Federated decentralized trusted dAta Marketplace for Embedded finance

D1.2 - Data Management Plan

Title	D1.2 - Data Management Plan
Revision Number	3.0
Task reference	T1.2
Lead Beneficiary	UPRC
Responsible	margy@unipi.gr; kiourtis@unipi.gr
Partners	AL, DAEM, GFT, IDSA, INNOV, INSO, KM, LXS, MOH, NUIG, UBI, UNP
Deliverable Type	R
Dissemination Level	PU
Due Date	2023-06-30 [Month 6]
Delivered Date	2023-06-29
Internal Reviewers	GFT, KM
Quality Assurance	UPRC
Acceptance	Coordinator Accepted
Project Title	FAME - Federated decentralized trusted dAta Marketplace for Embedded finance
Grant Agreement No.	101092639
EC Project Officer	Stefano Bertolo
Programme	HORIZON-CL4-2022-DATA-01-04

This project has received funding from the European Union's Horizon research and innovation programme under Grant Agreement no 101092639

Revision History

Version	Date	Partners	Description
0.1	2023-04-04	UPRC	TOC
0.2	2023-04-04	UPRC	Initial content
1.0	2023-06-20	UPRC	Version ready for internal review
1.1	2023-06-28	GFT, KM	Reviewed version
2.0	2023-06-28	UPRC	Final Version ready for QA
3.0	2023-06-29	UPRC	Final Version ready for submission

Definitions

Acronyms	Definition
4AML	4th Anti Money Laundering Directive
AAI	authentication authorization infrastructure
AI	Artificial Intelligence
BNPL	Buy Now Pay Later
CSV	Comma Separated Value files
DB	Data Base
DOI	Digital object identifier
DPO	Data Protection Officer
EC	European Commission
EE	Energy Efficient
ESG	Environmental, social, and corporate governance
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
FIBO	Financial Industry Business Ontology
FIGI	Financial Instrument Global Identifier
GA	Grant Agreement|General Assemply
GDPR	General Data Protection Regulation
IPR	Intellectual Property Right
IT	Information Technology
IoT	Internet of Things
JSON	JavaScript Object Notation
LC	Learning Centre
ML	Machine Learning
N/A	Not Available / Not Applicable
NDA	Non-Disclosure Agreement
PDDL	Public Domain Dedication and License
PDF	Portable Document Format (for Adobe Acrobat Reader)
PSD2	Second Payment Service Directive
PSR	Project Security Responsible
RDF	Resource Description Framework
SAX	Situation Aware Explainability
SQL	Structured Query Language
SUB	Subroutine
TB	Terabyte (1,000 gigabytes)

TBD	To Be Determined
TXT	Text|TeXT file
URL	Uniform Resource Locator
VDIH	Virtual Digital Innovation Hub
WP	Workpackage
XAI	eXplainable Artificial Intelligence

Executive Summary

The current deliverable (D1.2) defines and describes the overall Data Management Plan (DMP) of the FAME project. The DMP seeks to identify the best practices and specific standards for the generated data and assess their suitability for sharing and reuse in accordance with official guidelines. To this end, it aims to support the data management lifecycle for all data that will be collected, processed, or generated by the project in order to maximize its access.

FAME project's DMP will comply with the European Commission's (EC) Data Management Plan template [1], as it was refined for Horizon Europe, and will specify how the generated data will be easily discovered and accessed, ensuring open access and interoperability. It reflects the status of the data that is collected, processed, or generated, their respective data methodology and standards, whether and how these data will be shared and made open and how they will be curated and preserved.

This document also presents the available data assets and datasets to be shared within FAME project by its pilots and the software artefacts utilized by the technical partners, as well as the tools and procedures used to manage data according to policies and regulations. This is the first version of FAME's DMP produced, with a concrete final version becoming available on the last month of the project (M36).

Table of Contents

List of Tables

1 Introduction

A Data Management Plan (DMP) in European Union (EU) projects is developed to describe how the data will be gathered, processed and/or generated, as well as how it will be stored. It is an essential component of effective data management as it systematically describes how datasets will be used by project partners and what data assets will be generated within the scope of the project. This will encourage the project partners to make data management decisions based on FAME project and FAIR principles [2], while also facilitating the smooth operation of the project.

FAME project develops a secure federated data marketplace for embedded finance (EmFi), aiming to demonstrate the full potential of the data economy, and in this context the current deliverable focuses on the management of the data collected, processed, or generated within the project and its lifecycle by all its consortium partners. The DMP aims to monitor the generated data regarding their privacy and confidentiality, ensuring that the legal and ethical standards for data generation, use, storage, and sharing are met throughout the project lifecycle and in line with the overall management of FAME, as also foreseen in the Grant Agreement (GA) and Consortium Agreement (CA).

Information to produce this deliverable was gathered by the FAME Consortium through a questionnaire (ANNEX 1) with a variety of questions per data management aspect and in respect to the FAIR data principles. The structure of the DMP is also based on European Commission's (EC) Data Management Plan template. The DMP reflects a current picture of the Project based on the answers from each partner and should be viewed as a living document that will be supplemented and refined as new issues and information come to light during the project's lifecycle.

1.1 Objective of the Deliverable

FAME's DMP, aims to identify and organize all data collected, processed or generated throughout the project's lifecycle, as well as specific standards used, data management policies and potential best practices showcased by the consortium partners.

1.2 Insights from other Tasks and Deliverables

D1.2 is inherently connected with D1.1 - FAME Project Manual, Risk and Quality Management Plan (FPRM), which is a reference document for effective and successful management of the FAME project. The FPRM procedural framework relies on the GA and the CA and is structured around project management knowledge areas (scope, plans, governance, monitoring, communication, risks, quality assurance) in addition to style guidelines and best practices.

Despite the broader scope and goal of D1.1, which is to provide guidance to consortium partners related to project management, risk assessment and quality control, various information are included regarding the proper management of data within the project, whether it is administrative data, deliverables, or consortium partners inputs, as well as for the standards used in the project.

Moreover, this document is also related to other Work Packages (WP) and in particular with the ones involved with the pilot Use Cases (WP6 [T6.2, T6.3, T6.4]) and the work that is being currently done within the pilot co-creation workshops in WP2.

1.3 Structure

The rest of the deliverable is structured as follows:

- Section 2 provides the methodology of this work and how the FAIR principles will be applied in the project;
- Section 3 provides an overview of the FAME project which is required to proceed with the introduction of FAME's data assets and data spaces
- Section 4 introduces the various data assets in FAME project and provides a detailed overview of the data assets provides by consortium members and pilot partners;
- Section 5 focuses on the FAIR principles in FAME;
- Sections 6 briefly replies to the questions related to the allocation of resources;
- Section 7 and Section 8 cover all the ethical and security aspects.

2 Methodology

2.1 FAIR management of research data

Data-driven projects like FAME face several challenges, one of which is facilitating knowledge sharing by using data. In 2016, the FAIR Principles for Scientific Data Management and Stewardship[3] were published as guidelines to enhance digital asset findability, accessibility, interoperability, and reusability. The purpose of this section is to introduce best practices for making data FAIR in FAME project.

Among other things, the FAIR principles provide for a continuum of increasing reusability across numerous implementations for data and digital objects (including, data-related algorithms, tools, workflows, protocols, services and other kinds of digital and research objects).[4] These solutions describe capabilities and features of systems and services that can be used to create valuable research outputs and have them evaluated and widely reused, with appropriate credit attributed to the creator and users.[5]

Well-described, accessible, and standard data are essential for finding relevant data, performing machine analysis, and employing artificial intelligence (AI). As a quality standard, the FAIR principles have been widely accepted by EU policymakers and have ignited global debates about data stewardship in open science and data-driven research.[6] Additionally, they have prompted funding agencies to discuss how they will implement their plans. Currently, some of these requirements are just beginning to be developed, while others are already a mature set of guidelines.

This approach generally precedes implementation choices and does not necessarily imply a particular set of technical requirements, standards, or solutions. However, they emphasize “machine actionability”, which refers to the ability of a computational system to find, access, interoperate, and reuse data without or with minimal human intervention. A summary of the FAIR guiding principles can be found in Table 1 below:

Table 1 FAIR guiding principles

Findable:	○ Meta(data) are assigned a globally unique and persistent identifier ○ Data are described with rich metadata ○ Metadata clearly and explicitly include the identifier of the data it describes ○ Meta(data) are registered or indexed in a searchable source
Accessible:	○ Meta(data) are retrievable by their identifier using a standardized communication protocol ○ The protocol is open, free, and universally implementable ○ The protocol allows for an authentication and authorization procedure, where necessary ○ Metadata are accessible, even when the data are no longer available
Interoperable:	○ Meta(data) uses a formal, accessible, shared

	and broadly applicable language for knowledge representation o Meta(data) use vocabularies that follow FAIR principles o Meta(data) includes qualified references to other meta(data)
Reusable:	o Meta(data) are richly described with a plurality of accurate and relevant attributes o Meta(data) are released with a clear and accessible data usage license o Meta(data) are associated with detailed provenance o Meta(data) meet domain-relevant community standards

Simply put, the principle of Findability requires that data must be identified, described, registered, or indexed clearly and unequivocally. A unique and persistent identifier should be assigned to the data; the main characteristics of the data should be specified systematically, preferably in a standard format, and these should be indexed and stored in a public repository, such as a data archive. According to the principle of Accessibility, datasets should be accessible via a clearly defined access procedure, preferably by automated means. It involves establishing authentication and authorization procedures for access as well as automating data retrieval protocols if necessary. Interoperability requires data and metadata to be conceptualized, expressed, and structured according to standards that are both common and publicly available. It involves the use of standard data formats, variables, ontologies, and so forth. According to the principle of Reusability, data characteristics and their provenance should be described in detail according to domain-relevant community standards, with clear and accessible usage guidelines. As part of this process, accurate and relevant data descriptions must be provided, as well as licenses for access and use, community standards, and provenance for each dataset.[7]

Data life cycle management should use these four principles, which are closely interrelated. When the FAIR principles are applied to the workflow of a research project, it is not necessarily a requirement that data should be shared openly. Data that is FAIR is not the same as data that are open. A dataset that is open is one in which anyone can easily access, use, and share the data for any purpose without restrictions. However, the FAIR principles provide guidelines for sharing data, while at the same time respecting any legal, ethical, and contractual restrictions that might exist.[8]

The best practices listed here will help project partners produce high-quality data, which will in turn improve research output and impact. It will add more structure to the data, allowing other researchers to reuse it. It will also create more value for research and improve the peer-review process, ensuring high research integrity. As a result, applying the FAIR data principles will contribute to the transparency and reproducibility of FAME project.

FAIR principles are applied differently depending on the discipline and research method. However, there are several common elements to consider within search workflows such as:

- ***Documentation***: An appropriate documentation will ensure the data are understood by others by providing context to it.

- ***Metadata***: Data can be discovered more easily when there are metadata associated with it.
- ***Data formats***: There are a variety of formats that must be considered when combining datasets or making data readable by machines.
- ***Access to data***: The principle of access to data is one where the user decides whose data are available to whom under which conditions and how they can access it.
- ***Persistent identifiers***: The use of persistent identifiers enables others to find and refer to the data by having a link that stays with the data for a long time.[9]
- ***Data licenses***: The use of licenses refers to the legal agreement that defines how third parties may use the data.

Documenting research data can be a challenge. It is, therefore, important to determine a common strategy for structuring and documenting the data among all project partners, keep a detailed record of how the data were collected or generated, use a registration sheet, and assign an object identifier. This may be in the form of methods, instruments, or software.[10]

Project partners should make sure to document the entire "research process". It would be useful to know, for example, who has worked with the data. How were the data used? What was the outcome of the study? What is the relationship between these data and other datasets or publications that have been published? A well-documented research procedure provides context to the data, enabling other researchers to read and understand it in the future.[11]

Metadata – data that describes the data – is another significant type of documentation. In the context of research data or digital objects, metadata comprises descriptive, contextual, and provenance assertions about their properties.[12] A few examples include the location, date, time, file size, file format, keywords, and technical specifications. Data must be accompanied by metadata to be accessed, understood, and processed by both machines and humans. An algorithm, for example, recognizes text and correlates it with other content of a similar nature.

Researchers have always dealt with metadata, and many disciplines have developed standards for meta descriptions. As a method of ensuring the proper and accurate use and interpretation of data, metadata standards attempt to establish a common understanding of the meaning of the data to ensure the data are used correctly.

FAME project requires metadata standards to conduct technological research. The use of metadata can enhance consistency and the ability to process data automatically. In most data combinations, time spent preparing the data is devoted primarily to understanding it. As a result, having common metadata will reduce the time that will be spent on the preparation of data.

Well-designed metadata will ensure that data are easily findable, interoperable and reusable across systems, disciplines and languages. Metadata standards should be used within the FAME project research discipline, based on best practices or those that are widely accepted and used in the field. In the absence of any common standards and automatic metadata generation, project partners should include all the information relevant to the understanding of the datasets. Consider, for example, the preparation of a "Readme" file which contains metadata that describes the dataset in a manner that is likely to be understood by others.[13]

In order to collect and share data from other researchers, open data formats are useful. There are times when data are presented in a strange, binary format. Having data stored in commonly used

formats improves data interoperability and makes it easier for researchers and machines to reuse the data they store.

As a result of the FAIR principles, the reuse of research data is maximized. The metadata should be accessible if there are any restrictions, such as IPR or data protection. Information in the metadata should describe whether the data may be accessed, how they may be accessed, and who may be contacted for access requests. To ensure that data are easily accessible and findable, a persistent identifier for the data should be assigned to it.

“Digital object identifiers” (DOIs), for example – that are commonly used when referencing publications in scholarly journals – direct researchers to the general website where an article has been published, providing them with access if the data was made openly available.

A license determining how the data may be used by others should be provided when making data openly available. There are several types of data licenses. Creative Commons (CC) is the most used license. Depending on the license, researchers can either share data openly or restrictively. A CC license follows the “some rights reserved” model. When using a CC license, the copyright holder may retain a few rights from the bundle conferred by law. CC BY 4.0 may be the most desirable among the many CC licenses due to its ability to allow researchers to share (copy and redistribute) the content in any format or medium they choose. Additionally, they may make adaptations for free or even for commercial purposes. However, they must indicate if changes were made, add a link to the license, and provide appropriate credit to the original creator.

Among the CC bundle of licenses, an open-access license that does not restrict rights is Creative Commons No Rights Reserved (CC0). Through CC0, scientists are essentially able to waive their copyrights and related rights to their fullest extent. This allows them to place their works in the public domain worldwide so that other participants can freely utilize them and improve upon them. Data associated with this deed may be copied, modified, distributed, and communicated without prior authorization.[14] By contrast, the CC attribution, non-commercial, non-derivative license (CC-BY-NC-ND) restricts the use of the data so others cannot use it commercially or modify it in any way.

An alternative license scheme that can also be used is the Open Data Commons Public Domain Dedication and License (PDDL), which allows participants to freely share, modify, and use their work for any purpose. This license may be applied to databases or their content (“data”), either together or separately. Consequently, this license permits the exploitation of the database and its underlying data. PDDL aims to make data and databases as accessible to the public as possible. The objective is to eliminate all restrictions and legal requirements.[15]

Finally, generic data repositories such as Zenodo[16] and Figshare[17] facilitate the publication of FAIR data. Data repositories help to make data FAIR by exposing its metadata to the Internet, providing persistent identifiers and usage licenses. They are an invaluable tool for collaborating with other researchers and increasing the impact of a research project. Additionally, they can be used to preserve valuable data in the future.[18]

As a conclusion, metadata and persistent identifiers are necessary to make data findable. For data to be accessible, it is important to specify who has access to the data and how. It is recommended to provide access to metadata, for example through a data repository, if it is not possible to publish data openly due to data protection and IPRs restrictions. Common standards and open data formats

should be used to facilitate interoperability between datasets. Finally, by adding appropriate documentation to the datasets, others will be able to understand them more easily, and by using the appropriate license, the data can be reused in many ways.

This is also in line with Directive (EU) 2019/1024 on open data and re-use of public sector information (Open Data Directive) and Implementing Regulation (EU) 2023/138 laying down a list of specific high-value datasets and their arrangements for their publication and re-use and in particular, the applicable conditions for re-use and the minimum requirements for disseminating data via application programming interfaces (APIs).

2.2 Collection of information for FAME Data Management Plan

The Information for this deliverable was gathered by the FAME Consortium through a questionnaire (see ANNEX 1), which is based on the "Guidelines on FAIR Data Management" and corresponds to an updated template associated with them (Horizon Europe Data Management Plan Template), with some additional information asked. After various modifications the final version of the questionnaire was sent out to the partners by UPRC. While FAME project partners corresponded promptly to the questionnaire, at this is early stage of the project not all the requested information were available or detailed enough. As this is a living document, all changes are tracked throughout the project and will be included in the next version of the DMP.

3 FAME Project Overview

FAME is a joint effort of world-class experts in data management, data technologies, the data economy, and digital finance to develop, deploy and launch to the global market a unique, trustworthy, energy-efficient, and secure federated data marketplace for Embedded Finance (EmFi). The FAME marketplace will alleviate the proclaimed limitations of centralized cloud marketplaces towards demonstrating the full potential of the data economy. In this direction, the project will enhance a state-of-the-art data marketplace infrastructure (i.e., H2020 i3-Market marketplace) with novel functionalities in three complementary directions namely:

- Secure, interoperable, and regulatory compliant data exchange across multiple federated cloud-based data providers in-line with emerging European initiatives like GAIA-X.
- Decentralized, programmable, data assets trading and pricing leveraging blockchain tokenization techniques (including support for accruing data assets value in NFTs).
- Integration of trusted and Energy Efficient (EE) analytics based on novel technologies such as Quantitative Explainable AI, Situation Aware Explainability (SAX), incremental EE analytics, and edge analytics.

FAME will become operational in a federated cloud environment with multiple providers of EmFi data assets, including datasets, AI/ML models, and more. It will become interconnected with more than 12 data marketplaces that are operated by the project partners, as well as with other data infrastructures that will support the implementation of a plethora of pilots (WP6). Through this process, the data assets catalogue of the FAME marketplace targets to be populated with a critical mass of 1000+ data assets.

Furthermore, FAME will establish a Learning Center (LC) for tech and non-tech users, as this is a key prerequisite for unlocking the potential of the data economy. FAME will build a vibrant community of EmFi stakeholders around the FAME platform, which will serve as a catalyst for the sustainability of the project's results. In order to extract data-driven insights over FAME's assets found in the marketplace, a toolbox for trustworthy and energy efficient analytics will also be developed within the scope of WP5.

The development, implementation and provision of an Authentication and Authorization Infrastructure (AAI) introduced in WP3 will abstract data access, data management and data policies across federated data.

4 Data Summary

4.1 Data assets in FAME

4.1.1 Purpose of the data collection/generation and its relation to the project's objectives

FAME's goal is to specify, implement, and launch to the market a federated multi-cloud, multi-stakeholder, standards-based data marketplace platform that will be used to support discovery, exchange, pricing, and trading of data assets from multiple federated marketplaces. To successfully meet its objectives, FAME will engage in different types of data collection, generation and processing activities. Datasets from pilots and project partners will be collected, stored, organized, and made available for use.

4.1.2 Types and formats of the data assets that the project will generate/collect

The following different data types will be collected or generated within the lifecycle of FAME:

- **Administrative data:** This includes deliverables, recorded telcos, presentations, white papers and working documents produced within the scope of the project. To better manage and track generated documents, a repository for the FAME project has been set up to store documents and information of the project. The repository is based on Microsoft SharePoint, hosted on GFT Space and managed by an administration console. Access is granted only to trusted partners' accounts and is regulated by GFT IT department. Access from other email and gmail.com accounts can be enabled but are deprecated for security reasons. For increased security, access to the Repository is granted by GFT and is regulated by internal IT procedures. The SharePoint drive for the FAME project has been organized to store the documents and information of the project based on the structure of the Workplan (i.e., the structure reflects the work packages), while the repository can also be used as a collaboration platform. All Naming Conventions that should be followed by all partners within the SharePoint are described in Section 6.2.5 of D1.3 – Project Manual – Risk and Quality Management Plan.
- **Software Artifacts:** It includes a vast range of software code, analytics, data management and visualization tools and Explanaible AI (XAI) techniques that will be provided in an AI and Machine Learning (ML) analytics environment that will enable its users to produce analytical insights and AI models, leveraging FAME's available data assets.
- **Datasets:** A variety of datasets will be introduced either by the pilot partners and their various Use Cases, or by the marketplaces and data spaces integrated in the project's federated cloud environment**.** By design, FAME will provide full provenance and traceability. FAME will record metadata about the assets including its type, data volume, data quality features (e.g., completeness of values, readiness for machine learning use), user friendliness scores, carbon footprint and wastes associated with the assets, location specific characteristics (e.g., locality of data capture), timeliness of the asset, demand for the assets (e.g., number of searches and more), algorithmic metadata, and more. These features will be recorded in the blockchain to enable traceability of the status of the assets and of its use in specific applications, as well as the application of pricing and trading schemes using smart contract. The recording of the metadata of the asset in the blockchain will boost the security and the integrity of the data assets, leveraging the blockchain's tamper-proof properties.
- **Training material:** The implementation of Training Programs for EmFi Stakeholders within a Learning Management System (LMS) integrated in the FAME data marketplace.

Presentation of the integration or technical validation of the various components introduced within the scope of FAME.

- **Scientific/Research Output:** As part of our dissemination activities, it is foreseen for FAME partners to make collectively research publications in the context of the project to disseminate our results in International Conferences and Scientific Journals. Open access to all the peer reviewed scientific publications of the project will be provided. Selected publications will be made available with the highest standard (Gold Open Access). The rest publications will be made available in the project's web site, but also in OpenAIRE's Zenodo open access repository. More than five (5) open access journal publications and fifteen (15) conferences are targeted, with individual FAME academic partners and researchers aiming to further contribute, acknowledging the work done withing FAME project, in the fields of advanced AI algorithms, blockchain infrastructures for value trading, as well as energy efficient analytics.

Regarding the overall data format, based on the nature and purpose of the data collected, the data will be stored in a variety of types and formats such as docx, csv, txt, ppt, json, geojson, RDF, mp4, xml, xlsx, etc. In some cases, the datasets will not generate any new data. They will reuse datasets in the types and formats which are available within the project.

4.1.3 Origin of the data

The origin of the data depends on each dataset. Some of these data will be collected from partners and experts. Data will also be created from scratch using different software tools or collected and combined directly through the data lake that is coming from the Use Case scenarios and the associated marketplaces. The tables completed in sub-section 4.2 provide a more detailed description of the types and formats of the data for each dataset and software artifact produced.

Based on the project's goals and in order to achieve synergies and maximize the impact as well as the total number of data being available throughout the FAME federated marketplace, products/solutions from other H2020 projects, marketplaces and partners will be leveraged and are listed on the table below (Table 2):

Table 2 FAME Marketplaces and Data Spaces

Name	Description	Owner	Sector	Available Data Assets
AGORA	AGORA is a data-marketplace connecting Data Providers and Data Consumers for selling and acquiring Connected Vehicle and Home Building data under the standardized data model (CIDM, Common Industrial Data Model). It offers a secure and privacy preserving experience when selling or buying sharing big data, by having the full control over your data shared, to whom and for what purposes. Even though AGORA is targeted mainly to vehicles and home building information, it can be also an implemented in	ATOS	Automotive (initially)	***Open data sets***

	several environments such as smart health or smart energy. AGORA offers to cross-sectorial Data Consumers to combine, integrate and analyse all data at once – regardless of source, type, size, or format – to generate the insights needed to address a wide range of end-users and customers' needs and transform them into new business opportunities in the era of Data Economy.			
SecureIoT	The SecureIoT marketplace is a one stop-shop for Internet of Things (IoT) security solutions and an on-line community of IoT security solutions stakeholders. It facilitates access to IoT security solutions and knowledge, including datasets from the project testing phase.	Datasets owned by organization that uploaded them	Industry, IoT	***Datasets from 3 use cases, webinars, presentations***
IoT-Catalogue	Bringing IoT users and technology providers together from domain needs to IoT products (and back) via validated solutions with components, assembly guides and more.	UNP	IoT	***115 Use Cases, 2119 Components***
FINSEC	FINSEC (https://finsecurity.eu) training assets related to security aspects in the finance sector. The existing marketplace contains resources about Critical Infrastucture Protection	INNOV, GFT	Finance, Security	***100 resources including Training Presentations, Webinars, Whitepapers, Workshops, presentations and training workshops***
INFINITECH Marketplace	Solutions & Services for Finance. The INFINITECH marketplace was established to provide access to the project's solutions, accompanied a Virtualized Digital Innovation Hub (VDIH), aiming to support the innovators (FinTech/InsuranceTech) in their BigData, AI, Blockchain and IoT endeavours.	INFINITECH Project	Finance	***64 Total Assets (containing 18 Datasets, 16 ML algorithms and 8 blockchain assets among others), 85 3rd-Party Tools, 75 Accelerator Programmes, 2 Innov. Support Services, 213 Courses, 14 Workshops, 4 Hackathons, 4 Webinars, 6 How To Videos, 7 Whitepapers***
INFINITECH-Zenodo	Data for the INFINITECH Project - Tailored IoT & BigData Sandboxes and Testbeds for Smart,	INFINITECH Consortium	Finance	***N/A***

	Autonomous and Personalized Services in the European Finance and Insurance Services Ecosystem			
i3-MARKET	i3-MARKET develops a platform for interconnecting and federating heterogeneous marketplaces. It will therefore serve as a base component of the FAME platform: It will be enhanced with the project's AAI, analytics and data trading/monetization results.	i3-MARKET Project	Manufacturing, Climate, Automotive	***50.000 datasets***
Kaggle	Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access GPUs at no cost to you and a huge repository of community published data & code.	Kaggle Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043 USA	Finance, economy, industry, news, etc.	***A few hundred datasets; an additional thousands of datasets via the connected Mobilithek, i.e., the national mobility data access point***
Mobility Data Space	Mobility Data Space (Germany)	DRM Datenraum Mobilität GmbH	Mobility	***87 data assets***
PolicyCLOUD Data Marketplace	The Data Marketplace of the Policy Cloud project gives access to the Policy Cloud Data, Assets and Policies, for improving the modelling, creation and implementation of public and business decision making and policy, for developing common strategies and policies and, ultimately, for enhancing the citizens' quality of life.	UPRC	Smart Cities, Radicalization, Winery, Global Terrorism, Environment	
ADVANEO Data Marketplace	Advaneo has a vision to provide the most open possible access to data for all users serves as the basis for innovation and progress. Open data is a core aspect of the offering (around 2 million open data records from almost 20000 providers). The data marketplace serves as a data catalogue for the metadata provided by the data provider. In addition, it functions as a secure trading portal and offers functions for data processing and administration. It's a decentralized solution that gives users full control over their data at any time. This means that all relevant and possibly sensitive raw data always remain with the data provider and only is transferred directly – peer to	ADVANEO GmbH	Cross-sector	***2 million Open datasets + potentially some commercial datasets (number not disclosed)***

	peer – to the buyer in the event of a purchase. The data marketplace only stores the data descriptions (metadata) and has no access to your raw data at any time.	

4.2 Data assets provided by the Consortium members

The data assets introduced in this sub section have been generated based on the replies of the partners to the DMP Questionnaire (see ANNEX 1), as well as information from FAME's GA. A more detailed presentation of the data assets produced or collected within FAME will be available at the next iteration of this document.

Data Asset ID	**FAME_ATOS_CODE_1**	
Title	**Catalogue of AI/ML Techniques for EmFi**	
Data Overview	**Description**	A library of AI/ML techniques for EmFi use cases
	Partner(s)	**ATOS, ENG**
	Type	Software code/Product
	Related WP/Task	WP5 / T5.1
	Format	Python code (and other programming languages)
	Size	At this moment it is difficult to foresee the size of the data
	Data Collection Methodology	All this data will be collected by sharing it through email, Gitlab, or the SharePoint repository
	Pre-existing Data	Pre-existing data (i.e., software code, input to deliverables, publications and training material) can be used from other related projects (e.g., i3Market, INFINITECH) in which FAME partners have participated and have authorized us with access to it
	Required Tools	Many tools related to ML pipelines will be used during the project's lifetime
Open Data & Accessibility	**Storage & Backup**	This data will be stored in ATOS repositories as a backup strategy
	Data Provenance / Metadata	The provenance of the data will be documented by keeping a versioning history as well as metadata of the data (i.e., data related to their owner, source, date of creation, data type, etc.
	Access Control Responsible	ATOS
	Open Data	The name of ATOS will be included in the data that will be shared
	Scientific / Research Contribution	As part of FAME dissemination activities, it is foreseen to make research publications with other partners from FAME in the context of the project to disseminate our results in

		International Conferences and Journals. Whenever possible, an open-access policy will be promoted
Ethics & Security	**Data Protection Requirements**	All the data will be stored in trusted repositories, in their initial file format
	Personal Data	No
	Type of data	N/A
	Consent Status	N/A

Data Asset ID	FAME_ATOS_CODE_2	
Title	**Edge Data Management and EdgeAI Optimization**	
Data Overview	**Description**	Implementation of profiling techniques for cloud/edge computing applications in different profiles, such as real-time applications, low latency applications, applications requiring many data points, batch processing applications and more
	Partner(s)	**ATOS, ENG, LXS, JSI**
	Type	Software code/Product
	Related WP/Task	WP5 / T5.4
	Format	Python code (and other programming languages)
	Size	At this moment it is difficult to foresee the size of the data
	Data Collection Methodology	All this data will be collected by sharing it through email, Gitlab, or the SharePoint repository
	Pre-existing Data	Pre-existing data (i.e., software code, input to deliverables, publications and training material) can be used from other related projects (e.g., i3Market, INFINITECH) in which FAME partners have participated and have authorized us with access to it
	Required Tools	Many tools related to ML pipelines will be used during the project's lifetime
Open Data & Accessibility	**Storage & Backup**	This data will be stored in ATOS repositories as a backup strategy
	Data Provenance / Metadata	The provenance of the data will be documented by keeping a versioning history as well as metadata of the data (i.e., data related to their owner, source, date of creation, data type, etc.
	Access Control Responsible	ATOS
	Open Data	The name of ATOS will be included in the data that will be shared
	Scientific / Research Contribution	As part of FAME dissemination activities, it is foreseen to make research publications with other partners from FAME in the context of the project to disseminate our results in International Conferences and Journals. Whenever possible, an open-access policy will be promoted

Ethics & Security	**Data Protection Requirements**	All the data will be stored in trusted repositories, in their initial file format
	Personal Data	No
	Type of data	N/A
	Consent Status	N/A

Data Asset ID	FAME_DAEM_DATA_1	
Title	**Retail Services and Citizen Wallets Dataset**	
Data Overview	**Description**	The Athens pilot technical partners in collaboration with DAEM will develop a citizen wallet. The latter will foresee the creation of accounts for end-users for Pilot #2 Embedding - Finance Services in a Personalized Citizen Wallet validation. This technology will be developed by NOVO and UBI. Includes: -Incident Management Datasets -Incidents' Pictures -Parking Data -Real time data from 1,000 sensors
	Partner(s)	**DAEM, NOVO, UBI**
	Type	Dataset
	Related WP/Task	WP6/T6.2
	Format	
	Size	It's not possible to estimate the size
	Data Collection Methodology	All data sets are provided for Athens by NOVO.
	Pre-existing Data	Data from NOVO
	Required Tools	FAME marketplace
Open Data & Accessibility	**Storage & Backup**	FAME should have a horizontal technical strategy on backup and storage
	Data Provenance / Metadata	FAME Sharepoint must have provisioned the security of the project data by design. DAEM will follow this provision
	Access Control Responsible	NOVO and DAEM authorized personnel participated in the FAME project
	Open Data	Not available for re-use
	Scientific / Research Contribution	Any research publications that will be based on the Athens pilot data should include as co-authors DAEM and be openly available.
Ethics & Security	**Data Protection Requirements**	FAME will address this point at a consortium level, DAEM will contribute and follow the decided approach

	Personal Data	No
	Type of data	N/A
	Consent Status	N/A

Data Asset ID	FAME_INNOV_CODE_1	
Title	**Data Quality Assessment Engine**	
Data Overview	**Description**	Validated by Pilot 7 incoming data, this artifact assesses Data assets against their quality, including: (i) Raw sensor datasets; (ii) Pre-processed, filtered and labelled datasets; (iii) Predictive models (ML models) and algorithms. The different assets are audited against different characteristics, including their volume, completeness, locality and context, variety of data sources, use in industrial applications etc. A quality score is assigned to each of the assets to allow their comparative evaluation and to foster the implementation of trading and pricing schemes inside the marketplace.
	Partner(s)	**INNOV, JRC**
	Type	Software code/Product
	Related WP/Task	WP6/T6.4
	Format	The data formats will largely depend on the source and nature of the data and will include formats such as CSV, JSON, TXT for structured and semi-structured data, and DOCX or PDF for documents and reports.
	Size	he expected size is yet to be determined as it depends on the volume of data made available and generated during the project lifecycle. Estimated several GBs
	Data Collection Methodology	Data will be collected through various means, such as directly from databases (FINSEC's DB, MOH's Data Warehouse), surveys, user testing sessions, and software analytics tools.
	Pre-existing Data	INNOV will use pre-existing data from FINSEC and MOH. Moreover, it may use open-source data depending on the needs of the FAME project and will be identified as the project progresses.
	Required Tools	We anticipate needing various tools for data management, such as SQL for database management, Python for data processing and analysis, and tools like Plotly for data visualization. Coding tools like Git will be used for version control.
Open Data & Accessibility	**Storage & Backup**	Some data made available under NDA will be stored exclusively on local servers for processing. Any updates to these data will be versioned and backed up frequently. On the other hand, public data related to the FAME project will also be made available in the project's shared repository.
	Data Provenance / Metadata	INNOV will document the origin, movement, and processing of data according to applicable standards and FAME standardization requirements.

	Access Control Responsible	INNOV, along with the researchers and engineers that participate in the FAME project will check the correct execution of the overall access process.
	Open Data	Data will be catalogued and shared via established data repositories and/or FAME, ensuring its discoverability. Categories of data that will be made reusable include research findings and publications, analytical results, machine learning models, and potentially, anonymized data sets, subject to restrictions. While we aim to facilitate broad reuse of our data, certain restrictions are necessary to protect sensitive information, uphold privacy rights, and maintain proprietary interests.
	Scientific / Research Contribution	We anticipate that research publications will be produced based on the data collected, processed, and generated within the project. INNOV's intention is to make these publications openly available, wherever possible, through open access journals or institutional repositories, to maximize their reach and impact.
Ethics & Security	**Data Protection Requirements**	Data assets will be stored in at least two separate locations (i.e., private cloud and proprietary server). Depending on their type, requirements, and size, different storage and backup strategies will be implemented.
	Personal Data	Yes
	Type of data	INNOV intends to collect and process personal data such as names, contact information, and general user activity for registered users on the FINSEC platform.
	Consent Status	Data collection process has not started yet.

Data Asset ID	FAME_JRC_DATA_1	
Title	**Trade Lists**	
Data Overview	**Description**	Timeseries, historical data generated by algorithmic trading models containing date, time, type of order (open-close, buy-sell) and price
	Partner(s)	**JRC**
	Type	Dataset
	Related WP/Task	WP6 / T6.3
	Format	timeseries, csv, xls, html-links
	Size	1 MB
	Data Collection Methodology	ESG scores from different sources (e.g yahoo, etc) News data/ headlines(e.g. Kaggle Market place)
	Pre-existing Data	Market data as i)Tick data from TradeStation or ii) Daily or hourly data TradeStation From Kaggle Market place: https://www.kaggle.com/datasets/aaron7sun/stocknews

		Also data from Infinitech Marketplace-Zenodo and FINSEC Marketplace
	Required Tools	TradeStation
Open Data & Accessibility	**Storage & Backup**	Yes
	Data Provenance / Metadata	Yes
	Access Control Responsible	JRC department
	Open Data	Kaggle, Zenodo marketplace, INFINITECH (type: timeseries)-publicly accessible
	Scientific / Research Contribution	Quants, Modelling Experts, Traders and developers in order to test algorithms, trading models and usage on sentiment analysis
Ethics & Security	**Data Protection Requirements**	Beyond FAME no
	Personal Data	No
	Type of data	N/A
	Consent Status	N/A

Data Asset ID	FAME_MC_CODE_1	
Title	**Recommendation Engine**	
Data Overview	**Description**	The recommendation engine for families (UC#1), MC/SFS will create an algorithm to preview recommendation of financial products or shopping opportunities with Universo partners and across different entities in SONAE group. Data produced will be the algorithm itself, the recommendations provided for each client, client reaction to the recommendations and the dissemination materials of this use case (research paper, journal publications, presentations, etc.). With the BNPL interfaces test (UC#2) in this project, data will be produced according to the different front-end options to test, IT architecture design and data from client and merchants' interactions with the different interfaces tested.
	Partner(s)	**MC**
	Type	Software code/Product
	Related WP/Task	WP6 / T6.2
	Format	Formats expected: Python code Datasets in xls,csv or similar
	Size	Data relating to over 1 million customers (ex: transactions, digital behaviour, etc). Therefore, data used may reach from the tenths to hundred million rows; data generated around the million rows. These are rough estimates.
	Data Collection	MC/SFS will collect data from other partners in FAME that

	Methodology	prove to be useful for each use case; for example: individual or aggregate consumer behavioural data, macroeconomic indicators, etc. Data from Universo clients, and possibly from partners, will also be collected to develop and feed the recommendation engine, as well as the BNPL interfaces. Finally, data originating from client reactions to the application of each use case will also be collected. Data from MC/SFS and partners will be collected from existing processes or FAME. Collection of data will occur through our internal systems and external tools (such as email or git).
	Pre-existing Data	Universo clients' data and possibly data from partners. Data from FAME partners may also be useful, both for the model itself as well as proof of concept
	Required Tools	To access and visualize, no special tools are likely going to be required beyond common documental software. To process the data, python and cloud tools (Microsoft Azure) will be employed.
Open Data & Accessibility	**Storage & Backup**	Yes, provided by Azure, as well as following internal company backup processes.
	Data Provenance / Metadata	Yes, when possible. Training to standards may be required in early stages.
	Access Control Responsible	Universo - Corporate IT department.
	Open Data	Data will be discovered and shared through FAME. Data reusable or openly accessible will include the recommendation model, as well as publications.
	Scientific / Research Contribution	Yes. Open availability is foreseen when possible.
Ethics & Security	**Data Protection Requirements**	Yes, in our Azure private network. Formats will always include original format, as well as transformed when applicable.
	Personal Data	Yes
	Type of data	Data that proves to be useful for the recommendation engine, possibly including consumer behaviour when available. It is not clear at the moment.
	Consent Status	Yes, from Universo clients.

Data Asset ID	FAME_NRS_DATA_1	
Title	**Dataset based on analytics over environmental data**	
Data Overview	**Description**	Spatially downscaled climate risk indicators that take the form of various summaries of weather at latitude and longitude combinations for a variety of time horizons that relate to climate risk, for example total precipitation, expected maximum hourly precipitation for each month from now until 2055 and other similar indicators, the exact nature of which will be determined through discussion with stakeholders.
	Partner(s)	**NRS**
	Type	Dataset

	Related WP/Task	WP6/ T6.4
	Format	Stored as netCDF files, or, potentially grib files (if necessary – ideally netcdf).
	Size	This depends on ultimate ambition spatially but somewhere between 10 GB to 1 TB.
	Data Collection Methodology	All data will be collected via downloads from open data sources. We will collect climate model output from the CMIP5 or CMIP6 collection. In addition, we will also collect historical reanalysis data (ERA5), which can be thought of as historical weather information. We may also collect seasonal weather forecasts published on the copernicus data store.
	Pre-existing Data	All climate projections, reanalysis data and seasonal forecasts are publicly available preexisting data freely available for download from the Copernicus Climate Data Store. The license of these data is highly permissive and we are therefore allowed to build products on top of these data.
	Required Tools	Standard open source libraries for processing/visualizing these data exist in all major programming languages.
Open Data & Accessibility	**Storage & Backup**	This will have to be discussed with the FAME technical staff, but this should be rather straightforward. It's a large number of netcdf files. Furthermore, if data are lost, they can easily be redownloaded so a backup strategy does not have to be particularly involved.
	Data Provenance / Metadata	The netcdf files we download already adhere to the appropriate standards documenting their provenance thus there will be no additional work required by us.
	Access Control Responsible	Not relevant
	Open Data	he ambition is not to make the originally downloaded data available for reuse as anyone can download these data directly from their source.
	Scientific / Research Contribution	Potentially. All data will be referenced
Ethics & Security	**Data Protection Requirements**	Potentially. But again most of these data can be easily recreated if lost by redownloading from public repositories.
	Personal Data	No
	Type of data	N/A
	Consent Status	N/A

Data Asset ID	FAME_UBI_CODE_1	
Title	**Security Policies Management tool**	
Data Overview	**Description**	Supports access to the security policies of the underlying data marketplaces and data spaces, along with mechanisms for their consolidation at the level of the FAME federated platform.
	Partner(s)	**UBI, NRS**

	Type	Software code/Product
	Related WP/Task	WP3/T3.2
	Format	N/A
	Size	The expected size of the data will be in the range of MBs to GBs
	Data Collection Methodology	N/A
	Pre-existing Data	UBI will exploit the options of using open-source code from other parties depending on the needs of the project. In all cases, the relevant license permissions will be honored and the credits and links to the relevant sources will be documented.
	Required Tools	For the software code depending on the programming language utilize, several tools such as IntelliJ, Eclipse and PyCharm are available.
Open Data & Accessibility	**Storage & Backup**	UBI will exploit its own private cloud and proprietary server for the storage of the produced data where a backup strategy is applied on a corporate level. In addition to this, UBI will exploit the tools that will be utilized by the consortium such as Gitlab and SharePoint as provided by the project's management team.
	Data Provenance / Metadata	UBI will ensure the provenance of the generated data with the appropriate standards for the documentation of all the suitable metadata such as their origin, their versioning, list of contributors, licensing, etc. Tools like online source code repositories such as Gitlab and file repositories such as SharePoint will be leveraged following the project's standards and requirements.
	Access Control Responsible	UBI's IT administrators, along with the researchers and engineers that participate in the FAME project will check the correct execution of the overall access process.
	Open Data	All data (software code, deliverables, publications) that will be characterized as open will adhere to the OpenAIRE requirements and FAIR principles hence they will be reusable The audience of the reused data will be the research community as well as industrial and SMEs organisations
	Scientific / Research Contribution	UBI aspires to produce a series of publications based on the generated data and whenever possible they will be made openly available
Ethics & Security	**Data Protection Requirements**	Open access data will be stored in their initial format in Zenodo which guarantees long term preservation of the published data elements for at least the next 20 years. Generated data that will be stored on UBI's private cloud and proprietary server will be stored and managed with the highest security standards for a maximum of three years after the project ends.
	Personal Data	No
	Type of data	N/A
	Consent Status	N/A

Data Asset ID	FAME_UBI_CODE_2	
Title	**Authentication and Authorization Infrastructure Interfaces**	
Data Overview	**Description**	Leverages the AAI interfaces to access the data and annotate semantically the assets
	Partner(s)	**UBI**
	Type	Software code/Product
	Related WP/Task	WP3/T3.2
	Format	N/A
	Size	The expected size of the data will be in the range of MBs to GBs
	Data Collection Methodology	N/A
	Pre-existing Data	UBI will exploit the options of using open-source code from other parties depending on the needs of the project. In all cases, the relevant license permissions will be honored and the credits and links to the relevant sources will be documented.
	Required Tools	For the software code depending on the programming language utilize, several tools such as IntelliJ, Eclipse and PyCharm are available.
Open Data & Accessibility	**Storage & Backup**	UBI will exploit its own private cloud and proprietary server for the storage of the produced data where a backup strategy is applied on a corporate level. In addition to this, UBI will exploit the tools that will be utilized by the consortium such as Gitlab and SharePoint as provided by the project's management team.
	Data Provenance / Metadata	UBI will ensure the provenance of the generated data with the appropriate standards for the documentation of all the suitable metadata such as their origin, their versioning, list of contributors, licensing, etc. Tools like online source code repositories such as Gitlab and file repositories such as SharePoint will be leveraged following the project's standards and requirements.
	Access Control Responsible	UBI's IT administrators, along with the researchers and engineers that participate in the FAME project will check the correct execution of the overall access process.
	Open Data	all data (software code, deliverables, publications) that will be characterized as open will adhere to the OpenAIRE requirements and FAIR principles hence they will be reusable The audience of the reused data will be the research community as well as industrial and SMEs organisations
	Scientific / Research Contribution	UBI aspires to produce a series of publications based on the generated data and whenever possible they will be made openly available
Ethics & Security	**Data Protection Requirements**	Open access data will be stored in their initial format in Zenodo which guarantees long term preservation of the published data elements for at least the next 20 years. Generated data that will be stored on UBI's private cloud and proprietary server will be stored and managed with the

		highest security standards for a maximum of three years after the project ends.
	Personal Data	No
	Type of data	N/A
	Consent Status	N/A

Data Asset ID	FAME_UPRC_CODE_1	
Title	**XAI Scoring Framework**	
Data Overview	**Description**	A framework for scoring the explainability of the different models towards comparing alternative approaches, balancing performance vs. explainability trade-offs and supporting the project's trading and monetization schemes included in WP4.
	Partner(s)	
	Type	Software code
	Related WP/Task	WP5/T5.2
	Format	Software code will be firstly in the format of Java/Python language, and will be updated to other programming languages - if needed
	Size	The expected size of the data will be calculated from Kbytes to Gbytes.
	Data Collection Methodology	All this data will be collected through sharing it by email, Gitlab or the SharePoint repository.
	Pre-existing Data	Prexiisting data (i.e., software code, input to deliverables, publications and training material) will be used from other related projects (e.g., i3Market, INFINITECH) in which FAME partners have participated and have authorized us with access to it.
	Required Tools	Integrated Development Environments (IDEs), such as NetBeans, Eclipse, SPYDER, etc.
Open Data & Accessibility	**Storage & Backup**	Data will be stored locally, and on personal cloud repositories as a backup strategy
	Data Provenance / Metadata	The provenance of the data will be documented by keeping a versioning history as well as metadata of the data (i.e., data related with their owner, source, date of creation, data type, etc.).
	Access Control Responsible	The University of Piraeus Research Center, along with the researchers that participate in the FAME project will check the correct execution of the overall access process.
	Open Data	Data (e.g., Software code, publications, deliverables) may be used to whom may interest. The name of UPRC will be included to the data that will be shared. Data including software code, deliverables, and publications will be re-usable.
	Scientific / Research Contribution	PRC plans to make research publications with this data in the context of International Conferences and Journals. If feasible, this will be openly available in the form of Golden Access.

Ethics & Security	**Data Protection Requirements**	All the data will be stored in private cloud repositories, in their initial file format.
	Personal Data	No
	Type of data	Names - as well as personal details - of partners
	Consent Status	There have not been created/signed any specific consent forms.

Data Asset ID	FAME_UPRC_CODE_2	
Title	**FAME Dashboard**	
Data Overview	**Description**	**FAME Dashboard**
	Partner(s)	The dashboard provides a comprehensive overview of the project's outcomes and the status of all technical developments. It also specifies and implements OpenAPIs for accessing the functionalities of the data marketplace platform, including relevant documentation for the users of the dashboard and of the Open APIs
	Type	**UPRC**
	Related WP/Task	Software code
	Format	WP2/T2.4
	Size	Software code will be firstly in the format of Java/Python language, and will be updated to other programming languages - if needed
	Data Collection Methodology	The expected size of the data will be calculated from Kbytes to Gbytes.
	Pre-existing Data	All this data will be collected through sharing it by email, Gitlab or the SharePoint repository.
	Required Tools	Preexisting data (i.e., software code, input to deliverables, publications and training material) will be used from other related projects (e.g., i3Market, INFINITECH) in which FAME partners have participated and have authorized us with access to it.
Open Data & Accessibility	**Storage & Backup**	Integrated Development Environments (IDEs), such as NetBeans, Eclipse, SPYDER, etc.
	Data Provenance / Metadata	Data will be stored locally, and on personal cloud repositories as a backup strategy
	Access Control Responsible	The provenance of the data will be documented by keeping a versioning history as well as metadata of the data (i.e., data related with their owner, source, date of creation, data type, etc.).
	Open Data	The University of Piraeus Research Center, along with the researchers that participate in the FAME project will check the correct execution of the overall access process.
	Scientific / Research Contribution	Data (e.g., Software code, publications, deliverables) may be used to whom may interest. The name of UPRC will be included to the data that will be shared. Data including

		software code, deliverables, and publications will be re-usable.
Ethics & Security	**Data Protection Requirements**	UPRC plans to make research publications with this data in the context of International Conferences and Journals. If feasible, this will be openly available in the form of Golden Access.
	Personal Data	All the data will be stored in private cloud repositories, in their initial file format.
	Type of data	No
	Consent Status	Names - as well as personal details - of partners

4.3 Data assets provided by the Pilot partners

The tables regarding FAME's pilots have been created based on the pilot leaders replies to Section 8 of the DMP Questionnaire (Annex 1). During the writing of this document, due to the early circulation of the DMP Questionnaire, not all pilots have filled the template as the co-creation workshops of WP2 shaping the pilots use cases (T6.2-6.4) and data shared were still on-going. However, the available information gathered regarding pilots data assets have been included in Section 4.2. The next version of this deliverable will provide a detailed presentation of all use cases and their respective data assets.

- **Pilot #1: FaMLy – A powerful financial recommendation engine for families**

Pilot 1 UC1: Recommendation Engine for Families – SFS

Section	Description
ID	FAME-DS-P1-01
Title	Customer data based on SONAE's Universo Card / Social and demographic characteristics of customers
Description	Includes age, nationality, gender, marital status, number of dependents, monthly net salary, address, home type, profession, employer.
Owner	SFS
Licence / Privacy	
Data type	structured
Type of process (stream or static data)	static
Data format	.csv/ .xls/ sql / etc.
Data store	cloud
Data Security	access control
Regulatory Constraint Requirements	NDA required

Section	Description
ID	FAME-DS-P1-02
Title	Customer data based on SONAE's Universo Card /

	transactional and behavioural data
Description	Includes transactions (types, merchants, amounts, dates, location, etc.) and digital tracking data (client location, logins, app interaction, etc.)
Owner	SFS
Licence / Privacy	
Data type	structured
Type of process (stream or static data)	static
Data format	.csv/ .xls/ sql / etc.
Data store	cloud
Data Security	access control
Regulatory Constraint Requirements	NDA required

Section	Description
ID	FAME-DS-P1-03
Title	Customer text data
Description	Includes text written by customers, such as app reviews, answers to surveys, possibly social media posts
Owner	SFS
Licence / Privacy	
Data type	Semi-structured
Type of process (stream or static data)	static
Data format	.csv/ .xls/ sql / etc.
Data store	cloud
Data Security	access control
Regulatory Constraint Requirements	NDA required

Section	Description
ID	FAME-DS-P1-04
Title	Customer recommendations - output
Description	Includes client recommendations obtained from the recommendation engine
Owner	SFS
Licence / Privacy	
Data type	structured
Type of process (stream or static data)	static
Data format	.csv/ .xls/ sql / etc.
Data store	cloud
Data Security	access control
Regulatory Constraint Requirements	NDA required unless anonymized

Pilot 1: UC2: Buy Now Pay Later interfaces

Section	Description

ID	FAME-DS-P2-01
Title	Customer data based on SONAE's Universo FLEX / i. Social and demographic characteristics of customers ii. Characteristics of merchant and purchase
Description	includes age, nationality, gender, address, profession, employer. Includes type of activity, nº employees, purchase amount (nominal or aversage)
Owner	SFS
Licence / Privacy	
Data type	structured / unstructured
Type of process (stream or static data)	stream / static
Data format	.csv/ .xls/ etc.
Data store	locally / cloud
Data Security	anonymization / pseudonymization / access control
Regulatory Constraint Requirements	any data restrictions or constraints /NDA requirements

- **Pilot #5: ESG Scorecard Ranking & Sustainable Portfolio Optimisation**

Section	Description
ID	FAME-DS-P5-01-JRC
Title	Financial Markets TickData
Description	Timeseries, real market prices e.g. currencies, stocks
Owner	JRC
Licence / Privacy	-
Data type	structured
Type of process (stream or static data)	static
Data format	.csv
Data store	cloud
Data Security	-
Regulatory Constraint Requirements	No (during the project)

Section	Description
ID	FAME-DS-P5-02-JRC
Title	Trade Lists
Description	Timeseries, historical data generated by algorithmic trading models containing date, time, type of order (open-close, buy-sell) and price
Owner	JRC
Licence / Privacy	-
Data type	structured

Type of process (stream or static data)	static
Data format	.html or .csv
Data store	cloud
Data Security	Sensitive to re-engineering attempts
Regulatory Constraint Requirements	No (during the project)

- **Pilot #6: Embedding Climatic Predictions in Property Insurance Products**

Section	Description
ID	FAME-DS-P6-01
Title	CMIP climate projections
Description	Climate projections for a variety of scenarios and issuing organizations downloaded from the Copernicus Data Store (CDS)
Owner	These are publically available dataSFS
Licence / Privacy	These are publically available data with an open access license. We must simply acknowledge CDS
Data type	structured
Type of process (stream or static data)	static
Data format	netcdf
Data store	cloud
Data Security	N/A
Regulatory Constraint Requirements	N/A

Section	Description
ID	FAME-DS-P6-02
Title	ERA5 Reanalysis data
Description	Reanalysis data (“observed weather”) downloaded from CDS
Owner	These are publically available dataSFS
Licence / Privacy	These are publically available data with an open access license. We must simply acknowledge CDS
Data type	Structured
Type of process (stream or static data)	Static
Data format	Netcdf
Data store	Cloud
Data Security	Not relevant
Regulatory Constraint Requirements	Not relevant

Section	Description
ID	FAME-DS-P6-03
Title	Seasonal Forecast Data
Description	Seasonal Forecast products for a number of

	forecasting systems issued on a monthly basis and hosted on CDS.
Owner	These are publicly available dataSFS
Licence / Privacy	These are publicly available data with an open access license. We must simply acknowledge CDS
Data type	Structured
Type of process (stream or static data)	Static
Data format	Netcdf
Data store	Cloud
Data Security	Not relevant
Regulatory Constraint Requirements	Not relevant

Section	**Description**
ID	FAME-DS-P6-04
Title	Climate Risk Indices
Description	This dataset will provided a set of climate risk indices (whose exact formulation has yet to be finalized) that will give a user the ability to enter in a longitude, latitude and time horizon and receive a projected set of risk indices for this grouping, based on a statistical downscaling model which combines the climate projections, reanalysis data and seasonal forecast information, where the exact nature of the combination depends on the time horizon of interest.
Owner	NRS/GFT
Licence / Privacy	TBD
Data type	Structured
Type of process (stream or static data)	Static
Data format	Netcdf
Data store	Cloud
Data Security	TBD
Regulatory Constraint Requirements	No

5 FAIR Principles in FAME

The Federated Authentication and Authorization Infrastructure (AAI) developed in T3.1 implements an infrastructure over different types of data providers and their data infrastructures (i.e., data marketplaces, data spaces, databases etc.). The project leverages the existing i3-MARKET platform, which provides a baseline support for self-sovereign identities and access to data from federated marketplaces. The platform will be extended with interfaces that will adhere to the needs of FAME. Specifically, interfaces for data assets trading, pricing and data policy management are implemented, while also, existing interfaces are enhanced to support additional data exchange models and ontologies. FAME ontologies and models for EmFi are specified in T3.4 - Semantic Interoperability Middleware.

To secure findability and data reuse, a federated catalogue that will comprise data assets from the various federated data sources and marketplaces is introduced in T3.3. In this direction, the various data assets will be modelled according to the FAME models and will be structured according to embedded finance ontologies and taxonomies. The Federated Catalogue of Data Assets also specifies and implements APIs for accessing, searching, and querying the federated catalogue. These APIs will ensure a seamless and marketplace agnostic experience for the users of the FAME marketplace i.e., the users will be able to discover data assets across different marketplaces and data spaces.

The management of data policies in the FAME marketplace is handled in T3.2 - Unified Security Policy Management. The implementation of a security policies management tool aims to map FAME policies to the lower-level policies of the underlying providers, aligning to the FAIR principles. The tool will enable the specification of policies over the federated marketplaces. The starting point for the specifications is existing ontologies of the finance sector (e.g., FIBO, FIGI, H2020 INFINITECH ontologies). These ontologies will be linked with ontologies from other sectors (e.g., retail, smart cities, healthcare) in-line with the requirements of embedded finance use cases. Moreover, the task implements a middleware that will transform data from the formats and semantics of the individual underlying marketplaces and data spaces to the FAME ontologies. In this direction, the task will leverage the AAI interfaces to access the data and annotate semantically the assets.

Furthermore, search engine results ranking techniques to implement dynamic market mechanisms for data assets trading and monetization will be leveraged in the scope of T4.4 - Semantic Search for Trading and Valuation of Data Assets. In more detail, this task implements semantic search over the federated catalogue along with schemes for ranking the results according to relevance and value-based attributes of the data assets and is an important component in the overall data management of FAME and its federated data assets.

A wide range of data cleaning and interoperability standards-based techniques will be employed to detect and correct corrupt or inaccurate records and homogenize the collected data (complementing the functionalities of T3.4 for Interoperability).

Below, after a short presentation of each FAIR relevant sub-principle based on the “FAIR Guiding Principles for scientific data management and stewardship” and the “Three-point FAIRification Framework”, the corresponding Tasks that indicatively implements the referred mechanisms in compliance with FAIR principles are summarised in Table 3.

Table 3 FAIR sub-principles and the Tasks that will deploy the corresponding mechanisms ensuring FAIR Principles in FAME

FAIR SUB-Principles	FAME Tasks
F1. (Meta)data are assigned a globally unique and persistent identifier	T3.3, T3.4, T4.1
F2. Data are described with rich metadata	T3.3, T3.4, T4.1
F.3Metadata clearly and explicitly include the identifier of the data they describe	T3.3, T3.4, T4.1
F4. (Meta)data are registered or indexed in a searchable resource	T3.3
A1. (Meta)data are retrievable by their identifier using a standardized communication protocol	T3.1, T3.3, T3.4, T4.4
A1.1 The protocol is open, free, and universally implementable	T2.4, T3.1, T3.3
A1.2 The protocol allows for an authentication and authorization procedure, where necessary	T2.4, T3.1
A2. Metadata are accessible, even when the data are no longer available	T3.1, T3.3, T3.4, T4.4
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	T3.4
I2. (Meta)data use vocabularies that follow FAIR principles	T3.2, T3.4
I3. (Meta)data include qualified references to other (meta)data	T3.3, T3.4
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes	T3.2, T3.3, T4.1
R1.1. (Meta)data are released with a clear and accessible data usage license	T3.2, T3.5
R1.2. (Meta)data are associated with detailed provenance	T3.4, T4.1
R1.3. (Meta)data meet domain-relevant community standards	T3.4

5.1 Making data findable, including provision for metadata

- ***F1. (Meta)data are assigned a globally unique and persistent identifier***

This sub-principle refers to the uniqueness and persistence of the identifiers that FAME will develop. To ensure the findability of data and research outputs of the project, globally unique and persistent identifiers (PID) will be provided (e.g., DOIs).

- ***F2. Data are described with rich metadata***

This sub-principle dictates that metadata shall be extensive, with reliable information regarding the data's context, quality, condition and characteristics. Metadata of the project's datasets must be machine-readable to optimize their discovery. A URL connected to a document that contains machine-readable metadata and the format of the file must be provided.

- ***F3. Metadata clearly and explicitly include the identifier of the data they describe***

This sub-principle measures whether the metadata document includes the globally unique and persistent identifier for the digital resource. Because metadata and the datasets they describe are in different files, these two files must be associated by describing a dataset's globally unique and persistent identifier in the metadata file. Metadata shall contain the identifier of the digital resource it describes. As mentioned in the F1 description, many repositories address these issues. An

example is given though http://go-fair.org regarding this connection annotated in a formal manner using the foaf:primaryTopic predicate in the case of RDF metadata.

- ***F4. (Meta)data are registered or indexed in a searchable resource***

This sub-principle refers to finding digital resources using web-based search engines. Search engines must index the digital resource and its metadata for the users to find and reuse them. This can be provided by the persistent identifiers of the resource and URLs that give search results on various search engines.

5.2 Making data accessible

- ***A1. (Meta)data are retrievable by their identifier using a standardized communication protocol***
  - ***A1.1 The protocol is open, free, and universally implementable***

This FAIR criterion examines whether the access protocol is free (no-cost) and open-sourced, facilitating data reuse. These two criteria refer at least to the metadata. The project's datasets shall be available as open data in various FAIR repositories (e.g., Zenodo).

  - ***A1.2 The protocol allows for an authentication and authorization procedure, where necessary***

The conditions under which the data are accessible are defined in this sub-principle. Thus, "closed" or protected data are also FAIR. Furthermore, the specification of a protocol to access restricted content must be defined for content that is available under restriction. In cases where datasets are confidential, the project will consider authentication and authorization in specific accessing procedures.

- ***A2. Metadata are accessible, even when the data are no longer available***

This metric examines the existence of metadata even in the absence of data. This principle states that metadata will persist over time, continuing to provide descriptors of the data in connection with F3-F4 principles (metadata remains discoverable with a persistent identifier).

5.3 Making data interoperable

- ***I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.***

To ensure the automatic findability and interoperability of datasets, the project will adopt commonly used vocabularies and ontologies and a well-defined data model that describes the structure of the metadata (e.g., RDF knowledge representation model). T3.4 - Semantic Interoperability Middleware will provide the mechanisms for choosing a suitable schema ensuring schema interoperability through metadata annotation.

- ***I2. (Meta)data use vocabularies that follow FAIR principles***

Metadata should use community-accepted vocabulary and terminology that are themselves FAIR. The provenance descriptors and vocabulary used within datasets must be documented and resolvable using globally unique and persistent identifiers. In particular, the project will incorporate

and extend the Mobility Data Specification (MDS) and the outcomes of the Open Mobility Foundation (OMF).

- ***I3. (Meta)data include qualified references to other (meta)data***

Relationships within the various multidimensional datasets and (meta)data that will be used towards the FAME project, must have explicit and useful semantic meaning. An example of semantics of these relationships is: “objectA wasFoundInTheSamePlaceAs objectB”, which is an acceptable qualified reference. This principle addresses the specification of: a) interlinking between datasets, b) the implementation (if needed) of an additional dataset to complete the data, c) information about data that are stored in a different dataset.

5.4 Increase data re-use

- ***R1. (Meta)data are richly described with a plurality of accurate and relevant attributes***
  - ***R1.1. (Meta)data are released with a clear and accessible data usage license***

This sub-principle refers to legal interoperability by defining the conditions under which the metadata can be reused. These conditions are clarified through a license document for data and its associated metadata. This document will be provided by the IRI of license (e.g., URL) for both the data and metadata license.

  - ***R1.2. (Meta)data are associated with detailed provenance***

This sub-principle ensures that there is provenance information about the data, mainly regarding two topics of provenance: a) Who/what/when produced the data (i.e., whom to cite, who generated or collected the data), b) why/how the data was produced (i.e., description of the workflow that led to the data, to understand its context). This information will be delivered by T3.2, depicting the overall provenance of the data that will be processed during the project using ontologies and state-of-the-art approaches (e.g., PROV-O ontology). For instance, a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.

  - ***R1.3. (Meta)data meet domain-relevant community standards***

Many communities have established metadata guidelines to maximize their shared data's practical usability. As a result, they provide validation services to certify the compliance of the data. Furthermore, these communities offer several standards and requirements regarding context and provenance to improve the reusability of the data. Towards this direction, if such communities exist in the domain of Traffic/Transportation in the context of FAME, certifications or validation techniques (e.g., a Verisign signature) must be implemented. Towards this end, the project already collaborates with various standardization organizations such as GAIA-X and IDSA.

6 Allocation of Resources

The costs required for making the data collected/generated in the course of FAME activities FAIR are integrated within the budget of the project. These estimated costs will be needed to cover a set of specific data processing and data management activities, spanning from collection and documentation through storage and preservation over to sharing and re-use.

Regarding the data management responsibilities for administrative documents such us deliverables, presentations, etc, all WP Leaders, Task and Deliverable Leaders are described within FAME's GA.

7 Data Security

Data protection is a central issue for research ethics. A fundamental human right, enshrined in the EU Charter of Fundamental Rights, provides all individuals with control over the way information about them is collected and used. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) grant everyone the right to the protection of personal data concerning him or her and GDPR lays down rules relating to the protection of natural persons with regard to the processing of personal data.

Wherever personal data - information relating to an identified or identifiable natural person is processed[19] special care has to be taken. In research settings, data protection imposes obligations on researchers to provide research subjects with detailed information about what will happen to the personal data that they collect. Particular attention has to be paid to research involving sensitive data such as health data, which according to GDPR must not be processed unless the data subject has given explicit consent.[20]

When personal data is used, informed consent is the cornerstone of research ethics. The lawful bases for the processing of personal data related to stakeholders, under the GDPR, is that each data subject working for a Project partner has given consent to the processing of his or her personal data (Article 6 (1)(a)) and that the processing is necessary for the performance of a contract - namely, the data subjects' employment agreements with each Project partner (Article 6(1)(b)). After the end of the project, files containing personal information of data subjects working for Project partners will be maintained by each Project partner. Any partner will have the right to continue to maintain its copy of the contact data of employees working for the FAME Project partners unless said that employees request a deletion of the contact data. Mailing lists of the project will be deleted only after the very final payment and assessment from the European Commission. Data subjects' contact details will be shared only with Project members and only for the time needed to execute the Grant Agreement and/or complete the Project. Authorship information may be made publicly available with the consent of the data subjects once the application becomes publicly or commercially available.

All partners responsible for processing data have the responsibility, hence ensure that the data remains protected under all necessary security controls (including backup policies and integrity checks) and access controls (identification, authentication, authorization) within their infrastructure. In the unfortunate event of personal data breach, the project partners will notify without delay their competent national supervisory authorities as well as the data subject(s) that may be affected by the breach. At the same time, they will document any personal data breaches and all related information.

A data controller is required to notify the supervisory authority within 72 hours of becoming aware of a data breach[21] if it represents a risk to data subjects. Any breach of personal data must be documented by the controller, detailing the facts related to the breach, the consequences and the remedial actions taken.

Regulatory Compliance is considered a horizontal task in FAME, affecting all consortium partners. As part of the project, all partners are required to ensure that appropriate measures are taken to prevent and minimize the impact of incidents concerning the security of the platform, the network

and information systems. Furthermore, all participants in the project should implement a suitable continuity plan in order to ensure that the activities of the project continue in a safe manner. Whenever there is a breach of personal information, it is necessary to assess the impact of the incident to mitigate adverse effects and prevent future incidents from occurring. Several factors need to be considered when assessing the significance of a personal data breach, including the number of affected individuals, the extent of the impact on their rights and freedoms, as well as on the project activities, and the duration of the incident.

Each partner shall have a Data Protection Officer (DPO) and GFT has appointed a Project Security Responsible (PSR). In case of a potential data breach, the PSR officer should take mitigation actions in accordance with all the DPOs.

8 Ethical Aspects

FAME project will be carried out in line with the highest ethical standards and the applicable EU, international and national law on ethical principles. Ethics are considered significant in the FAME project and are a distinct part of the Grant Agreement (Annex 5), while the ethical requirements are the sole purpose of WP8, whose sole objective is to ensure ethical compliance with the ethical requirements set in the project.

This reflects to the report of the ethical standards of FAME project which corresponds to T1.4 and D1.3 Ethical Management and Regulatory Compliance Framework (M12), which develops the project's social, legal, and ethical activities framework, including the project's methodology for handling social science and gender issues. The framework will ensure the compliance of the project's technologies to applicable laws and regulations (e.g., 4AML, PSD2, MiFIDII, GDPR), as well as to emerging regulations (e.g., AI regulation Act). It also deals with issues like data sources specification and co-creation, identification of the environmental dimension of the research, and social impact assessment. Moreover, the task will identify ethics and risks associated with the project's technologies, along with guidelines to mitigate them. For the work of the task to be performed and to establish a solid ethical framework, the task leader will appoint an Ethics and Legal Expert with knowhow on financial regulations. It will also mobilize the FAME Ethical Board (EB), which will audit the research activities against ethical and regulatory compliance. The designed ethical framework also includes frequent interactions with the use cases towards ensuring their compliance to regulatory requirements based on the FAME technologies and tools.

Once the project's AI tools and software components have reached a more mature phase, a new round of questionnaires will be circulated to the consortium partners focusing primarily on the AI ACT proposal [22] and the Assessment List for Trustworthy Artificial Intelligence (ALTAI).[23]

References

1. Data Management Plan Template, available at : https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf (retrieved 12/04/23)
2. Go FAIR, FAIR Principles, https://www.go-fair.org/fair-principles/ (retrieved 12/04/23)
3. Mark Wilkinson et al., 'The FAIR Guiding Principles for scientific data management and stewardship' Sci Data3, 160018 (2016), available at https://doi.org/10.1038/sdata.2016.18
4. EOSC Declaration 2017, p. 1., available at: https://eosc-portal.eu/sites/default/files/eosc_declaration.pdf (retrieved 28/04/23)
5. Mons Barend et al., 'Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud', Information Services & Use, 37(1):50, 2017. DOI: 10.3233/ISU-170824 (retrieved 12/04/23)
6. European Commission, 'Turning FAIR Into Reality', Final Report and Action Plan from the European Commission Expert Group on FAIR Data, available at: https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf (retrieved 12/04/23).
7. European Commission, 'Turning FAIR Into Reality', Final Report and Action Plan from the European Commission Expert Group on FAIR Data, available at: https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf (retrieved 12/04/23)
8. Martínez-Lavanchy, et al. (2019). 'FAIR Principles'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), Research Data Management (eLearning course). doi: 10.11581/dtu:00000049
9. Chue Hong et al., Six Recommendations for Implementation of FAIR Practice (2020) http://doi.org/10.5281/zenodo.4065549
10. Martínez-Lavanchy, et al. (2019). 'FAIR Principles'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), Research Data Management (eLearning course). doi: 10.11581/dtu:00000049
11. Ibid.
12. Gary Berg-Cross, Raphael Ritz and Peter Wittenburg, RDA Data Foundation and terminology – DFT; Results RFC (v1.5) (2015), p. 2, available at: https://www.rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf
13. Martínez-Lavanchy, et al. (2019). 'FAIR Principles'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), Research Data Management (eLearning course). doi: 10.11581/dtu:00000049
14. Creative Commons, CC0 1.0 Universal (CC0 1.0), available at: https://creativecommons.org/publicdomain/zero/1.0/
15. Open Knowledge Foundation, Open Data Commons: Legal Tools for Open Data, available at: https://opendatacommons.org/licenses/pddl/1-0/
16. Zenodo, available at: https://zenodo.org/record/999150#.Y1lPHVJBx_Q (retrieved 30/05/23)
17. Figshare, available at: https://figshare.com/ . (retrieved 30/05/23)
18. Martínez-Lavanchy, et al. (2019). 'FAIR Principles'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), Research Data Management (eLearning course). doi: 10.11581/dtu:00000049
19. Processing, according to Regulation (EU) 2016/679 of the European Parliament (GDPR), means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by

transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

20. GDPR Art 9(2)(a).
21. The GDPR defines a “personal data breach” in Article 4(12) as: a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed.”
22. Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206, (retrieved 30/05/23)
23. Assessment List for Trustworthy Artificial Intelligence, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (retrieved 30/05/23)

Annex I

Questionnaire completed by each FAME consortium partner

This questionnaire is intended for all partners collecting, processing, or generating data in the scope of FAME project, with the contribution of each partner included and shaping D.1.2- Data Management Plan. For each data category/data type (these two terms are used as synonymous) you plan to utilize, please provide a separate answer to the following questions, clarifying its underlying attributes.

1. Data Summary: The following questions aim to provide an overview of what types of data will be generated, collected and/or shared within FAME project.

- What type of data will you produce or generate during the Project?
- What type of data will you collect during the Project?
- How will you collect the data? In what formats?
- Will the provenance of the data be thoroughly documented using the appropriate standards?
- Will you use pre-existing data? From where?
- What is the expected size of the data that you intend to generate or re-use?
- Are there tools or software needed to create/process/visualize the data?
- Is there a storage and backup strategy in place?

2. Data Organization, Documentation and Metadata: The following questions are intended to understand the plan for organizing, documenting, and using descriptive metadata to assure quality control and findability of the respective data.

- What standards will be used for documentation and metadata (e.g., Digital Object Identifiers)?
- Do you use any best practices/guidelines for managing the data to publish (i.e., make available to third parties)?
- Do you use any tool for checking that the data are well formatted?
- What directory and file naming convention will be used?
- What project and data identifiers will be assigned?
- Is there a community standard for metadata sharing/integration?
- Will metadata be offered in such a way that it can be harvested and indexed?

3. Data Accessibility: The following questions aim to identify any data access and ownership concern.

- What steps will be taken to protect privacy, security, confidentiality, intellectual property or other rights?
- Does your data have any access concerns? Describe the process someone would take to access your data.
- Who checks the correct execution of the access process (e.g., lab, University, funder)?
- What procedures have you developed for the safe transfer of personal or sensitive data?
- Do you plan to make any research publications based on the data collected, processed or generated within the project? If yes, is it going to be openly available?

4. Data Interoperability: What data and metadata vocabularies, standards, formats or methodologies will be followed to make your data interoperable and facilitate data exchange.

- Will your data include qualified references to other data (e.g., other data from your project, or datasets from previous research)?
- Will your data use a formal, accessible, shared, and broadly applicable language for knowledge representation?
- 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?

5. Data Reuse and Sharing: The following questions are intended to clarify how the collected data will be released for sharing and to evaluate their reproducibility.

- If you allow others to reuse your data, how will the data be discovered and shared? List the categories of data that will be made re-usable or openly accessible.
- Will the process of data generation be reproducible? What would happen if collected data got lost or became unusable later?
- What is the audience for the reused data? How are they potentially utilizing the data?
- Any restrictions on who can re-use the data and for what purpose?

6. Data Security and Preservation

The following questions are intended to clarify how the collected data will be preserved and archived.

- Will the data be safely stored in trusted repositories for long term preservation and curation? In what format?
- 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)?
- Who decides what data or what categories of data will be kept and for how long?
- Who will maintain the data for the long-term?

- Have you prepared a formal risk assessment addressing each of the major risks to data security and potential solutions?
- Any special privacy or security requirements (e.g., personal data, high-security data)?
- The GDPR requires personal data not be kept longer than necessary for the purpose for which it was stored. What protocol(s) will you put in place to ensure you delete personal data that is no longer required to be stored?

7. Ethical Aspects:

- What types of personal data do you intend to collect, generate or process?
- What types of sensitive data do you intend to collect, generate or process?
- Will any of the data subjects be children or vulnerable people?
- Will you be collecting personal or sensitive data from people who have not given their explicit consent to participate in the Project?
- If you collected personal data, as defined by the GDPR, which of the six Art. 6.1 bases will you rely on for the processing of each category of personal data?
  - http://www.privacy-regulation.eu/en/article-6-lawfulness-of-processing-GDPR.htm
- If you collected sensitive data, as defined by the GDPR, which of the ten Art. 9 bases will you rely on for the processing of each category of sensitive data?
  - http://www.privacy-regulation.eu/en/article-9-processing-of-special-categories-of-personal-data-GDPR.htm
- Have you already gained consent for data preservation and sharing from any data subject(s)?
- Will you engage in large scale or big data processing?
- Will any entity (including any service provider) outside of the E.U. have access to personal or sensitive data? If yes, please provide further information.

Pilot Data template filled by each FAME Pilot Leader

Pilot Templates should be filled once by each respective pilot leader, replicating the template as many times required, with additional information asked regarding the responsible for determining each pilot's collected/generated data.

Section	Description
ID	
Title	

Description	
Owner	
Licence / Privacy	
Data type	
Type of process (stream or static data)	
Data format	
Data store	
Data Security	
Regulatory Constraint Requirements	