERTH_AutoDW_Task_3.1: Readme files with explanations on installing and running the services. Online swagger page for the API, detailing what each endpoint does and how to trigger it.



CTL: Readme files and instructions will be shared but internally to allow reusing the data in a different domain if needed.

DBC: We will provide comprehensive documentation, including a readme file, methodology description, codebooks, data cleaning procedures, and variable definitions to ensure the data can be reused and analyzed correctly.

BOSCH: Documentation should include explanation of the alert sources, classification criteria, and validation methods.

KInIT_AIEF: Codebooks from manual analysis and prompts used for automated analysis.

KInIT_AIoD: OpenAPI specification is available.

KInIT_T2.2_Survey:

KInIT_T2.2_Annotation: Codebooks from manual analysis and prompts used for automated analysis.

DAS_Thread_Summaries: Readme on methods, format, field meanings and origin

DAS_Post Metadata: Readme on methods, format, field meanings and origin

DAS_Post Text data: Readme on methods format, field meanings, and origin

DAS_Stakeholder_Diversity: None.

DAS_ETD-Hub_Software: N/A - software output not a dataset

F.8. Will your data be made freely available in the public domain?

CERTH_AutoDW_Task_3.1: No.

CTL: No

DBC: Yes, the data will be made freely available under an open license, ensuring that others can access and reuse it.

BOSCH: In clarification as explained in B.1 point

KInIT_AIEF: Yes.

KInIT_AIoD: It already is through the AIoD platform.

KInIT_T2.2_Survey: No.

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: No.

DAS_Post Metadata: No.



DAS_Post Text data: No.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: No.

F.9. Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

CERTH_AutoDW_Task_3.1: Not applicable. AutoDW doesn't create any data, only associated metadata for every dataset saved.

CTL: Yes

DBC: Yes, the data will be licensed using standard reuse licenses to allow for easy reuse in accordance with the Grant Agreement.

BOSCH: In clarification as explained in B.1 point

KInIT_AIEF: Yes.

KInIT_AIoD: N/A, managed by the AIoD platform.

KInIT_T2.2_Survey: No

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: N/A - software output not a dataset

F.10. Will the data produced in the project be useable by third parties, in particular after the end of the project?

CERTH_AutoDW_Task_3.1: It depends on the owner of the data.

CTL: Yes, for 2 years after the end of the project.

DBC: Yes, the data will be reusable by third parties, even after the project's completion, provided they follow the specified usage terms and license.

BOSCH: In clarification as explained in B.1 point

KInIT_AIEF: Yes.

KInIT_AIoD: Yes, but it's managed by the AIoD platform.

KInIT_T2.2_Survey: No.



KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: No.

DAS_Post Text data: No.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: N/A - software output not a dataset

F.11. Will the provenance of the data be thoroughly documented using the appropriate standards?

CERTH_AutoDW_Task_3.1: Yes.

CTL: Yes

DBC: Yes, the provenance of the data will be thoroughly documented using recognized standards to ensure that data lineage is clear and traceable.

BOSCH: Yes, the external source of alerts and internal classification process should be documented.

KInIT_AIEF: Yes.

KInIT_AIoD: Managed by the AIoD platform.

KInIT_T2.2_Survey: No

KInIT_T2.2_Annotation: Yes.

DAS_Thread_Summaries: Yes.

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: N/A

DAS_ETD-Hub_Software: N/A - software output not a dataset

F.12. Briefly describe relevant data quality assurance processes.

CERTH_AutoDW_Task_3.1: We will implement functionalities for data validation and data cleaning, metadata documentation versioning capabilities and standardized data entry.

CTL: We will implement version control (e.g. Git). Moreover we will implement automated quality control (e.g. when data is collected ensure that all the fields are present, like the alert field).

DBC: We will implement data validation and cleaning procedures to ensure the accuracy and consistency of the data. This includes checks for completeness, consistency, and adherence to predefined standards.

BOSCH: Data is validated with few null values. The quality metric of particular importance is recall for the "Relevant" class, as it is important not to miss relevant cases that compliance officers need to investigate further.

KInIT_AIEF: Annotations will be cross-checked by multiple annotators.

KInIT_AIoD: Managed by the AIoD platform.

KInIT_T2.2_Survey: N/A

KInIT_T2.2_Annotation: Annotations will be cross-checked by multiple annotators.

DAS_Thread_Summaries:

Schema Conformity: Ensure that each record has the required fields (e.g., post ID, user ID, timestamp, theme).

Integrity Constraints: Confirm that referenced threads actually exist.

Human-in-the-Loop: For summarisation check AI outputs.

Logging and Version Control: Maintain logs of data processing steps and version control for changes to cleaning scripts or data schema.

Anonymisation/De-identification: Remove or mask personally identifiable information (PII) and sensitive content.

Policy Enforcement: Ensure compliance with relevant regulations (e.g., GDPR) during storage and processing.

Regular Audits: Periodically sample post metadata to identify emerging issues like unexpected data formats.

DAS_Post Metadata:

Schema Conformity: Ensure that each record has the required fields (e.g., post ID, user ID, timestamp, theme).

Metadata Checks: Validate timestamps, and post-to-thread relationships.

Integrity Constraints: Confirm that referenced users and threads actually exist.

Logging and Version Control: Maintain logs of data processing steps and version control for changes to cleaning scripts or data schema.

Anonymisation/De-identification: Remove or mask personally identifiable information (PII) and sensitive content.

Policy Enforcement: Ensure compliance with relevant regulations (e.g., GDPR) during storage and processing.

Regular Audits: Periodically sample post metadata to identify emerging issues like unexpected data formats.

DAS_Post Text data:

Schema Conformity: Ensure that each record has the required fields (e.g., post ID, user ID, timestamp, content).

Metadata Checks: Validate timestamps, post-to-thread relationships.

Logging and Version Control: Maintain logs of data processing steps and version control for changes to cleaning scripts or data schema.

Anonymisation/De-identification: Remove or mask personally identifiable information (PII) and sensitive content.

Policy Enforcement: Ensure compliance with relevant regulations (e.g., GDPR) during storage and processing.

Regular Audits: Periodically sample posts to identify emerging issues like new spam patterns or unexpected data formats.

DAS_Stakeholder_Diversity: N/A

DAS_ETD-Hub_Software: N/A - software output not a dataset

RESOURCES ALLOCATION

G.3. What is the costs for making data or other research outputs FAIR? (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

The cost involve servers with appropriate setup to handle saving of large datasets (in terms of storage and protection against drive failures etc.).

CERTH_AutoDW_Task_3.1: Appropriate software and technical expertise of personnel to operate it so that it covers the needs for archiving, search and retrieval, re-use and distribution and security.

CTL: No direct costs are associated with making IRIS data available. Indirect costs include electricity bills for storing and maintaining the data to CTL servers, that do not exceed 1K Euros per year.



DBC: The costs include storage, long-term archiving, data management tools, and ensuring compliance with the FAIR principles. Additional costs may also arise from providing access, security measures, and supporting data reuse.

BOSCH: It needs further evaluation

KInIT_AIEF: N/A

KInIT_AIoD: There're no project-related costs, the data is managed and maintained by the AIoD platform.

KInIT_T2.2_Survey: There are no project-related costs.

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: <€4000 (~€1,400 for storage)

DAS_Post Metadata: <€2,000 (~€700 for storage)

DAS_Post Text data: <€2,000 (~€700 for storage)

DAS_Stakeholder_Diversity: N/A

DAS_ETD-Hub_Software: N/A

G.4. How will they be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

CERTH_AutoDW_Task_3.1: CERTH infrastructure and technical expertise can cover the costs related to making the data or other research outputs FAIR.

CTL: The indirect costs are covered from the CTL capital.

DBC: The costs will be covered as part of the Horizon Europe grant, as they are compliant with the project's data management and open access requirements.

BOSCH: This would need to be addressed as part of the project funding arrangements.

KInIT_AIEF: N/A

KInIT_AIoD: N/A, see G.1

KInIT_T2.2_Survey: N/A, see G.1

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: Horizon Europe grant

DAS_Post Metadata: Horizon Europe grant



DAS_Post Text data: Horizon Europe grant

DAS_Stakeholder_Diversity: N/A

DAS_ETD-Hub_Software: N/A

ETHICS

H.3. Are there, or could there be, any ethics or legal issues that can impact data sharing?

CERTH_AutoDW_Task_3.1: No, ALFIE is a tool which prioritizes ethical AI development hence, the AutoDW cannot be utilized to create and unethical AI solution.

CTL: Yes, the data contains sensitive information that include physiological attributes of the participants (i.e. face, heart-rate).

DBC: Potential ethical and legal issues could include privacy concerns if personal data is involved, or intellectual property considerations that may require a delay in sharing certain datasets.

BOSCH: Yes, the data contains information about entities (organizations, companies) and there are contractual considerations with the provider services that need clarification.

KInIT_AIEF: We have to maintain existing licenses regarding the AI ethics frameworks

KInIT_AIoD: We're not aware of any.

KInIT_T2.2_Survey: Data will include personal information (GDPR applicable).

KInIT_T2.2_Annotation: We have to maintain existing licenses regarding the AI ethics frameworks

DAS_Thread_Summaries: Yes, including damaging legitimate interests – therefore data will not be shared

DAS_Post Metadata: Yes - may be linkable to other datasets to identify participants

DAS_Post Text data: Yes – if stakeholders reveal excessive personal information/selfdox

DAS_Stakeholder_Diversity: Yes substantial – therefore data will not be shared

DAS_ETD-Hub_Software: Yes, including damaging legitimate interests – therefore data will not be shared

H.4. Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?

CERTH_AutoDW_Task_3.1: Yes, however only if those are present in the original data source.

CTL: Yes.

DBC: Yes, informed consent will be included in questionnaires to ensure that individuals are aware of how their data will be used and preserved in the long term.

BOSCH: Not applicable as the data is about organizations rather than individuals.

KInIT_AIEF: N/A

KInIT_AIoD: We're not aware of such obligations.

KInIT_T2.2_Survey: Yes.

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: N/A

DAS_Post Metadata: Yes.

DAS_Post Text data: Yes.

DAS_Stakeholder_Diversity: N/A – data will not be shared or retained, but there will be a thorough disclosure and informed consent document anyway for use in the project

DAS_ETD-Hub_Software: Yes there will be a privacy policy and disclosures on the website

OTHER

I.3. Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management?

CERTH_AutoDW_Task_3.1: No.

CTL: For the moment No.

DBC: Yes, we will follow other relevant national, funder, or sector-specific data management procedures to ensure compliance with all applicable requirements.

BOSCH: Not explicitly mentioned, but Bosch likely has internal compliance data management procedures that would be followed.

KInIT_AIEF: N/A

KInIT_AIoD: We're not aware of any.

KInIT_T2.2_Survey: We're not aware of any.

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: No.

DAS_Post Metadata: No.

DAS_Post Text data: No.

DAS_Stakeholder_Diversity: No.

DAS_ETD-Hub_Software: No.

I.4. I.2. If yes, which ones (please list and briefly describe them)?

DBC: We will adhere to procedures outlined by the EU Open Data Directive, Horizon Europe guidelines, and other relevant data management policies to ensure proper handling of data throughout the project.

BOSCH: Bosch internal compliance processes.

KInIT_AIEF: N/A

KInIT_AIoD: N/A

KInIT_T2.2_Survey:

KInIT_T2.2_Annotation: N/A

DAS_Thread_Summaries: N/A - No to I.1.

DAS_Post Metadata: N/A - No to I.1.

DAS_Post Text data: N/A - No to I.1

DAS_Stakeholder_Diversity: N/A - No to I.1

DAS_ETD-Hub_Software: N/A - No to I.1