D6.1 Data Management Plan

Project Title	Multi-Attribute, Multimodal Bias Mitigation in AI Systems
Contract No.	101070285
Instrument	HORIZON-RIA Research and Innovation Actions
Thematic Priority	HORIZON-CL4-2021-HUMAN-01-24 Tackling gender, race and other biases in AI
Start of Project	1 November 2022
Duration	36 months

MAMMOth is a Horizon Europe Research and Innovation Project co-funded by the European Union under Grant Agreement ID: 101070285 and an UK Research and Innovation grant 10041914.

Deliverable title	**Data Management Plan**
Deliverable number	D6.1
Deliverable version	V1.0
Contractual Date of delivery	30.04.2023
Actual Date of delivery	25.04.2023
Nature of deliverable	Data Management Plan
Dissemination level	Public
Partner Responsible	CERTH
Author(s)	Stavroula Rizou (CERTH), Symeon Papadopoulos (CERTH)
Reviewer(s)	Evren Yalaz (TRI IE), Ian Slesinger (TRI UK)
EC Project Officer	Evangelia Markidou
Abstract	This deliverable reflects the first version of the Data Management Plan of the project, which will be maintained and updated internally throughout the project, by adhering to the Guidelines on Data Management in Horizon Europe.
Keywords	Artificial intelligence, data management plan, data protection, data security, ethics, FAIR data, GDPR, metadata, open data, open repositories, research data

Copyright

Revision History

Version	Date	Modified By	Comments
V0.1	15/11/2022	Stavroula Rizou (CERTH)	First draft Table of Content
V0.2	03/02/2023	Stavroula Rizou (CERTH), Symeon Papadopoulos (CERTH)	First draft sent to partners for contributions
V0.3	28/02/2023	CERTH, EXUS, IDnow	Partners contribution - MAMMOth datasets in Section 3
V0.4	17/03/2023	CSH, CSI, UniBw	Partners contribution - MAMMOth datasets in Section 3
V0.5	17/03/2023	Stavroula Rizou (CERTH)	Further updates and revisions. Additions to the other research outputs in subsection 2.3
V0.6	20/03/2023	Stavroula Rizou (CERTH)	Additions to the types and formats of data in subsection 2.1
V0.7	07/04/2023	Evren Yalaz (TRI IE), Ian Slesinger (TRI UK)	Peer review comments
V0.8	19/04/2023	Stavroula Rizou (CERTH)	Updated version based on internal review comments
V1.0	24/04/2023	Symeon Papadopoulos (CERTH)	Final version to submit

Glossary

Abbreviation	Meaning
AI	Artificial Intelligence
DMP	Data Management Plan
DoA	Description of Action
DOI	Digital Object Identifiers
DPO	Data Protection Officer
FAIR	Findability, Accessibility, Interoperability, and Reusability
GDPR	General Data Protection Regulation
PID	Persistent identifier
URL	Uniform Resource Locator
WP	Work Package

Table of contents

Index of tables

Executive summary

The MAMMOth data management plan (DMP), which corresponds to deliverable D6.1 provides documentation on the datasets that are generated or re-used during the lifetime of the MAMMOth project, implementing the FAIR principles (Wilkinson et al., 2016). Apart from compliance with the FAIR principles, the current document also takes into consideration other research outputs besides data. In parallel, the DMP entails the issues that arise from the allocation of resources, data security, and ethics and legal considerations.

The DMP is based on the "Data Management Plan Template" (version 1.0, 5th May 2021), which was released by the European Commission. The current document constitutes the first version of the MAMMOth DMP, which will evolve alongside other developments as the project progresses, and will support the accomplishment of the project's research objectives. As a result, the goal of the initial DMP is to set up the governance and technical framework of data management, in accordance with Article 17 of the Annotated Grant Agreement.

For the purposes of the project implementation, datasets include the following categories: research data and AI training data consisting of different data types (structured records, free text, images, academic profile data, etc.), operational data mainly in the form of text (user data, log files), and user evaluation data (responses to user evaluation questionnaires, interviews, and co-creation workshops etc.); existing open datasets, which are publicly available online.

More specifically, the presented datasets have been classified based on their relevance to the type of project activity: a) datasets related to the MAMMOth toolkit and its research components, b) datasets related to the three use cases, c) datasets related to communication, dissemination and exploitation and d) datasets related to project management.

This document provides a comprehensive approach to how data will be managed during and after the project. First, it will present the overarching methodology of the data management plan. Second, it will address in a more specific way how the MAMMOth data management plan applies specific principles and relevant considerations for the management of research data. Finally, it will set out what are the datasets that will be used in the project, and specify how they will be managed.

1 Introduction

This deliverable reflects the Data Management Plan for the MAMMOth project. The MAMMOth DMP is a living document, and as such it will be regularly reviewed and updated. According to section 2.2, Annex 5, of the Annotated Grant Agreement an updated DMP deliverable must also be produced mid-project (M18) and at the end of the project (M36). Therefore, DMP will be updated periodically and made internally available to the MAMMOth consortium. The updates to the DMP will reflect the substantial changes over the course of the project, including the generation of new data, changes in legislation, changes in the consortium, etc.

The presented datasets and the guiding questions of the data management refer to research and non-research data during and after the project lifetime, including the data quality, the data storage, the ethical and legal requirements, the long-term preservation, the selected repositories, and the usage licenses. The starting point of the recording for the DMP was four months after the start of the project (M4), by sharing the "Template for the presentation of the data management plan for a specific dataset" with the MAMMOth consortium. The mapping of the datasets represents the current status, when completing the questionnaires, in the context of the project by every consortium partner.

Setting as a central axis all the re-used and generated datasets by the project, the DMP analyses and investigates the key components of the data management policy that will be adopted, targeting also the awareness of the MAMMOth consortium for the data management considerations. Therefore, the DMP intends to present the implementation of the DMP of the MAMMOth consortium and to raise awareness on how to comply with the core data management requirements.

D6.1 is divided into four sections. Section 2 documents the MAMMOth DMP methodology, and presents the overall MAMMOth data management pathway. Section 3 illustrates and analyses the management plan for the specific datasets, which are generated or re-used within MAMMOth in accordance with the guiding questions in the defined methodology in Section 2. Section 4 concludes the deliverable.

2 Data management methodology

The methodology of the MAMMOth data management is based on the "Data Management Plan Template" (version 1.0), which was released by the European Commission on May 5, 2021 and was provided under the reporting templates in the reference documents of the Funding and Tenders portal of the European Commission, as defined by the *HE Programme Guide: V2.0, Section 16, page 40*. This template has been enriched with detailed examples, where appropriate, to accommodate its efficient establishment and implementation.

In particular, the MAMMOth DMP revolves around the following seven key topics:

1. Data summary;
2. FAIR data
    a. Making data findable, including provisions for metadata;
    b. Making data accessible;
    c. Making data interoperable;
    d. Increase data re-use;
3. Other research outputs;
4. Allocation of resources;
5. Data security;
6. Legal and ethical requirements;
7. Other issues.

More specifically, in the following subsections of Section 2, as a first step, we describe the specified guiding questions, associated with the seven key requirements, and as a second step, the overview of MAMMOth data management is presented. Section 3 refers to the responses to these questions for each identified dataset.

2.1 Data summary

The guiding questions of the data summary requirements point to the following main topics:

- Re-use of existing data
- Types and formats of data that the project generates or re-uses
- Purpose of the data generation or re-use in relation to the objectives of MAMMOth
- Expected size of the data
- Data origin
- Data utility

In this subsection, the description of the data that will be generated or re-used is defined, including the short justification of their origin, the kind and volume of data which are included, the purposes of the processing of these data in the context of the project, as well as their broader utility.

Will you re-use any existing data and what will you re-use it for?
The datasets of MAMMOth include both new data, which are collected or produced via the project activities, and existing data that are re-used. The data are processed for the research activities of the project (in developing an innovative fairness-aware AI-data driven foundation) and in general for the overall implementation of the project.

What types and formats of data will the project generate or re-use?
MAMMOth will generate and re-use various types of data, including images, target labels, graphs, node attributes, tabular data, node communities, research papers' spreadsheets, documents, survey data, interview data, videos, and audios. In terms of the data format, the project entails mainly jpeg, png, csv, json, mp4, xls, doc, pptx, txt and pdf. In particular, with regards to graphs, the graphs' format refers to mm, mtx, snap, attributed graphs to gml, node attributes to csv and node communities to snap format. Section 3 details the data types and formats of the identified datasets in MAMMOth at the time of writing this document.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?
MAMMOth focuses on multi-discrimination mitigation for tabular, network and multi-modal data. This goal will be achieved via the generation or re-use of the following dataset types: **Data related to user requirements** analysis and co-creation for the development and deployment of the MAMMOth toolkit that would allow data scientists and stakeholders to conduct reliable analysis of data and extract trustworthy, explainable results. **Data for the extraction of multicriteria attributes from multimodal data and multidimensional bias mitigation methods**, will be processed by the project's technical partners. In particular, these data are essential for developing bias-aware multi-modal methods and tools for bias analysis in multi-attribute graphs, identifying bias encoded in multimodal data. In parallel, the extension of the existing definitions of fairness and new methods for identifying and mitigating bias will be developed using multi-dimensional processes. **Data related to bias assessment** as well as training and evaluation of algorithms in order to address bias across all phases of the development of AI systems, to evaluate and mitigate AI bias, to ensure reliability, traceability and explainability of AI solutions. **Data related to the evaluation** of the MAMMOth toolkit through the project's three use cases, namely algorithmic decision-making in finance, face verification for identity authentication and ranking algorithms for citation searches; the goal is to be assessed by real-life data with a large impact on society. **Data related to communication, dissemination and exploitation of the project,** in order to ensure project visibility, and to develop an exploitation plan for the MAMMOth toolkit, to raise awareness about preventing gender and intersectional bias and improving the scientific basis of future policies addressing explainable and trustworthy AI.

Data related to the project management activities, including the administrative processes, the project coordination and monitoring by respecting scientific principles as well as ethical, legal and data protection requirements under the Horizon Europe framework.

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

The detailed size of every dataset generated or re-used is presented in Section 3.

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

The provenance of all the datasets, which are processed in MAMMOth, is documented in Section 3. In general, the crucial sources refer to:

- Open repositories such as GitHub[1] and Zenodo[2]
- Surveys and interviews to collect community feedback
- Co-creation workshops
- Consortium administrative data
- Existing datasets of the use case partners
- Web (e.g., news articles & comments)
- Questionnaires filled in by project partners
- Other shared data by external researchers.

To whom might your data be useful, outside your project?

The MAMMOth datasets presented in this DMP are necessary for implementing the project's objectives and expected impacts. Furthermore, the datasets, which are processed in the context of the project, would have an impact on AI developers, software engineers and especially researchers with a focus on trustworthy and explainable AI, bias-preventing AI solutions and in general on the development or/and auditing similar concepts and frameworks with MAMMOth. One of the main MAMMOth goals in this direction is the incorporation of MAMMOth tools within existing machine learning libraries and tools.

2.2 FAIR data

The FAIR principles (Wilkinson et al., 2016) are addressed in this document by achieving the following points: 2.2.1, 0, 2.2.3, and 2.2.4.

[1] https://github.com/
[2] https://zenodo.org/

2.2.1 Making data findable, including provisions for metadata

This point includes the following issues:

- Will data be identified by a persistent identifier?
- Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?
- Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?
- Will metadata be offered in such a way that it can be harvested and indexed?

Initially, the majority of the data generated or re-used by the project will be made discoverable and shared. The guiding questions for this point, are presented below:

Will data be identified by a persistent identifier?

Publicly available datasets that are re-used in MAMMOth will be identifiable from their sources. This information will be provided on a per dataset basis in Section 3. In addition, the generated datasets from MAMMOth partners, unless otherwise justified, will be uploaded to open repositories (e.g., Zenodo), making them identifiable. The MAMMOth datasets that are solely processed internally, for confidentiality reasons, are identifiable by consortium partners (totally or partially).

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?

The data of the project, in general, includes version controls, and clear folder structures, enabling well-ordered research data that can be re-used. The metadata production for the internal datasets uses a “readme” text file and clear file headers. In addition, naming conventions are followed for the dataset documentation, according to the “Template for the presentation of datasets in the data management plan”:

MAMMOth_<DatasetNo>_<DatasetCategory>_<DatasetTitle>_V<VersionNumber>
Partner: <Short name of partner processing this data>

The research datasets that will be shared via open repositories, and will therefore adhere to the metadata standards that have been provided by these repositories. The selection of trusted repositories for the research data and scientific publications of the project contributes to the adoption of metadata standards that are broadly accepted by the scientific community. It should be mentioned that the Zenodo repository, which will be enforced in MAMMOth, fulfils all mandatory metadata requirements of the open science requirements in the Horizon Europe Model Grant Agreement, according to a recent study that has been prepared for the European Research Council Executive Agency (Jahn et al., 2023).

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

Representative keywords will be provided in cases where this is possible.

Will metadata be offered in such a way that it can be harvested and indexed?
The metadata of the datasets, which will be shared through repositories, will be indexed in a searchable resource.

2.2.2 Making data openly accessible

The second element of the FAIR data refers to:

- Repository
- Data
- Metadata

The following questions are directed by the three key pillars regarding data access: repositories, data and metadata; they reflect the general approach of the project to these directions.

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?
As stated in the DoA, open research datasets and accompanying metadata within MAMMOth, are uploaded to the Zenodo repository, which is also connected to OpenAIRE, aiming at maximum accessibility. It should be mentioned that Zenodo is provided as an example for open repositories in the AGA – Annotated Model Grant Agreement (V0.2). In addition, Zenodo "has been identified as trusted because it fulfils all the essential characteristics required - policy, (open) access and PID assignment, metadata requirements", according to the study, which was prepared for the European Research Council Executive Agency, regarding the examination of the repositories that comply with the open science requirements of the Horizon Europe Model Grant Agreement (Jahn et al., 2023). Discoverability is reinforced by Zenodo's Digital Object Identifiers (DOI), which are granted to uploaded data.

Will all data be made openly available?
Apart from the publicly available datasets that are re-used in MAMMOth, there are also generated datasets within MAMMOth. In general, the research datasets and accompanying metadata will be openly available. However, there are cases, especially in terms of the existing datasets of the use case partners of the project, in which datasets cannot be shared either with the MAMMOth consortium or to the public. There are two cases in which open access cannot be provided, by this time, in accordance with Article 17 of the *MAMMOth Grant Agreement (ID 101070285)*. In particular, a number of datasets cannot be shared due to the GDPR provisions over sensitive personal data. Furthermore, a part of the existing datasets of the use case partners cannot be shared due to confidentiality reasons, related to the beneficiary's legitimate interests.

Will the data be accessible through a free and standardised access protocol?
Through the selected repositories (e.g., Zenodo, GitHub, etc.), free and easy access is provided to the research results and publications of the project.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?
In line with Article 17, Annex 5 of the *MAMMOth Grant Agreement (ID 101070285)*, the open access data will be deposited under Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a license with equivalent rights. More specifically CC-BY 4.0 (Creative Commons Attribution 4.0 International License) license, or any latest recommended suite, will be considered.

How will the identity of the person accessing the data be ascertained?
The identity examination for access to the MAMMOth data depends on the type of the dataset. In particular, this obligation does not apply to open access datasets that are generated in the project or publicly available existing datasets. For the necessary internal data processing by the project's partners, due to reasons that have been already defined, access control procedures are set that define access rights, role-based access and provide secure access with username/password credentials.

Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?
At the time of writing this DMP, no such issue has been raised.

Will metadata be made openly available and licensed 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?
In line with Article 17, Annex 5 of the AGA – Annotated Model Grant Agreement (V0.2), the openly available metadata, which are produced in the context of MAMMOth's research data, will be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded). Regarding accessibility, it is provided through the included metadata and search facilities of Zenodo, or relevant trusted repositories.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?
In general, the openly shared datasets by the project will be continuously available and findable, as well as the respective metadata.

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)?
This question is addressed in relation to each dataset in Section 3.

2.2.3 Making data interoperable

The third element of the FAIR data refers to:

- Interoperability
- Use of vocabularies
- Qualified references

Aiming at interoperable data includes the use of standard vocabularies and formats both for data and metadata, allowing data exchange, combining them in an automatic way, and distinguishing the metadata from the research data files.

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?
Apart from the data that will be deposited in trusted repositories, rich metadata in line with the FAIR principles will follow specific standards. In particular, MAMMOth is following the OpenAIRE guidelines for metadata[3] regarding interoperable standards.

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?
In case it would be impractical to avoid the usage of uncommon ontologies or vocabularies, effort will be made to provide mappings or/and publications of the generated ontologies/vocabularies.

[3] OpenAIRE guidelines for metadata: https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm.

Will your data include qualified references[4] to other data (e.g., other data from your project, or datasets from previous research)?
Efforts will be made to ensure that data and metadata will include qualified references to other data or/and metadata, if such correlations have been identified.

2.2.4 Increase data re-use

The fourth factor of the FAIR data refers to:

- Documentation
- License
- Usable by third parties after the end of project
- Data provenance
- Quality control measures

The overall plan of the project for making its research outputs available for use and re-use by other researchers, is reflected in the following questions.

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.)?
This will be examined on a case-by-case basis, depending on the dataset. In general, MAMMOth will encourage the re-use of its datasets, by providing "readme" text files, file headers, or/and any other documentation that could contribute in this direction.

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?
According to Article 17, Annex 5 of the *Grant Agreement of MAMMOth (ID 101070285)*, the research data of MAMMOth will be deposited under the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a license with equivalent rights, unless the restrictive conditions defined in this Article are applied. More specifically CC-BY 4.0 (Creative Commons Attribution 4.0 International License) license, or any latest recommended suite, will be considered.

[4] *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?
The generated research datasets within MAMMOth, which will be shared via open repositories, will be available for re-use after the end of the project as well.

Will the provenance of the data be thoroughly documented using the appropriate standards?
The datasets that are processed in MAMMOth ensure that the information about the data (including versioning, sources, etc.) is specified and documented in order for the provenance to be traced. In addition, it should be mentioned that the generated research datasets within MAMMOth, which will be shared via open repositories, ensure that the metadata contain detailed information about the provenance of the data, following the standards of trusted repositories.

Describe all relevant data quality assurance processes.
In general, automatic data cleaning techniques will be applied to improve the data quality. Furthermore, peer review of data will be employed. For datasets with questionnaire data, manual quality control will be implemented to ensure data quality. In particular, the quality assurance processes constitute a component of Task 6.2 and will be described in the internal Quality Control Plan of the project.

2.3 Other research outputs

The DMP plan takes into consideration the existence of other research outputs, apart from data, that may be generated or re-used throughout the project lifetime. Other research outputs could be digital (such as software, workflows, protocols, and models) or physical (such as new materials, and samples).

- Existence of other research outputs, apart from data
- Apply FAIR principles to other research outputs

The following questions are clarifying the issue of other research outputs, except from data, in the project:

Are there any other research outputs that may be generated or re-used throughout the project?
The research outputs apart from data (based on current evidence) that will be generated or re-used within MAMMOth would include software, code examples, code libraries and documentation, protocols, data formats, algorithms and models.

Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?
As mentioned above, MAMMOth will use GitHub's repository or similar services. This repository will also be used to deposit research outputs that will be generated in MAMMOth other than data. GitHub, in combination with Zenodo, can be used in order to provide unique identifiers (e.g., DOI or versions), pointing at discoverable, reusable and accessible research elements. Furthermore, the respective license, in line with the obligations set out in the Grant Agreement, will be provided to these research outputs [Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a license with equivalent rights], aiming at re-using the research outputs of MAMMOth.

2.4 Allocation of resources

This point addresses the following issues:

- Costs for making data FAIR
- Reimbursement of the costs
- Responsible for data management
- Long-term preservation

Regarding the individual questions for the allocation of resources, our overall DMP approach is summarised below. The detailed answers for each dataset are presented in Section 3, except for the indication of the DMP responsible party, which constitutes a consistent element and is addressed solely in the current Section.

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, staff time, etc.)?
As the preparation, delivery and scheduled update of the DMP are covered in terms of personnel effort by the relevant task T6.3 of the project, and in particular, by the processes for addressing the requirements of the FAIR principles, the allocation of the relevant resources refers to these project management activities. Furthermore, regarding the repository charges, it should be mentioned that the selected trusted repository (Zenodo) by the project is provided without any charges. As a result, no additional costs will be allocated.

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).
If additional costs, related to research data or/and output management, arise in the next stages of the project's lifetime, they will be covered in terms of their nature and in accordance with the Grant Agreement.

Who will be responsible for data management in your project?
The data management of MAMMOth will be led by CERTH, directed by the identified datasets of the MAMMOth consortium. CERTH is responsible for implementing the DMP, and for ensuring it is reviewed and, if necessary, revised. CERTH is the lead beneficiary of the D6.1 Data Management Plan.

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)?
The internal data (generated or re-used), which cannot be shared publicly, due to certain conditions, will be stored and used up to five years after the last project payment or until the data subject's consent is withdrawn (whichever is sooner) in the case of personal data, after which point it will be deleted. The research datasets, that will be shared via open repositories, will remain available after the project ends without cost.

2.5 Data security

This essential part of the DMP addresses the:

- Security measures
- Repositories policies and procedures

In the context of the project, the following data security issues are examined:

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)?
Data security is aimed at ensuring the "integrity" and "availability" of personal data [Article 5 par.1 (f) GDPR] (Rizou et al., 2020). They should be accessible to the responsible parties and should not be changed or deleted by unauthorised persons, implementing the triplet of "confidentiality, integrity and availability" (Wolters P. T. J., 2017). According to Article 5 par.1 (f) GDPR, the level of data security inside the project depends on the defined risks for the data subjects (in case of unauthorised access or disclosure, accidental deletion or destruction of the data), in line with the principle of proportionality. Furthermore, it is important to mention the issue of the existence of mixed datasets (where data security measures have been implemented) in the context of the research activities of MAMMOth. In general, a mixed dataset consists of both personal and non-personal data (e.g., anonymised data) (EC, 2019). If the non-personal data part and the personal data part are "inextricably linked", GDPR compliance is taken into consideration within MAMMOth for the whole mixed dataset, even if the personal data represent only a small part of this dataset. In terms of data storage and archiving, MAMMOth datasets are either openly accessible or accessible only by the project's partners, following the principle "as open as possible, as closed as necessary". In the first case,

the data will be uploaded to a trustworthy open repository, while in the second case, the datasets are stored in the partners' organisation premises.

Open repository
Datasets to be openly accessible will be deposited in Zenodo, where MAMMOth has created a MAMMOth community[5], or other repositories, which also provide a data security strategy. Zenodo servers are managed via OpenStack and Puppet configuration management system. Apart from that, GitHub or other similar platforms will be used.

MAMMOth's file repository
CERTH, as project coordinator, is hosting the datasets related to project management and selected research activities in their premises, based on Google Drive, where stored data is encrypted in-transit and at-rest (according to the application's provided information). Every partner of the consortium (only the relevant staff that was indicated by their organisations) has access to this repository.

The security measures, which are adopted by CERTH, are presented as follows:

- Regular back-ups of the project repository, based on Google Drive, in order to maintain the integrity of the data are conducted on a weekly basis by CERTH. The servers are securely located and monitored in the main building of the Information Technologies Institute of CERTH. The local network is firewall protected from all external traffic and access to CERTH's servers is granted through password-protected SSH with only selected trusted employees maintaining privileged accounts.
- The data that are stored in Google Drive are non-confidential and do not include sensitive personal data. The access is encrypted with HTTPS and a user login is required to access any of their content.

MAMMOth partners' servers
MAMMOth partners, especially the technical partners that process data on a large scale, have significant experience in data security and protection both in terms of their institutional operations and in the context of their participation in other EU-funded projects. The partners, apart from their existing operational policies, consider the data security and protection issues within the project, by consulting the Research Ethics Committees (both CERTH's and UNIBO's have already provided their ethics approval and advice in March 2023) and by raising constantly relevant questions to the Ethical and Legal Adviser of the project. Each partner is responsible for the day-to-day enforcement of the appropriate security measures, according to the principle of accountability.

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

Personal data will be stored and used for up to five years after the last project payment or until consent is withdrawn (whichever is sooner), after which point it will be deleted. Regarding the process of data deletion, paper files must be destroyed in a secure manner (including a shredder). Electronic files must be deleted with specialised software, suitable to remove all data from the material; All personal data will be deleted from the databases of the consortium with proper software and/or hardware procedures rendering unauthorised restoration impossible (e.g., DiskWipe). In case any partner has a parallel legal obligation to further process any kind of data, those partners will further process the data on the given legal basis, while all other partners shall carry out the deletion process.

[5] https://zenodo.org/communities/mammoth/?page=1&size=20

2.6 Legal and ethical requirements

The legal and ethical requirements, within the DMP, are identified in the following main issues:

- Ethics or legal issues
- Informed consent

An overview of the ethical and legal considerations within MAMMOth, is presented below. The following presentation is not exhaustive, as it results from the specific questions in the DMP template, guiding the legal and ethical analysis.

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).

The assessment of ethical and data protection considerations, from the earliest stages of MAMMOth, is conducted from a horizontal perspective in relation to its tasks. The following paragraphs present the generic MAMMOth approach to ethical and data protection requirements; firstly, clarifications are provided from a procedural point and then the ethical and legal considerations are presented separately, from a substantive point of view.

Regarding the procedural framework of the 2.6 point of the DMP, having as a starting point the part A - section 4 "Ethics & security", the project's ethical and data protection implications has been carried out through the dedicated Task (T6.3), which is entitled: "Ethical, legal and data management". Trilateral Research (TRI) constitutes the consortium party, which is responsible for the ethical and legal management with extensive experience in leading respective efforts, holding the role of the project's "Ethical and Legal Adviser". More specifically, in the context of this task, an internal Legal, Ethics and Data Protection Compliance Protocol providing guidance for the MAMMOth consortium was prepared and shared with the partners. Following the release of the internal Legal, Ethics and Data Compliance Protocol on M5, an internal webinar was organised on M6 by the responsible partners (CERTH, TRI) of T6.3 in order to raise awareness for the ethical and legal implications of the project, to guide the MAMMOth consortium, and to respond to partners' questions and doubts regarding ethical and legal issues. Furthermore, the designation of a Data Protection Officer[6] by each partner's organisation was investigated and recorded, in relation to the existence of personal data processing within MAMMOth.

In terms of ethical considerations, within the project, the mapping of the relevant aspects of the tasks has been carried out from the earliest stages of the project. In particular, decisive issues have been examined, concerning all project tasks, in order to obtain approval of a Research Ethics Committee, where applicable and to clarify different aspects of the involvement of human participants. In general, the internal Legal, Ethics and Data Protection Compliance Protocol (as part of T6.3), which is a living document and complementary to this DMP, monitors all partners' research activities in terms of the individual project's tasks.

Data protection requirements within MAMMOth are directed by the General Data Protection Regulation (2016/679). In general, GDPR sets out provisions regarding scientific research in Article 5, Article 9 par. 2, Article

[6] Article 37 of GDPR.

89, Recital 50 and Recitals 156-163[7]. In particular, in relation to the legitimate purposes of the personal data collection, in the context of MAMMOth, the data processing includes both data that were initially collected outside the context of the project and data that are collected from the project's activities as well. In the first case, data processing is based on Article 5 par. 1 (b) and Recital 50 GDPR resulting in the processing of the pre-existing datasets, without the necessity of the examination of a legal basis for the activities of this research project, as they are considered compatible with the initial purposes. In the second case, the informed consent of all data subjects consists of the legal basis for this data processing [Article 6 par 1 (a)]. Sensitive personal data will be processed for the project's research purposes on addressing race, gender, sex and ethnicity-related biases. The processing of sensitive personal data is permitted, for the activities of the research project, according to Article 9 par. 2 (j) GDPR, in line with the technical and organisational measures of Article 89 par. 1 GDPR.

As for the right to be informed, it should be mentioned that in case of further data processing for a different purpose, the data subjects must be informed, according to Articles 13 par. 3 and Article 14 par. 4 GDPR before further processing, whether the secondary purpose is compatible or not. In case of the data collected through the project's research activities, this information is provided (Referred documents: Informed Consent, Participant Information Sheet) for the data subjects to be able to exercise their other data rights as well. In cases where personal data have not been obtained from the data subject, according to Article 14 par.5 (b) the obligation to provide information does not apply if it "…*proves impossible or would involve a disproportionate effort, in particular for processing for scientific research purposes when the conditions of Article 89 are satisfied or when this is likely to render impossible or seriously impair the achievement of the objective of that processing*". In the context of MAMMOth, the controller of each dataset assesses the effort to provide the information to data subjects against the impact on the data subject if they are not provided with the information. This balancing takes into consideration the number of data subjects, the age of the data and any appropriate safeguards adopted (Recital 62 GDPR). Furthermore, the current public document presents the data management crucial aspects of the project's datasets (including the existing ones), contributing to data transparency.

Regarding the cross-border data flows from the EEA to third countries, the analysis of the participation of a UK associated partner in MAMMOth should be presented. In particular, Trilateral Research UK is involved in the MAMMOth project as an associated entity of Trilateral Research Ireland. Given that Trilateral Research UK is not involved in data analysis or technology development/deployment activities, only personal data stemming from project activities (e.g., involved researchers' contact details) will be exported to the UK. Since 31 January 2020, the UK is no longer an EU country and the EU regulations (including GDPR) applied by 31 December 2020. From 1 January 2021 the data flows continued to be free from EEA to UK, according to the *Trade and Cooperation Agreement.* On June 28, 2021, the EC adopted an adequacy decision for the UK[8], ensuring the continued free flow of personal data from individuals inside the EEA to the UK. As a result, the data flows within MAMMOth are conducted freely from the beginning of the project (1 November 2022) among all the partners, including the UK associated partner.

[7] https://www.dpa.gr/el/enimerwtiko/thematikes_enotites/eidikoiskopoi/paideiaereuna (accessed 25 July 2022)

<u>Informed consent:</u> Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?
The project's activities that require the processing of personal data, especially personal data that will be initially collected for the purposes of the project (see the previous question of the 2.6 for further details), are accompanied by the documents of the given Informed Consent and the Participant Information Sheet. The Participant Information Sheet provides information about the crucial aspects of data processing, implementing Articles 13 and 14 of GDPR, as the provided information enables data subjects to exercise their other rights (Articles 15-22 GDPR) as well. In addition, it is important to mention that the Information sheet, in line with the principle of transparency (Recital 39 GDPR), uses clear and plain language. The issues of data sharing and long-term preservation have been addressed by the Informed Consent Form and the Participant Information Sheet that have been prepared. More specifically, regarding the aspect of the defined time limits for the preservation of personal data, the information, that accompanies the consent, is specified as five years after the end of the project in order to be able to fulfil reporting obligations to the European Commission. In terms of the consent, the purposes of the data processing are being defined along with the context (including the specific data controller with their contact details, security measures, the data subjects' rights in terms of the processing of their data and the withdrawal of consent). In particular, the liabilities of GDPR in terms of the data subjects' rights, which are mentioned, are the following: ● Rights to access their personal data (Article 15 GDPR), and the right for these data to be in a portable form (Article 20). ● Right to rectify their personal data (Article 16 GDPR). ● Right to restrict the processing of their personal data (Article 18 GDPR). ● Right to demand the erasure of data subjects' personal data (Article 17 GDPR). ● Right to complain to a supervisory authority (Article 21 GDPR), by also providing reference/information about the competent European Data Protection Authority.

2.7 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)?
Apart from this analysis of the general DMP methodology, every organisation implements data protection and security measures, while making research data FAIR, to each component of the project and is responsible for the day-to-day enforcement, according to the principle of accountability. The existence of other issues, which are not included in the previous questions, is examined on the basis of the presented datasets in Section 3.

3 MAMMOth datasets

This section presents all the datasets that have been processed at the time of writing this deliverable for achieving the project's objectives. The publicly available datasets, processed by the MAMMOth consortium, have been grouped, in cases where the content of the data is similar, considering also the partner who processes each dataset. Therefore, this aggregation facilitates the publicly available dataset documentation in MAMMOth and their re-use for further similar research purposes. The "Template for the presentation of datasets in the data management plan[8]" consists of categorised questions and relevant definitions or/and examples for the essential inputs. The following Table 1 indicates the questions, to which every partner of the MAMMOth consortium that processes datasets, at the time of writing this deliverable, responded to. The main pillars of the template refer to the data summary, the requirements for making data FAIR, the potential other research outputs (e.g., software) of MAMMOth, the allocation of resources in the context of data management, the data security, the ethical and legal requirements, as well as other issues in relation to DMP.

Table 1. Template for the presentation of datasets in the data management plan

DMP component	**MAMMOth_DatasetNo_DatasetCategory_DatasetTitle_Version** **Partner: Short name of partner processing this data**
1. Data Summary	**Re-use of existing data: 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)* **Type/format: What types and formats of data will the project generate or re-use?** *(examples of types: databases, spreadsheets, textual (documents), image, audio, video, and/or mixed media), (examples of format: pdf, xls, doc, txt, or rdf)* **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** **Expected size: What is the expected size of the data that you intend to generate or re-use?** (*e.g., bytes, and/or in numbers of objects, files, rows, and columns)* **Data origin: What is the origin/provenance of the data, either generated or re-used?** **Data utility: To whom might your data be useful, outside your project?**
2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** *(e.g., persistent and unique identifiers such as Digital Object Identifiers)* **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** *(examples of metadata standards: DDI, TEI, EML, MARC, CMDI)* *(In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how)* **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?**

[8] This template was created by the authors of the D6.1 and is based on the "***Data Management Plan Template***" (version 1.0, 5th May 2021), which was released by the European Commission, and on the "**Science Europe. (2021). *Practical Guide to the International Alignment of Research Data Management - Extended Edition***".

	<u>Findable metadata:</u> Will metadata be offered in such a way that it can be harvested and indexed?
2.2 Making data openly accessible	**<u>Repository:</u>** **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?** **<u>Data:</u>** **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 standardised 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)?** **<u>Metadata:</u>** **Will metadata be made openly available and licensed 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 <u>long</u> will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** **Will <u>documentation</u> 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	**<u>Interoperability:</u> 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?** **<u>Use of vocabularies:</u> 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?**

	Qualified references: Will your data include qualified references[9] to other data (e.g., other data from your project, or datasets from previous research)?
2.4 Increase data re-use	**Documentation: 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.)?** **License: 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?** [*Article 17, Annex 5 of the Grant Agreement of MAMMOth: Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a license with equivalent rights*] **Usable by third parties after the end of project: Will the data produced in the project be useable by third parties, in particular after the end of the project?** **Data provenance: Will the provenance of the data be thoroughly documented using the appropriate standards?** **Quality control measures: Describe all relevant data quality assurance processes.** *(These measures may include calibration, repeated samples or measurements, standardised data capture, data entry validation, peer review of data, or representation with controlled vocabularies.)*
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** *[Such outputs can be either digital (e.g., software, workflows, protocols, models, etc.) or physical (e.g., new materials, antibodies, reagents, samples, etc.).]* **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?**
4. Allocation of resources	**Costs for making data FAIR: 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, staff time, etc.)?** **Reimbursement of the costs: 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).** **Long-term preservation: 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)?**

[9] *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/*)*

5. Data security	**Security measures: 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)?** *[Such security measures can include anonymisation of personal data, pseudonymisation, encryption]* **Repositories policies and procedures: Will the data be safely stored in trusted repositories for long-term preservation and curation?**
6. Legal and ethical requirements	**Ethics or legal issues: 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).* **Informed consent: 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)?**

The MAMMOth datasets have been classified, in line with the project structure, to:

a) datasets related to the MAMMOth toolkit and its research components (RES),
b) datasets related to the three use cases (USEC),
c) datasets related to communication, dissemination and exploitation (DISEX) and
d) datasets related to project management (MGT).

Nine datasets have been identified in total, at the time of writing this deliverable (Table 2).

Table 2. Summary of the MAMMOth datasets

DMP component	Relevant sub-section/table
Datasets related to MAMMOth toolkit and its research components (RES)	**3.1**
MAMMOth_001_RES_BiasAssessmentAndMitigation_V1	Table 3
MAMMOth_002_RES_BiasToolkitDevelopment_V1	Table 4
MAMMOth_003_RES_GraphSignalProcessingBiasAssessmentandMitigation_V1	Table 5
MAMMOth_004_RES_Multi-DimensionalBiasAssessmentandMitigation_V1	Table 6
Datasets related to the three use cases (USEC)	**3.2**
MAMMOth_005_USEC_FaceVerif_V1	Table 7
MAMMOth_006_USEC_DebtCollection_V1	Table 8
MAMMOth_007_USEC_AcademicNetworks_V1	Table 9
Datasets related to communication, dissemination and exploitation (DISEX)	**3.3**
MAMMOth_008_DISEX_SocialMediaContent_V1	Table 10
Datasets related to project management (MGT)	**3.4**
MAMMOth_009_MGT_ConsortiumAdminData_V1	Table 11

3.1 Datasets related to MAMMOth toolkit and its research components (RES)

At the time of writing this deliverable four datasets are reported in relation to the development of the MAMMOth toolkit and its research components.

Table 3. Bias assessment and mitigation dataset

DMP component	**MAMMOth_001_RES_BiasAssessmentAndMitigation_V1** **Partner: CERTH**
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** We will re-use FairFace, UTKFace, Biased MNIST, Celeb A & ImageNet (subsets) datasets (all publicly available) for bias assessment and mitigation. **Type/format: What types and formats of data will the project generate or re-use?** image (jpeg, png formats), target labels (csv format) **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** Bias assessment, training and evaluation of algorithms developed in the framework of tasks T2.1 & T2.4. They are related to O2, O3 and O4 specific objectives of the project. **Expected size: What is the expected size of the data that you intend to generate or re-use?** More than 400,000 images with a size of around 5GB in total. **Data origin: What is the origin/provenance of the data, either generated or re-used?** ● FairFace, UTKFace, CelebA, ImageNet: web ● Biased MNIST: synthetic **Data utility: To whom might your data be useful, outside your project?** N/A (these data are publicly available)
2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** No DOI, only URLs ● FairFace: https://drive.google.com/file/d/1Z1RqRo0_JiavaZw2yzZG6WETdZQ8qX86/view ● UTKFace: https://susanqq.github.io/UTKFace/ ● Biased MNIST: http://yann.lecun.com/exdb/mnist/ ● CelebA: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html ● ImageNet: https://github.com/hendrycks/natural-adv-examples and https://github.com/clovaai/rebias **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A **Findable metadata: Will metadata be offered in such a way that it can be harvested and indexed?** N/A

2.2 Making data openly accessible	**<u>Repository:</u>** **Will the data be deposited in a trusted repository?** N/A **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** N/A **<u>Data:</u>** **Will all data be made openly available?** They are already publicly available **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** N/A **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A **<u>Metadata:</u>** **Will metadata be made openly available and licensed 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?** N/A **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** N/A **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)?** N/A

2.3 Making data interoperable	**Interoperability: 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?** N/A **Use of vocabularies: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?** N/A **Qualified references: Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**Documentation: 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.)?** N/A **License: 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?** N/A **Usable by third parties after the end of project: Will the data produced in the project be useable by third parties, in particular after the end of the project?** N/A **Data provenance: Will the provenance of the data be thoroughly documented using the appropriate standards?** N/A **Quality control measures: Describe all relevant data quality assurance processes.** N/A
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** Fair models and software related to training and inference will be open. **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** ● Developed software will have an identifier (GitHub URL) ● Models and software will be publicly available through GitHub ● Documentation will be provided in the form of readme files ● The license under which our research outputs will be is Creative Commons ● Models and software will be usable by third parties after the end of the project

4. Allocation of resources	**<u>Costs for making data FAIR:</u> 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, staff time, etc.)?** N/A **<u>Reimbursement of the costs:</u> 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).** N/A **<u>Long-term preservation:</u> 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)?** N/A
5. Data security	**<u>Security measures:</u> 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)?** The datasets will be stored in CERTH's premises and protected. As the data security (confidentiality, integrity and availability) level, according to Article 5 par.1 (f) GDPR, is based on the defined risks for the data subjects (in case of unauthorised access or disclosure, accidental deletion or destruction of the data), the security measures for the research datasets, regarding bias assessment and mitigation, will be proportionate to the dealing risks. State-of-the-art IT security measures and CERTH's policies mitigate most of the risk of illegitimate access, contributing to the ability to detect promptly an incident and thus restore the data (data availability). Sensitive personal data will be processed for the essential scientific purposes of MAMMOth [Article 9 par. 2 (j), Article 89 par. 1 GDPR]. **<u>Repositories policies and procedures:</u> Will the data be safely stored in trusted repositories for long-term preservation and curation?** N/A
6. Legal and ethical requirements	**<u>Ethics or legal issues:</u> Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** As the datasets refer to publicly available data that have been initially collected outside the context of the project, further data processing, for the research purposes of MAMMOth, is considered compatible with the initial purposes of the data collection. **<u>Informed consent:</u> Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** N/A
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)?** No

Table 4. Bias toolkit development dataset

DMP component	MAMMOth_002_RES_BiasToolkitDevelopment_V1 Partner: EXUS
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** We will re-use Adult Census Income and German Credit Data and other datasets previously mentioned (all publicly available) for toolkit testing purposes. **Type/format: What types and formats of data will the project generate or re-use?** image (jpeg, png formats), target labels (csv format) **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** Testing the MAMMOth toolkit developed in the framework of tasks T1.2 and T1.3. **Expected size: What is the expected size of the data that you intend to generate or re-use?** About 50,000 instances. **Data origin: What is the origin/provenance of the data, either generated or re-used?** Data are available on web. **Data utility: To whom might your data be useful, outside your project?** N/A (these data are publicly available)
2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** No DOI, only URLs • Adult Census Income: https://archive.ics.uci.edu/ml/datasets/adult • German Credit Data: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29 **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A **Findable metadata: Will metadata be offered in such a way that it can be harvested and indexed?** N/A
2.2 Making data openly accessible	**Repository:** **Will the data be deposited in a trusted repository?** N/A **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** N/A

	Data: **Will all data be made openly available?** Already openly available. **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** N/A **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A **Metadata:** **Will metadata be made openly available and licensed 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?** N/A **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** N/A **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)?** N/A
2.3 Making data interoperable	**Interoperability: 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?** N/A **Use of vocabularies: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?** N/A **Qualified references: Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**Documentation: 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.)?** N/A

	<u>License:</u> 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? N/A **<u>Usable by third parties after the end of project:</u> Will the data produced in the project be useable by third parties, in particular after the end of the project?** N/A **<u>Data provenance:</u> Will the provenance of the data be thoroughly documented using the appropriate standards?** N/A **<u>Quality control measures:</u> Describe all relevant data quality assurance processes.** N/A
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** N/A **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** • Developed software will have an identifier (GitHub URL) • Software will be publicly available through GitHub • Documentation will be provided in the form of readme files • The license under which our research outputs will be is Creative Commons • Software will be usable by third parties after the end of the project
4. Allocation of resources	**<u>Costs for making data FAIR:</u> 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, staff time, etc.)?** N/A **<u>Reimbursement of the costs:</u> 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).** N/A **<u>Long-term preservation:</u> 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)?** N/A
5. Data security	**<u>Security measures:</u> 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)?** N/A **<u>Repositories policies and procedures:</u> Will the data be safely stored in trusted repositories for long-term preservation and curation?**

	N/A
6. Legal and ethical requirements	**<u>Ethics or legal issues:</u> Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** As the datasets refer to publicly available data, that have been initially collected outside the context of the project, further data processing, for the research purposes of MAMMOth, is considered compatible with the initial purposes of the data collection. **<u>Informed consent:</u> Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** N/A
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)?** No

Table 5. Graph signal processing bias assessment and mitigation dataset

DMP component	**MAMMOth_003_RES_GraphSignalProcessingBiasAssessmentandMitigation_V1** **Partner: CERTH**
1. Data Summary	**<u>Re-use of existing data</u>: Will you re-use any existing data and what will you re-use it for?** We will re-use SocioPatterns, Pokec, PolBooks, PolBlogs, FacebookCircles, TwitterCircles datasets (all publicly available) for bias assessment and mitigation of graph signal processing algorithms. **<u>Type/format:</u> What types and formats of data will the project generate or re-use?** graphs (mm, mtx, snap format), attributed graphs (gml format), node attributes (csv format), node communities (snap format). The project will convert all data to csv and snap format to assist experimentation. **<u>Purpose</u>: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** Bias assessment, as well as training, tuning, and evaluation of algorithms developed in task T2.3 that combine graph and tabular modalities. They are related to O2, O2, O3, O4, O5 specific objectives of the project. **<u>Expected size</u>: What is the expected size of the data that you intend to generate or re-use?** Each dataset comprises 100-1 million nodes and 200-31 million edges stored in a total of 1GB. **<u>Data origin</u>: What is the origin/provenance of the data, either generated or re-used?** ● SocioPatterns, Pokec, PolBooks, PolBlogs, FacebookCircles, TwitterCircles: web ● pygrank-datasets: synthetic **<u>Data utility</u>: To whom might your data be useful, outside your project?** N/A (these data are publicly available)

2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** No DOI, only URLs ● SocioPatterns: http://www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks ● Pokec: https://snap.stanford.edu/data/soc-Pokec.html ● PolBooks: http://www-personal.umich.edu/~mejn/netdata/polblogs.zip ● PolBlogs: http://www-personal.umich.edu/~mejn/netdata/polbooks.zip ● FacebookCircles: https://snap.stanford.edu/data/ego-Facebook.html ● TwitterCircles: https://snap.stanford.edu/data/ego-Twitter.html ● pygrank-datasets: https://github.com/maniospas/pygrank-datasets Dataset discovery and hydration will also be provided via unique identifiers within developed software. **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A **Findable metadata: Will metadata be offered in such a way that it can be harvested and indexed?** N/A
2.2 Making data openly accessible	**Repository:** **Will the data be deposited in a trusted repository?** N/A **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** N/A **Data:** **Will all data be made openly available?** They are already publicly available. **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** N/A **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A

	Metadata: **Will metadata be made openly available and licensed 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?** N/A **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** N/A **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)?** N/A
2.3 Making data interoperable	**Interoperability: 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?** N/A **Use of vocabularies: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?** N/A **Qualified references: Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**Documentation: 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.)?** N/A **License: 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?** N/A **Usable by third parties after the end of project: Will the data produced in the project be useable by third parties, in particular after the end of the project?** N/A **Data provenance: Will the provenance of the data be thoroughly documented using the appropriate standards?** N/A **Quality control measures: Describe all relevant data quality assurance processes.** N/A

3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** Fair graph signal processing algorithms and software implementations that will be open. **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** ● Developed algorithms will be described in yaml format ● Descriptions of developed algorithms will have an identifier (GitHub URL) ● Algorithms and software will be publicly available through GitHub ● Documentation will be provided in the form of readme files ● The license under which our research outputs will be is Creative Commons ● Developed algorithms will be shared under Apache License 2.0 ● Algorithms and software will be usable by third parties after the end of the project
4. Allocation of resources	**Costs for making data FAIR: 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, staff time, etc.)?** N/A **Reimbursement of the costs: 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).** N/A **Long-term preservation: 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)?** N/A
5. Data security	**Security measures: 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)?** The datasets will be stored in CERTH's premises and protected. As the data security (confidentiality, integrity and availability) level, according to Article 5 par.1 (f) GDPR, is based on the defined risks for the data subjects (in case of unauthorised access or disclosure, accidental deletion or destruction of the data), the security measures for the research datasets, regarding bias assessment and mitigation, will be proportionate to the dealing risks. State-of-the-art IT security measures and CERTH's policies mitigate most of the risk of illegitimate access, contributing to the ability to promptly detect an incident and thus restore the data (data availability). Sensitive personal data will be processed for the essential scientific purposes of MAMMOth [Article 9 par. 2 (j), Article 89 par. 1 GDPR]. In addition, it should be mentioned that these data are anonymised, according to their sources, and thus data subjects are not or no longer identifiable (Recital 26 GDPR). **Repositories policies and procedures: Will the data be safely stored in trusted repositories for long-term preservation and curation?** N/A

6. Legal and ethical requirements	**Ethics or legal issues: Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** As the datasets refer to publicly available data, that have been initially collected outside the context of the project, further data processing, for the research purposes of MAMMOth, is considered compatible with the initial purposes of the data collection. **Informed consent: Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** N/A
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)?** No

Table 6. Multi-dimensional bias assessment and mitigation dataset

DMP component	**MAMMOth_004_RES_Multi-DimensionalBiasAssessmentandMitigation_V1** **Partner: UniBw**
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** We will re-use publicly available datasets like: ACS-PUMS, ADULT, BANK, CREDIT, COMPAS, and others (see our recent survey on datasets for fairness-aware machine learning link) for bias assessment and mitigation. **Type/format: What types and formats of data will the project generate or re-use?** tabular data mainly, csv format. **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** Bias definition, assessment as well as training and evaluation of algorithms in the framework of tasks T3.1 & T3.2. They are related to O1, O2, O3 and O4 specific objectives of the project. **Expected size: What is the expected size of the data that you intend to generate or re-use?** Different datasets have different cardinalities, the largest dataset is from ACS-PUMS which consists of about 7M instances for all five tasks. **Data origin: What is the origin/provenance of the data, either generated or re-used?** All datasets are available on the Web. **Data utility: To whom might your data be useful, outside your project?** N/A (these data are publicly available).

2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** No DOI, only URLs ● ACS-PUMS: https://www.census.gov/programs-surveys/acs/microdata.html ● ADULT, BANK, CREDIT, COMPAS: https://archive.ics.uci.edu **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A **Findable metadata: Will metadata be offered in such a way that it can be harvested and indexed?** N/A
2.2 Making data openly accessible	**Repository:** **Will the data be deposited in a trusted repository?** N/A **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** N/A **Data:** **Will all data be made openly available?** They are already publicly available. **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** N/A **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A **Metadata:** **Will metadata be made openly available and licensed 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?** N/A **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?**

	N/A **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)?** N/A
2.3 Making data interoperable	**<u>Interoperability:</u> 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?** N/A **<u>Use of vocabularies:</u> 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?** N/A **<u>Qualified references:</u> Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**<u>Documentation:</u> 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.)?** N/A **<u>License:</u> 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?** N/A **<u>Usable by third parties after the end of project:</u> Will the data produced in the project be useable by third parties, in particular after the end of the project?** N/A **<u>Data provenance:</u> Will the provenance of the data be thoroughly documented using the appropriate standards?** N/A **<u>Quality control measures:</u> Describe all relevant data quality assurance processes.** N/A
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** Fair models and software related to training and inference will be open. **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** ● Developed software will have an identifier (GitHub URL)

	● Models and software will be publicly available through GitHub ● Documentation will be provided in the form of readme files ● The license under which our research outputs will be is Creative Commons ● Models and software will be usable by third parties after the end of the project ● "Datasheets for datasets" (if not available) and "Model cards" will be created following the datasheets for datasets template (Gebru et al, 2018) and model cards template (Mitchel et al, 2019).
4. Allocation of resources	**Costs for making data FAIR: 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, staff time, etc.)?** N/A **Reimbursement of the costs: 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).** N/A **Long-term preservation: 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)?** N/A
5. Data security	**Security measures: 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)?** N/A **Repositories policies and procedures: Will the data be safely stored in trusted repositories for long-term preservation and curation?** N/A
6. Legal and ethical requirements	**Ethics or legal issues: Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** As the datasets refer to publicly available data that have been initially collected outside the context of the project, further data processing, for the research purposes of MAMMOth, is considered compatible with the initial purposes of the data collection. **Informed consent: Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** N/A
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)?** N/A

3.2 Datasets related to the three use cases (USEC)

At the time of writing this deliverable three datasets are reported in relation to the three use cases of the project.

Table 7. FaceVerif dataset

DMP component	MAMMOth_005_USEC_FaceVerif_V1 Partner: IDnow
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** We will reuse existing data collected from the industrial flow of IDnow for research purposes, if the user has given its explicit consent. The initial purpose of the data processing is considered compatible with the further data processing in the context of MAMMOth, according to IDnow's organisational protocol for obtaining permission for further use of a user's data for other purposes. **Type/format: What types and formats of data will the project generate or re-use?** jpeg images: selfies and ID photos extracted from ID documents. **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** The purpose of the data use is the evaluation of the methods developed during MAMMOth to prevent, mitigate and solve biases in AI, on the specific use case of face verification. It will contribute mainly to the specific objective O8 of the project: *Study AI biases in case-by-case basis – providing insights on high risk applications: demonstrate the developed methods and extract insights of bias, in finance, face recognition and research.* **Expected size: What is the expected size of the data that you intend to generate or re-use?** Training data: The data will be grouped by subject: one ID photo + 1 to 4 selfie(s) / subject. Data from ~10 000 subjects will be used. Test data: It will consist of pairs of ID photo/selfie. Data from ~10 000 subjects will be used. **Data origin: What is the origin/provenance of the data, either generated or re-used?** The data will be extracted from IDnow's database of ID documents and face images coming from IDnow's industrial flow. **Data utility: To whom might your data be useful, outside your project?** This data include biometric information and therefore are classified as personal data of "special categories" under GDPR, Article 9. The processing of this highly sensitive data may entail higher ethical risks. Therefore, it will not be made available outside of IDnow.
2. FAIR data **2.1 Making data findable, including provisions for metadata**	This Section is not applicable since we cannot share our data outside of the company. **Findable data: Will data be identified by a persistent identifier?** N/A **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A

	<u>Findable metadata:</u> Will metadata be offered in such a way that it can be harvested and indexed? N/A
2.2 Making data openly accessible	**<u>Repository:</u>** **Will the data be deposited in a trusted repository?** No deposit on a repository due to the GDPR provisions (no sharing of personal data). **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** N/A **<u>Data:</u>** **Will all data be made openly available?** The dataset consists of images of faces. This data is considered as personal data (belonging to "special category" of data) by the GDPR, Article 9. The sharing of this data without anonymisation or pseudonymisation is therefore forbidden. Considering the nature of this data, it is inherently identifiable and therefore it is not possible to anonymise or pseudonymise it. **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** The data will only be used by IDnow's authorised employees through a role-based access. **How will the identity of the person accessing the data be ascertained?** Only a restricted number of IDnow employees have access to the data. They can only access it through a personal account with a two-factor authentication. **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** No **<u>Metadata:</u>** **Will metadata be made openly available and licensed 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?** The data will not be made available, therefore no metadata will be shared either. **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** The data will not be made available and findable. **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)?** N/A

2.3 Making data interoperable	**Interoperability: 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?** N/A (the data will remain confidential) **Use of vocabularies: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?** N/A **Qualified references: Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**Documentation: 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.)?** The documentation will only be internal to IDnow. **License: 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?** No **Usable by third parties after the end of project: Will the data produced in the project be useable by third parties, in particular after the end of the project?** No, because the data cannot be shared outside of IDnow. **Data provenance: Will the provenance of the data be thoroughly documented using the appropriate standards?** No (no data sharing) **Quality control measures: Describe all relevant data quality assurance processes.** It should be mentioned that the data quality constitutes one of the challenges of the relevant use case of the project. Mitigation of internal quality assurance processes are governed by mitigation strategy including both manual and automated processes on the data.
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** Yes. We will reuse our current face verification algorithms, and potentially generate new algorithms or newly trained algorithms free of biases. **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** The algorithms used in the frame of MAMMOth belongs to IDnow and will remain confidential (no disclosure outside of IDnow).

4. Allocation of resources	**Costs for making data FAIR: 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, staff time, etc.)?** No costs are planned since the data will remain confidential. **Reimbursement of the costs: 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).** N/A **Long-term preservation: 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)?** The data are preserved for 10 years after their collection or until the withdrawal of the consent (whichever comes earlier). The identification of the data to preserve and the duration or the preservation are decided by the research team in agreement with the legal team, which sets the legal framework for this preservation.
5. Data security	**Security measures: 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)?** IDnow deploys several security measures to ensure the protection of the personal data collected for research purposes. Personal and sensitive data collected by IDnow is encrypted and stored on a completely Internet-tight platform for a fixed duration of 10 years. The encryption keys are owned only by the four persons in charge of the data storage management. The storage is performed in IDnow premises. Only a limited number of persons have access to the storage room, mainly for maintenance. The system administrators have no direct access to the data due to the encryption. The access to the data is restricted to the people who actually process it for research purposes (researchers, engineers, and technicians). In addition to these security measures, all employees follow a yearly training about GDPR. **Repositories policies and procedures: Will the data be safely stored in trusted repositories for long-term preservation and curation?** The data are only stored on IDnow's premises to avoid any leakage of the data, on hard drives designed for long term preservation. Data are duplicated (1 to 1) and backed up each week (encrypted back-up).
6. Legal and ethical requirements	**Ethics or legal issues: Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** Yes, the data in this dataset is personal and cannot be shared for legal reasons (GDPR). **Informed consent: Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** The data is used for research purposes and will be processed in the frame of MAMMOth only if the user has given its explicit consent that his/her data can be used for the purpose of improving IDnow's algorithms. They are informed of the data processing as well as the rights they have from a pop-up window which refers via a hypertext link to the privacy and personal data processing policy. In this policy, the contact information of IDnow's DPO is explicitly given so that the users can exercise their rights.

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)?** No

Table 8. Debt collection dataset

DMP component	**MAMMOth_006_USEC_DebtCollection_V1** **Partner: EXUS**
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** We will use historical datasets of customers for the training of AI models. Time period, variables and features that will be used will be defined based on the project's needs. **Type/format: What types and formats of data will the project generate or re-use?** The initial format of the data will be csv files, which are afterwards loaded into tables in a PostgreSQL database. **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** Identify, possible sources or cases of data biases. **Expected size: What is the expected size of the data that you intend to generate or re-use?** The data are up to tens of gigabytes in size (~ 40-50 GB), depending on the years that the clients use EXUS Financial Suite product. **Data origin: What is the origin/provenance of the data, either generated or re-used?** The data are generated by the EXUS EFS software, which tracks customer payments, customer days passed due (how many days have passed since the payment had to be made), agent-customer interactions, etc. **Data utility: To whom might your data be useful, outside your project?** The data will not leave EXUS premises and will not be shared among the consortium partners.
2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** No **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A **Findable metadata: Will metadata be offered in such a way that it can be harvested and indexed?** N/A

2.2 Making data openly accessible	**<u>Repository:</u>** **Will the data be deposited in a trusted repository?** N/A **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** N/A **<u>Data:</u>** **Will all data be made openly available?** No, they will not be publicly available. **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** Role-base access on these data will be available only for EXUS members. **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A **<u>Metadata:</u>** **Will metadata be made openly available and licensed 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?** N/A **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** N/A **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)?** No
2.3 Making data interoperable	**<u>Interoperability:</u> 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?** N/A **<u>Use of vocabularies:</u> 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? N/A **Qualified references: Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**Documentation: 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.)?** N/A **License: 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?** No **Usable by third parties after the end of project: Will the data produced in the project be useable by third parties, in particular after the end of the project?** N/A **Data provenance: Will the provenance of the data be thoroughly documented using the appropriate standards?** N/A **Quality control measures: Describe all relevant data quality assurance processes.** N/A
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** Results from using the MAMMOth algorithms from the data will be published in scientific papers. **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** The algorithms used in the frame of MAMMOth belong to EXUS and will remain confidential (no disclosure outside of EXUS).

4. Allocation of resources	**Costs for making data FAIR: 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, staff time, etc.)?** No costs are planned since the data will remain confidential. **Reimbursement of the costs: 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).** N/A **Long-term preservation: 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)?** N/A
5. Data security	**Security measures: 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)?** Data are pseudonymised and all Azure cloud security guidelines are followed. **Repositories policies and procedures: Will the data be safely stored in trusted repositories for long-term preservation and curation?** N/A
6. Legal and ethical requirements	**Ethics or legal issues: Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** Data will not be shared. **Informed consent: Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** Data will be processed based on the privacy policy (data processing purpose and time period of the processing are indicated) signed by the clients of the dataset providers (EXUS clients); no further questionnaires will be used.
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)?** No

Table 9. Academic networks dataset

DMP component	**MAMMOth_007_USEC_AcademicNetworks_V1** **Partner: CSH**
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** We will re-use publicly available datasets from Semantic Scholar Open Research Corpus with papers published for 19 different fields, and Elsevier datasets for geographical information of the authors. **Type/format: What types and formats of data will the project generate or re-use?** Paper information and researchers' institution of affiliation (json and csv format). The

	project will convert all data to csv and format to the data in a standardised way to assist experimentation. **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** Bias assessment in academic ranking to understand how underrepresented communities in science are affected when using algorithms personalisation and online research engines, and to test proposed strategies for the mitigation of identified biases. These are related to O1, O2, O2, O3, O8 specific objectives of the project. **Expected size: What is the expected size of the data that you intend to generate or re-use?** The dataset has information for around 200 million papers taking around 1TB. **Data origin: What is the origin/provenance of the data, either generated or re-used?** ● Semantic Scholar open research corpus ● Scopus/Elsevier API **Data utility: To whom might your data be useful, outside your project?** N/A (these data are publicly available)
2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** ● Semantic scholar open research corpus: https://www.semanticscholar.org/product/api ● Elsevier/Scopus: https://dev.elsevier.com/ **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A **Findable metadata: Will metadata be offered in such a way that it can be harvested and indexed?** N/A
2.2 Making data openly accessible	**Repository:** **Will the data be deposited in a trusted repository?** We expected to deposit the enriched dataset in the Open Science framework: https://osf.io/ **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **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?** N/A **Will the data be accessible through a free and standardised access protocol?**

	N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** N/A **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A **Metadata:** **Will metadata be made openly available and licensed 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?** N/A **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** N/A **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)?** Descriptions of how the dataset was collected and enriched (GitHub URL).
2.3 Making data interoperable	**Interoperability: 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?** N/A **Use of vocabularies: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?** N/A **Qualified references: Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**Documentation: 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.)?** We will provide readme files, variable definitions, and units of measurement. **License: 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?** Yes

	<u>Usable by third parties after the end of project:</u> Will the data produced in the project be useable by third parties, in particular after the end of the project? Yes **<u>Data provenance:</u> Will the provenance of the data be thoroughly documented using the appropriate standards?** Yes **<u>Quality control measures:</u> Describe all relevant data quality assurance processes.** N/A
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** N/A **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** 1. Descriptions of developed algorithms will have an identifier (GitHub URL) 2. Algorithms and software will be publicly available through GitHub 3. Documentation will be provided in the form of readme files 4. The license under which our research outputs will be is Creative Commons 5. Developed algorithms will be shared under Apache License 2.0 6. Algorithms and software will be usable by third parties after the end of the project
4. Allocation of resources	**<u>Costs for making data FAIR:</u> 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, staff time, etc.)?** N/A **<u>Reimbursement of the costs:</u> 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).** N/A **<u>Long-term preservation:</u> 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)?** N/A
5. Data security	**<u>Security measures:</u> 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)?** N/A **<u>Repositories policies and procedures:</u> Will the data be safely stored in trusted repositories for long-term preservation and curation?** We expected to deposit the enriched dataset in the Open Science Framework: https://osf.io/

6. Legal and ethical requirements	**Ethics or legal issues: Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** As the datasets refer to publicly available data, that have been initially collected outside the context of the project. Further data processing, for the research purposes of MAMMOth, is considered compatible with the initial purposes of the data collection. **Informed consent: Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** N/A
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)?** No

3.3 Datasets related to communication, dissemination and exploitation (DISEX)

In the scope of the communication, dissemination, and exploitation activities of MAMMOth, the following dataset has been identified.

Table 10. Social media content dataset

DMP component	MAMMOth_008_DISEX_SocialMediaContent_V1 Partner: CSI
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** image (jpeg, png formats) & Video (mov, mp4 formats) **Type/format: What types and formats of data will the project generate or re-use?** png & mp4 **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** Social Media Content Distribution and Brand Awareness in all digital media platforms. **Expected size: What is the expected size of the data that you intend to generate or re-use?** Maximum size is 10MB and uploaded online. **Data origin: What is the origin/provenance of the data, either generated or re-used?** All relevant content is created from an online design tool provided by https://www.canva.com/. **Data utility: To whom might your data be useful, outside your project?** Useful to MAMMOth community and social media audience.

2. FAIR data **2.1 Making data findable, including provisions for metadata**	**<u>Findable data:</u> Will data be identified by a persistent identifier?** N/A **<u>Metadata creation</u>: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **<u>Search keywords</u>: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?** N/A **<u>Findable metadata:</u> Will metadata be offered in such a way that it can be harvested and indexed?** N/A
2.2 Making data openly accessible	**<u>Repository:</u>** **Will the data be deposited in a trusted repository?** Yes **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** Yes **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** Yes **<u>Data:</u>** **Will all data be made openly available?** Social Media Content can be found on all social media platforms **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** N/A **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A **<u>Metadata:</u>** **Will metadata be made openly available and licensed 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?** N/A **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** N/A **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)? N/A
2.3 Making data interoperable	**Interoperability: 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?** N/A **Use of vocabularies: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?** N/A **Qualified references: Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**Documentation: 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.)?** N/A **License: 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?** N/A **Usable by third parties after the end of project: Will the data produced in the project be useable by third parties, in particular after the end of the project?** We do not share user generated content in our social media content. **Data provenance: Will the provenance of the data be thoroughly documented using the appropriate standards?** N/A **Quality control measures: Describe all relevant data quality assurance processes.** N/A
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** N/A **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?** N/A

4. Allocation of resources	**Costs for making data FAIR: 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, staff time, etc.)?** N/A **Reimbursement of the costs: 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).** N/A **Long-term preservation: 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)?** N/A
5. Data security	**Security measures: 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)?** The following security measures for data recovery are in place: https://eu.siteground.com/blog/siteground-is-gdpr-compliant/#Processing_of_the_data_uploaded_on_your_account SiteGround has implemented and maintains appropriate technical and organisational security measures to protect personal data against accidental or unauthorised loss, destruction, alteration, disclosure or access, and against all other unlawful forms of processing. The security measures are in accordance with the Standard Contractual Clauses and include provisions for personnel and confidentiality, physical security, system access control, services access control, transmission control, input control, network control, incident response, system logs, reliability and backup, data destruction, and subprocessor security. SiteGround regularly tests its disaster recovery plans, maintains security incident management policies and procedures, and monitors a variety of communication channels for security incidents. Before onboarding Sub-processors, SiteGround conducts due diligence of the security and privacy practices of Sub-processors to ensure they provide a level of security and privacy appropriate to their access to data and the scope of the services they are engaged to provide. SiteGround also enhances and implements changes in the Services during the term of the Agreement, while providing security controls that deliver a level of security protection that is not materially lower than that provided as of the Effective Date. (Source: https://eu.siteground.com/viewtos/data_processing_agreement?scid=3&lang=en) **Repositories policies and procedures: Will the data be safely stored in trusted repositories for long-term preservation and curation?** We are not collecting data so N/A.
6. Legal and ethical requirements	**Ethics or legal issues: Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** We do not collect data so N/A. And if we repost any content we state the source. **Informed consent: Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?** Yes

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)?** No

3.4 Datasets related to project management (MGT)

In the scope of the management of the project, the following dataset has been identified.

Table 11. Consortium admin dataset

DMP component	MAMMOth_009_MGT_ConsortiumAdminData_V1 Partner: CERTH
1. Data Summary	**Re-use of existing data: Will you re-use any existing data and what will you re-use it for?** No. Data in the context of project management refer exclusively to the MAMMOth project and its consortium. **Type/format: What types and formats of data will the project generate or re-use?** The collected data consist, in terms of data type, of spreadsheets, documents and in terms of their format, of xls, doc, pptx, txt and pdf. **Purpose: What is the purpose of the data generation or re-use and its relation to the objectives of the project?** This dataset includes business contact information of the MAMMOth consortium members, including names, organisation, emails, office postal addresses, office phone numbers, financial data and any other related administrative data that is considered necessary and proportionate for the management of the project and more specifically for the fulfilment of the tasks of *WP6-Project Management*. **Expected size: What is the expected size of the data that you intend to generate or re-use?** <500 MB **Data origin: What is the origin/provenance of the data, either generated or re-used?** The data were collected from the MAMMOth partners by CERTH via email communications or via Google Docs Editors suite applications. **Data utility: To whom might your data be useful, outside your project?** In general, the data of the WP6 will be processed by the project's partners. A part of that data might be processed solely by CERTH (project coordinator) and some exclusively by the coordinator and the European Commission.
2. FAIR data **2.1 Making data findable, including provisions for metadata**	**Findable data: Will data be identified by a persistent identifier?** No. This dataset is not and is not going to be findable from external parts. **Metadata creation: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?** N/A **Search keywords: Will search keywords be provided in the metadata to optimise the**

	possibility for discovery and then potential re-use? N/A **<u>Findable metadata:</u> Will metadata be offered in such a way that it can be harvested and indexed?** N/A
2.2 Making data openly accessible	**<u>Repository:</u>** **Will the data be deposited in a trusted repository?** No. This dataset does not include digital research data that have been generated in the project and, in line with the FAIR principles, must be deposited in a trusted repository, according to Article 17 (Annex 5) of the Annotated Grant Agreement. **Have you explored appropriate arrangements with the identified repository where your data will be deposited?** N/A **Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?** N/A **<u>Data:</u>** **Will all data be made openly available?** No. This dataset does not include digital research data that have been generated in the project and, in line with the FAIR principles, should be openly available, according to Article 17 (Annex 5) of the Annotated Grant Agreement. **Will the data be accessible through a free and standardised access protocol?** N/A **If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?** N/A **How will the identity of the person accessing the data be ascertained?** N/A **Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?** N/A **<u>Metadata:</u>** **Will metadata be made openly available and licensed 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?** No. This dataset does not include digital research data, whose metadata must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles, according to Article 17 (Annex 5) of the Annotated Grant Agreement. **How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?** N/A

	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)? N/A
2.3 Making data interoperable	**<u>Interoperability:</u> 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?** This dataset does not include digital research data that have been generated in the project and, in line with the FAIR principles, should be interoperable. **<u>Use of vocabularies:</u> 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?** N/A **<u>Qualified references:</u> Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?** N/A
2.4 Increase data re-use	**<u>Documentation:</u> 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.)?** No. This dataset is not going to be discoverable and shared. **<u>License:</u> 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?** N/A **<u>Usable by third parties after the end of project:</u> Will the data produced in the project be useable by third parties, in particular after the end of the project?** N/A **<u>Data provenance:</u> Will the provenance of the data be thoroughly documented using the appropriate standards?** N/A **<u>Quality control measures:</u> Describe all relevant data quality assurance processes.** N/A
3. Other research outputs	**Are there any other research outputs that may be generated or re-used throughout the project?** There are no other research outputs in the context of the project management dataset. **Are there any questions pertaining to FAIR data section above, that can apply to the management of the other research outputs? Will you provide sufficient detail on how your research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles?**

	N/A
4. Allocation of resources	**Costs for making data FAIR: 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, staff time, etc.)?** N/A **Reimbursement of the costs: 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).** N/A **Long-term preservation: 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)?** N/A
5. Data security	**Security measures: 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)?** As the data security (confidentiality, integrity and availability) level, according to Article 5 par.1 (f) GDPR, is based on the defined risks for the data subjects (in case of unauthorised access or disclosure, accidental deletion or destruction of the data), the security measures for the project management datasets will be proportionate to the dealing risks. The data of the project management is stored in the project's file repository based on Google Drive. The access is given, after user login, only to the relevant staff (confidentiality). Moreover, regular backups are scheduled, in the context of WP6, to minimise the risks of the impacts of deletion or destruction of the data (integrity). In parallel, state-of-the-art IT security measures and company policies mitigate most of the risk of illegitimate access, contributing to the ability to detect promptly an incident and thus restore the data (availability). Sensitive personal data (Article 9 par. 1 GDPR) are not included in the project management dataset. **Repositories policies and procedures: Will the data be safely stored in trusted repositories for long-term preservation and curation?** This dataset is stored in the project's file repository based on Google Drive, where stored data is encrypted in-transit and at-rest (according to the application's provided information).
6. Legal and ethical requirements	**Ethics or legal issues: Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?** This dataset, which constitutes administrative data, does not raise issues, related to ethics requirements. The dataset is not going to be shared with external parties. Initially, the internal Legal, Ethics and Data Compliance Protocol (as an integral part of the T6.3), which is complementary to the DMP, addresses the general approach of the project regarding ethical and data protection issues. **Informed consent: Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?**

	N/A
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)?** No

4 Conclusions

The aim of the MAMMOth DMP has been addressed by recording information about the datasets created so far that have either been generated or re-used by the project. This document also outlines the strategy and methodology for data management in MAMMOth. The implementation of the data management plan is governed by a collaborative and comprehensive approach across the project partners.

MAMMOth datasets have been classified into four categories based on their relation to the project type activity. These are datasets for the MAMMOth toolkit and its research components, datasets for the three use cases outlined in the project GA, datasets for communication, dissemination and exploitation, and datasets for the project management of MAMMOth. Nine datasets have been identified in total. However, it should be mentioned that public datasets that are re-used in MAMMOth have been grouped where possible in order to present them in a more concise way.

Furthermore, this DMP will be updated, at least mid-project and at the end of the project, and will be available to the MAMMOth consortium in order to accompany the project's evolution.

5 References

1. Article 29 Working Party. (2018) *Guidelines on transparency under regulation 2016/679. WP260*.
2. ERC Scientific Council. (2022). *Open Research Data and Data Management Plans: (V4.1).*
3. European Commission. (2019). *Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union*. Brussels.
4. European Commission. (2021). *AGA – Annotated Model Grant Agreement: (V0.2).*
5. European Commission. (2021). *Data Management Plan Template: (V1.0).*
6. European Commission. (2022). *HE Programme Guide: (V2.0).*
7. Jahn, N., Laakso, M., Lazzeri, E., & McQuilton, P. (2023). Study on the readiness of research data and literature repositories to facilitate compliance with the Open Science Horizon Europe MGA requirements (V1.0). Zenodo. https://doi.org/10.5281/ZENODO.7728015.
8. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text With EEA relevance)*.
9. Rizou, S., Alexandropoulou-Egyptiadou, E., & Psannis, K. E. (2020). GDPR interference with next generation 5G and IoT networks. *IEEE Access*, *8*, 108052-108061.
10. Science Europe. (2021). *Practical Guide to the International Alignment of Research Data Management - Extended Edition*.
11. Wiewiorówski, W. (2020). A preliminary opinion on data protection and scientific research. *Brussels, Belgium: European Data Protection Supervisor*.
12. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9.
13. Wolters, P. T. J. (2017). The security of personal data under the GDPR: a harmonized duty or a shared responsibility?. *International Data Privacy Law*, *7*(3), 165-178.

MAMMOth is a Horizon Europe Research and Innovation Project co-funded by the European Union under Grant Agreement ID: 101070285 and an UK Research and Innovation grant 10041914.