D7.2 Data Management plan

Deliverable information	
WP	WP7
Document dissemination level	CO Confidential, only for members of the consortium
Deliverable type	R Document, report
Lead	BTF
Contributors	All partners
Document status	Draft under review
Document version	V0.2
Date	04/10/2021

Document History

Version	Release date	Summary of changes	Partner
V0.1	01/09/2021	First draft released	BTF
V0.2	04/10/2021	reviewed	BTF, UCM

Project information

Project start date: 1st of March 2021

Project Duration: 36 months

Project website: https://isee4xai.com/

iSee consortium

UCM	UNIVERSIDAD COMPLUTENSE DE MADRID	SPAIN
RGU	ROBERT GORDON UNIVERSITY	SCOTLAND
BT	BT FRANCE	FRANCE
UCC	UNIVERSITY COLLEGE CORK	IRELAND

Contact: hello@isee4xai.com

Project Coordinator:

Professor M Belen Díaz Agudo
Instituto de Tecnología del Conocimiento
Facultad de Informática,
Universidad Complutense de Madrid
C/ Profesor Jose Garcia Santesmases, 9
Ciudad Universitaria
28040 – Madrid, Spain

Summary

The document details the *Data Management Plan* (DMP) for the iSee project. It describes how the different data produced or consumed by the project are handled during the entire lifetime of the project. The scope of this document is to provide a reference document describing how data assets are organized, secured, made available, transported and protected, from their generation to destruction.

This *Data Management Plan* seeks at addressing the following two objectives:

1. This document provides a comprehensive overview of all data used, collected or generated in the project and a characteristization of the data. It ensures that the appropriate governance policies are in place to meet legal, security, regulatory or ethical requirements.

2. Most European research funding agencies (such as the ANR[1] in France) require all projects funded in 2019 onwards to produce a Data Management Plan (DMP). This is intended to support European and international alignment efforts on the structure of open research data, and is guided by the principle: “as open as possible, as closed as necessary”.

The data management plan evolves along with the project with continuous adaptations as the data and project change.

[1] Agence National pour la Recherche (French national Research Agency)

Table of Contents

List of data sources used in this project

Multiple data are collected and produced during the development and through the lifetime of the iSee project. Some of these data may contain confidential or sensitive information as well: for instance, the data brought by the partners might contain intellectual property or personal information and therefore the overall management of these data must be carefully designed to respect privacy and regulatory requirements.

This document describes the management principles (governance, organization, location, access rights, security and compliance policies) of the following data used within the iSee project:

- **Project storage area** used by the iSee consortium members to share general information during the development of the project.
- **Repository for the source codes** of the different iSee software modules published under an open-source licence.
- **The iSee website** (http://www.isee4xai.com): general information about the iSee project with blogs, information about partners, use cases and contacts details.
- **Deliverable document in their final version** : The documents highlight the main progress and results at each milestone. They are not aimed at being public. The access to these documents should be under the control of iSee consortium. Documents should be stored in a centralised place though with adequate access rights.
- **Research Papers and published literature:** Academic and papers published by the iSee consortium members and made available to the XAI and AI community.

Project storage area[2]

Data collection How will new data be collected or produced and/or how will existing data be re-used? What data (for example the kind, formats, and volumes), will be collected or produced?	This area is a general repository used by the iSee consortium members to share general information during the development of the project, such as meeting minutes, working documents, internal presentations,... This area is delivered by a shared Google Drive service. Most documents will be stored using the Microsoft Office formats (xlsx, docx, pptx, …) and/or Google Office formats. Contents stored in this area are not meant to stay after the project is delivered and this area will be archived and deleted when no longer in use.

[2] https://drive.google.com/drive/folders/1Hcq7q_NEJgrXXu0-s-rPy5FjERONtfHA

Documentation and data quality What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data? What data quality control measures will be used?	Data is organized in a hierarchical structure with folders and sub-folders. At the root level, one folder is created for each work package (WP1 to WP7). Data is only available to the iSee consortium members (UCM, RGU, BT, UCC) and is not shared externally. Users are identified with their own personal or enterprise Google account.
Storage and backup during the research process How will data and metadata be stored and backed up during the research? How will data security and protection of sensitive data be taken care during the research	No metadata is used in this service, and no personal information is kept in this storage area. Google delivers this service as a fully managed service and is responsible for ensuring the data is backed up and secure. However, no SLA is provided for this service. This is an accepted risk as a) no confidential data and b) no critical information are stored on this area.
Legal and ethical requirements, code of conduct If personal data are processed, how will compliance with legislation on personal data and on security be ensured? How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable? What ethical issues and codes of conduct are there, and how will they be taken into account?	Not applicable as no personal or confidential data is stored in this area.

Data sharing and long-term preservation How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons? How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)? What methods or software tools are needed to access and use data?	Not applicable as data is not shared outside the iSee consortium members.
Data management responsibilities and resources Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)? What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?	This area is a self-managed service offered by Google and its management responsibility is shared across all iSee consortium members.

Repository for the source codes[3]

Data collection How will new data be collected or produced and/or how will existing data be re-used? What data (for example the kind, formats, and volumes), will be collected or produced?	The repository for the source codes is the storage area used by the version control system. This system is responsible for managing changes to the source codes of the different iSee components. Data on this repository can be accessed by a source code management software (such as Github for instance) or through a web portal. Data can also be made available to anybody (public repository) or restricted to the iSee consortium members (private repository).

[3] https://github.com/isee4xai

Documentation and data quality What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data? What data quality control measures will be used?	The following metadata are tracked by the system: ● date, user/owner of the change ● files impacted by the change ● name or number of the release ● other tags that can be added by users.
Storage and backup during the research process How will data and metadata be stored and backed up during the research? How will data security and protection of sensitive data be taken care during the research	This service is offered as a SaaS (software-as-a-service) offering by Github.com. As this is delivered as a managed service, Github.com is responsible for backup and ensuring data is protected from unauthorized access.
Legal and ethical requirements, code of conduct If personal data are processed, how will compliance with legislation on personal data and on security be ensured? How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable? What ethical issues and codes of conduct are there, and how will they be taken into account?	Not applicable as no personal or confidential data is stored in this area.

Data sharing and long-term preservation How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons? How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)? What methods or software tools are needed to access and use data?	Data stored on public repositories are available to everyone without restrictions. *Note that Github.com drives different initiatives to ensure long-term availability of the public repositories. One of these initiatives involves storing public repositories in the GitHub Arctic Code Vault at the Arctic World Archive in Svalbard. The GitHub Arctic Code Vault stores the latest revision of the project's default branch (potentially excluding large binary files depending on the overall size of the repository) at the time of capture. This "cold storage" is designed to last for 1,000 years.*
Data management responsibilities and resources Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)? What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?	The management responsibility for public repositories is shared across all iSee consortium members.

Research Papers and published literature[4]

Data collection How will new data be collected or produced and/or how will existing data be re-used? What data (for example the kind, formats, and volumes), will be collected or produced?	In order to encourage the dissemination of research and academic knowledge about Explainable AI, the iSee project members embrace the principles of Open Science: open source, open data, open access. All publications, as well as their related datasets and digital assets, will be published and referenced on Zenodo (https://zenodo.org/) . A copy of the articles will be available on the iSee website for reference, depending on the journal or conference copyrights clearance.

[4] https://isee4xai.com/project/
Zenodo.org account : username = isee4xai , email = iseechistera@gmail.com, password is the same as mailbox

Documentation and data quality What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data? What data quality control measures will be used?	The following metadata are used for each publication: • Author(s) name • Date of Publication • DOI (Digital Object Identifier) All digital artefacts related to papers will also be published and identified (with a DOI) and will be kept alongside the publications to ensure research can be fully understood and reproduced.
Storage and backup during the research process How will data and metadata be stored and backed up during the research? How will data security and protection of sensitive data be taken care during the research	All data stored on Zenodo are hosted on CERN's EOS service , a platform for storing large amounts of physics data and user files used at the Large Hadron Collider (LHC). Multiple copies of the data are generated and a tape storage service is used to ensure long-term data retention and security. Metadata and persistent identifiers in Zenodo are stored in a database operated by the CERN' with 12-hourly backup cycle with one backup sent to tape storage once a week. Note that no sensitive or confidential information will be part of the data shared on this service.
Legal and ethical requirements, code of conduct If personal data are processed, how will compliance with legislation on personal data and on security be ensured? How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable? What ethical issues and codes of conduct are there, and how will they be taken into account?	Not applicable as no personal or confidential data is stored in this area.

Data sharing and long-term preservation How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons? How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)? What methods or software tools are needed to access and use data?	Zenodo states that "*all stored data "will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least."* In case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories.
Data management responsibilities and resources Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)? What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?	Zenodo is responsible for all data management activities.

ANNEX

General information about data used in the iSee project

Name	Description	Location / link	Intended audience & access rights	How data is protected (backup)	Contains Personal data[5] ?
iSee Google Drive -	Online drive used by the iSee consortium members to share general information during the development of the project.	Link	Access restricted to iSee consortium members NO confidential / sensitive / personal data	Built-in data Protection service from Google.	No
Github repository **(public & private repositories).**	Main source code repository for the iSee software modules published under an open-source licence.	Link	Read access : no restriction Write access: iSee consortium members	Built-in data protection service from github.com	No
iSee Website (general information)	iSee project main website - general information about the iSee project with blogs, information about partners, use cases and contacts	Link	Read access: no restriction Publication rights: iSee consortium members	OVH web cloud backup policy	No

[5] 'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Name	Description	Location / link	Intended audience & access rights	How data is protected (backup)	Contains Personal data[6] ?
iSee XAI modules	Webserver delivering the following XAI modules and components: (individual component is to be detailed in year 2 version) ● iSee IDT (Interactive Design Tool) ● Isee CE² (Cockpit Evaluation Environment ● Case-based reasoning engine (iSee Onto) ● User analytics API ● XAI engine API ● Interaction engine ● iSee Cockpit Evaluation forms (?) ● Datasets from partners (when applicable) ● AI Models from partners (when applicable) Front-end - servers hosted in BT DCs. Backend - K8s/VM on servers hosted in BT data-centres. Data layer - servers hosted in BT DCs(?) or AWS (?)	*tbd*	*Read access : iSee consortium* *Write access : iSee consortium and platform administrators*	*No data about AI model or confidential datasets from partners will be stored in iSee. Cockpit and Case onto engine will save data on the iSee server. Data of cockpit will be questions/responses in an aggregated format*	*Yes and data may not be anonymisable.*

[6] 'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

LinkedIn Group	LinkedIn used to create, grow and manage an on-line community for professionals with interests in XAI to share experience, insights and build connections.	Link	Available to all LinkedIn users. New members must be invited to join the group by existing members.	Builtin LinkedIn data protection service	Yes - privacy policy managed by LinkedIn (see policy here)
Research Papers and published literature	Academic and research papers published by the iSee consortium members and made available to the XAI and AI community	*Zenodo - Research. Shared.*	*Stored on Zenodo.org from the isee4xai user profile*	*Builtin Zenodo data management service*	*No*
iSee general deliverables	Project results and progress main document built by iSee consortium members. Made available on CHIST-ERA website when open access, and on Zenodo website (access control) + A reference to our own website	*Zenodo - Research. Shared.*	Read access: no restriction Publication rights: iSee consortium members	*Builtin Zenodo data management service*	No