



# **D7.1: Data Management Plan**



This project has received funding from the European Union's Horizon Research and Innovation Actions under Grant Agreement N° 101093216.

<b>Title:</b>	<b>Document version:</b>
D7.1: Data Management Plan	1.0

<b>Project number:</b>	<b>Project Acronym</b>	<b>Project Title</b>
101093216	UPCAST Project	UPCAST Project

<b>Contractual Delivery Date:</b>	<b>Actual Delivery Date:</b>	<b>Deliverable Type*-Security*:</b>
M6 (June 2023)	M8 (August 2023)	Report -Confidential

<b>Responsible:</b>	<b>Organization:</b>	<b>Contributing WP:</b>
Richard Stevens	IDC	WP7

<b>Authors (organization):</b>
Richard Stevens (IDC)
George Konstantinidis (SOT)
Luis-Daniel Ibáñez (SOT)
Nevena Raczko (IDC)
All Partners managing data (All)

**Abstract:**

The UPCAST project is a pioneering endeavour that enhances data sharing across diverse sectors providing plugins to increase efficiency and effectiveness of data marketplaces through a unified platform. Leveraging advanced AI and data management technologies, the plugins address complex challenges in digital marketing, healthcare, public administration, and genomics research. The project integrates cutting-edge tools to facilitate seamless data discovery, processing, privacy enforcement, pricing, and environmental impact assessment. UPCAST promotes open science, gender neutrality, and adheres to ethical AI principles. With a comprehensive data governance structure, it optimizes data utilization while ensuring privacy and compliance. By fostering cross-sector collaboration, UPCAST accelerates innovation and empowers decision-making for a data-driven future.

The Upcast Data Management Plan outlines comprehensive strategies for data collection, sharing, and protection within the collaborative project. It establishes clear procedures for data sharing agreements, access controls, and authentication mechanisms among partners and external stakeholders. The plan emphasizes compliance with data protection regulations, including GDPR, through robust security measures, encryption during storage and transfer. Ethical and legal implications are diligently considered, ensuring responsible data handling. The plan also defines retention periods, secure data disposal methods, and a structured review process to adapt to evolving regulations. Overall, the plan underscores the commitment to ethical, secure, and transparent data management practices within the Upcast project.

## Keywords:

Data Management Plan, Data Governance, Data Acquisition, Data Storage, Data Quality, Data Privacy, Data Integration, Data Sharing, Data Security, Data Lifecycle

---

## REVISION HISTORY

Revision:	Date:	Description:	Author (Organization)
V0.1	01.04.2023	Outline and Table of Contents	Richard Stevens (IDC)
V0.2	12.05.2023	Content added to general project wide sections	Richard Stevens (IDC) George George Konstantinidis Luis-Daniel Ibáñez (SOT) Sofaklis Efremidis (MAG)
V0.3	24.05.2023	Pilot specific data and Data management described	Milan Vukovic, (NIS) Nenad Stojanovic (NIS) Paraskevi Tarani (MDAT) Fernando Persles (JOT) Olga Papadodima (NHRF), Georgios V. Lioudakis (ABO) Alexandros Lemperos (CAC)
V0.4	20.06.2023	General comments and detail added	All Partners
V0.5	08.07.2023	Peer Review comments integrated	Sofoklis Efremidis (MAG)
V1.0	25.07.2023	Release Candidate	Richard Stevens (IDC)
V1.x	?	Considering Amendment	TBD



This project has received funding from the European Union's Horizon Research and Innovation Actions under Grant Agreement N° 101093216.

More information available at <https://upcastproject.eu/>

## COPYRIGHT STATEMENT

The work and information provided in this document reflects the opinion of the authors and the UPCAST Project consortium and does not necessarily reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information it contains. This document and its content are property of the UPCAST Project Consortium. All rights related to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the UPCAST Project Consortium and are not to be disclosed externally without prior written consent from the UPCAST Project Partners. Each UPCAST Project Partner may use this document in conformity with the UPCAST Project Consortium Grant Agreement provisions.

# INDEX

<b>1</b>	<b>INTRODUCTION .....</b>	<b>7</b>
<b>2</b>	<b>DATA SUMMARY .....</b>	<b>8</b>
2.2	Reuse of data.....	8
2.2.1	Purpose and Contribution.....	9
2.2.2	Data Utility Beyond the Project.....	9
2.3	Operational Data.....	10
2.3.1	Performance and Operation of Systems.....	10
2.3.2	Operational Communication and Collaboration Data .....	10
2.4	User Data .....	11
2.5	Economic Data .....	11
2.6	Social Data.....	12
2.7	Performance Data .....	12
2.8	Regulatory and Compliance Data.....	13
<b>3</b>	<b>FAIR DATA .....</b>	<b>13</b>
3.1	Making data findable, including provisions for metadata .....	14
3.2	Making data accessible .....	14
3.3	Making data interoperable.....	15
3.4	Increasing data reusability.....	15
<b>4</b>	<b>OTHER RESEARCH OUTPUTS .....</b>	<b>15</b>
<b>5</b>	<b>DATA GOVERNANCE AND ROLES.....</b>	<b>16</b>
5.1	Data Steward .....	16
5.2	Data Board .....	17
5.3	Data Custodians .....	17
5.4	Data Users .....	17
5.5	Data Ethics Officer .....	18
<b>6</b>	<b>DATA COLLECTION AND PROCESSING FOR OPERATIONAL PURPOSES .....</b>	<b>18</b>
6.1	Data Processing for dissemination purposes .....	18
6.2	Data Management in Transferability and Training.....	19
6.3	Surveys.....	19
6.3.1	Survey Data Management Processes .....	20
6.3.2	Survey Data Security Measures:.....	20
6.3.3	Survey GDPR Compliance:.....	20
<b>7</b>	<b>DATA MANAGEMENT FOR UPCAST PLUGINS.....</b>	<b>21</b>
7.1	Resource Specification .....	21
7.2	Resource Discovery.....	21
7.3	Data Processing Workflow .....	22
7.4	Data Pricing .....	22
7.5	Environmental Impact.....	24
7.6	Privacy and Usage constraints.....	25
7.7	Negotiation .....	25
7.8	Data Integration.....	26
7.9	Federated Machine Learning.....	27
7.10	Safe and Secure Exchange .....	28
7.11	Monitoring.....	28
<b>8</b>	<b>DATA MANAGEMENT IN THE CONTEXT OF PILOTS.....</b>	<b>28</b>
8.1	Pilot Case 1.1: Digital Marketing Data and Resources (JOT).....	28

8.2	Pilot Case 1.2: Digital Marketing Data and Resources (Cactus)	30
8.3	Pilot Case 2: Biomedical and Genomic Data Sharing (NHRF-ICTAbovo)	31
8.4	Pilot Case 3: Sharing Public Administration for Climate across Thessaloniki Cities (MDAT+OKFGR)	33
8.5	Pilot Case 4: Health and Fitness Data Trading (Nissatech)	34
<b>9</b>	<b>DATA SHARING AND ACCESS</b>	<b>35</b>
9.1	Data Sharing Agreements between Partners	35
9.2	Access Controls and Authentication Mechanisms	36
9.3	Data Dissemination and Publication Policies	36
9.4	Intellectual Property Rights and Data Ownership Considerations	36
<b>10</b>	<b>DATA SECURITY AND PRIVACY</b>	<b>36</b>
10.1	Security Measures	37
10.1.1	Data Encryption during Storage and Transfer	37
10.1.2	Access Controls and User Authentication	37
10.2	Privacy	37
10.2.1	Compliance with Relevant Data Protection Regulations	37
10.2.2	Data Breach Response and Incident Management Procedures	37
<b>11</b>	<b>DATA RETENTION AND DISPOSAL</b>	<b>37</b>
<b>12</b>	<b>DATA MANAGEMENT PLAN REVIEW</b>	<b>38</b>
<b>13</b>	<b>CONCLUSION</b>	<b>38</b>
	<b>ANNEX I ACRONYMS &amp; ABBREVIATIONS</b>	<b>40</b>
	Acronyms	40
	Abbreviations	40

## LIST OF FIGURES

Figure 1	Newsletter subscription form	19
----------	------------------------------	----

## LIST OF TABLES

<b>Table 1</b>	<b>Acronyms</b>	<b>40</b>
Table 2	Abbreviations	41

# 1 Introduction

The purpose of this Data Management Plan (DMP) is to outline the strategies and procedures for the management, sharing, and protection of data within the collaborative European Horizon Europe **Upcast** project. The project aims to develop plugins that will improve efficiency and effectiveness of Data Exchange Platform like Data Marketplaces and their piloting it in the following sectors: Public Administrations, Healthcare and healthcare data, sports data of individuals' training, and the use of public web search data regarding the media sector. Data, managing data and taking advantage of data is a central part of the Upcast Project. The Upcast DMP lays out comprehensive strategies for the FAIR (Findable, Accessible, Interoperable, Reusable) collection, sharing, protection, and ethical handling of data and is the consortium's strategic data management roadmap, outlining the project's commitment to adhering to the highest standards of data integrity, security, and privacy. It is not simply a required deliverable but a vital enabler of project success. It should be mentioned here that the project is undergoing expansion and new partners and new Data Marketplaces are poised to begin integrating their data and perform transactions with partner's data and this document is expected to be a living document and to be updated as the various project partners access, manipulate, transfer and store additional new types of data or employ new techniques to manage that data. The Upcast DMP is therefore not confined to a static vision but embodies a document that is expected to evolve with the project. Data management follows a structured review process is a special session of project consortium meetings ensuring its relevance and effectiveness throughout the project's lifecycle. This forward-thinking approach empowers the project to promptly adapt to changing regulations, emerging technologies, and unforeseen ethical and legal challenges. In particular, this document will be reappraised and amended where applicable with the entrance of any new partner or adaptation to any new data marketplace in the consortium.

With the Upcast project focusing on data exchange scenarios spanning diverse sectors such as digital marketing, biomedical research, climate data sharing, and health and fitness analytics, the DMP takes a tailored approach to address the unique challenges posed by each pilot case. Each pilot case necessitates specific data types, sharing agreements, access controls, and ethical considerations, and the DMP intricately navigates through these variables to provide a unified yet adaptable framework.

In the subsequent sections of the DMP, we delve into the core components that constitute the backbone of the single plugins or technical components and the data management of the pilot scenario implementations. We address the data sharing agreements meticulously established between project partners and potential external stakeholders. The DMP meticulously outlines the access controls and authentication mechanisms that safeguard the data's confidentiality and integrity.

Data dissemination and publication policies are scrutinized, ensuring that our commitment to ethical principles remains steadfast. The DMP delves into the intricate web of intellectual property rights and data ownership considerations, paving the way for a harmonious collaboration where innovation is balanced with equitable recognition.

The paramount importance of data security and privacy is given the spotlight, with the DMP detailing encryption during storage and transfer, stringent access controls, and adherence to stringent data protection regulations such as GDPR. In an age where data

breaches are a looming threat, the Upcast DMP offers a scheme to our project's data assets.

Ethical and legal implications are critically evaluated, considering the potential risks and benefits of data exchange and management. We approach this aspect with a keen eye on compliance with applicable regulations and guidelines, upholding our commitment to responsible research practices.

Lastly, the DMP concludes with an emphasis on data retention and disposal, delineating the retention periods for different data types and outlining secure methods for data anonymization or destruction. As we tread the path of innovation, we are equally steadfast in our commitment to protecting privacy and respecting data subjects' rights.

In essence, the Upcast DMP is a realistic assurance of our intention to responsibly and legally perform data management today and modify the approach in the future. It is not just a roadmap; it's a manual to ensure ethical, secure, and transparent approach to navigating the complex terrain of data exchange and innovation in a multi-partner collaborative Horizon Europe project where significant amounts and types of data are exchanged.

## 2 Data Summary

This section is produced along the guidelines for data management plans published by the European Commission<sup>1</sup>. The section gives an overview of the data used in the project including expected reuse of data, types of data, provenance and sharing of the data and the volume of each of the types of data used in the project. For the convention of this DMP generic descriptions of the size of the datasets in question are defined as "small," "medium," and "large". **Small Data Sets** have a relatively low volume, measured in megabytes (MB) or gigabytes (GB). These data sets are manageable using common software tools and can often be analyzed on a single machine. Small data sets will be used in prototyping the plugins where the data can be easily loaded into memory for processing. **Medium Data Sets**: Medium data sets are expected in Upcast during final demonstration and are larger in volume than small data sets but are still manageable for analysis on a single machine. They can range from a few gigabytes to tens or hundreds of gigabytes. While they may require more computational resources than small data sets, Upcast may require processing data using distributed computing frameworks or cloud services. This will change the security and privacy as well as access control and need or particular attention to data management policies and the introduction of new actors in this DMP should external services be required for data management. Upcast currently doesn't consider applicable the use of **Large Data Sets** consisting of several terabytes (TB) or petabytes (PB) (even though the genomics pilot could potentially evolve in this direction). As these would require distributed computing frameworks, and additional analytical frameworks such as Apache Hadoop or Apache Spark, to efficiently process and analyze the data which would introduce new data management requirements this data plan will be amended to establish appropriate processes if the case arises.

### 2.2 Reuse of data

---

<sup>1</sup>OpenAIRE. (2021). Data Management Plan Template Version 1.0 05 May 2021. Retrieved from [https://www.openaire.eu/images/Guides/HORIZON\\_EUROPE\\_Data-Management-Plan-Template.pdf](https://www.openaire.eu/images/Guides/HORIZON_EUROPE_Data-Management-Plan-Template.pdf)

In the context of the upcast project where data is expected to be shared and monetised in Data Marketplaces and enrich pilot specific capabilities the reuse of data and existing datasets is of central importance. The value that the pilots can achieve is of course based on the focus and specificity of data they can bring into their operations and simply using the data they generate will lead to business as usual. Acquiring and reusing data from public sources and from clients and data marketplaces is where the expected value add can be achieved, thus data reuse is of core importance to the plugins described in Chapter 7 and the pilots described in chapter 8 of this document. The project acknowledges the potential of leveraging publicly available datasets and domain-specific resources relevant to the project's focus areas. While specific datasets considered for reuse were evaluated, some were not chosen due to data quality concerns or scope mismatch. However, select publicly available datasets and research project datasets and results that align with the project's research questions will be reused for comparative analysis, benchmarking, and contextualization of the project's findings.

### 2.2.1 Purpose and Contribution

The reuse of existing data sources and generated datasets serves to validate the project's hypotheses, enhance predictive models, and facilitate cross-validation. The reuse of contextual data from projects including for example (others are currently being identified and will be added here as they become known): Datamarket Austria, DataPorts i3 Market will enable the project to benchmark its results against established baselines and industry standards.

- **Data Market Austria (DMA)**<sup>2</sup>: This is a national initiative that aims to establish a data-driven innovation ecosystem in Austria. It involves multiple stakeholders from academia, industry, public sector and civil society. It also supports the development of data products and services, as well as data literacy and skills.
- **DataPorts**<sup>3</sup>: This is a Horizon 2020 project that focuses on creating a data platform for cognitive ports, which are ports that use data and AI to optimize their operations and logistics. The project involves 13 partners from 7 countries, and covers various aspects of data sharing, such as governance, security, interoperability and monetization.
- **i3-MARKET**<sup>4</sup>: This is another Horizon 2020 project that aims to create an integrated, interoperable and intelligent data marketplace for industrial data. The project involves 15 partners from 9 countries, and addresses the challenges of data quality, trust, privacy and value creation. You can find their research data on their website:

Where publicly available and suited to the needs of Upcast we will use these and other projects data sources and ensure compliance to this DMP and eventual backwards rights and obligations from the data sources. This reuse enables the project to contextualize its findings in real-world scenarios of these initiatives and compare to existing research, enhancing the practical applicability of its outcomes.

### 2.2.2 Data Utility Beyond the Project

---

<sup>2</sup> <https://www.ait.ac.at/en/research-topics/datascienceartificialintelligence/projects/dma/>

<sup>3</sup> <https://dataports-project.eu/>

<sup>4</sup> <https://www.i3-market.eu/>

The data reused and generated by Upcast hold potential utility beyond the project's immediate scope. Researchers, professionals, and stakeholders in domains beyond the scenarios in upcast (Public Administrations, Healthcare and healthcare data, sports data of individuals' training, and the use of public web search data regarding the media sector) can benefit from the availability of these datasets. The data's utility lies in applications such as validation of industry practices, informed decision-making, cross-disciplinary research. By making these datasets accessible and reusable, Upcast aims to contribute to broader advancements in additional fields beyond those originally expected in our Description of Action and allow other research projects to leverage our findings thus enriching the research communities understanding of the Data Marketplace domain.

Overall, the strategic reuse of data aligns with the project's commitment to efficient resource utilization, robust analyses, and meaningful impact. The incorporation of both generated and reused data enhances the project's credibility, enriches its analyses, and strengthens its contribution to the field.

## 2.3 Operational Data

### 2.3.1 Performance and Operation of Systems

Data from the performance and operation of systems or processes involved in the pilot cases described in Chapter 7. This can include sensor data, operational logs, or real-time monitoring data. Useful internally to measure the advance of the project and to detect any problems that could be corrected. Not meant to be shared externally.

**Operational Data** regarding **Performance and Operation of Systems** is expected to be **retired** and **deleted** where it is shared across several partners and when it is no longer needed in the strict sense of the execution of the Description of Action. However the primary data owners that collected the data may choose to retain the data if it was generated in their organisations and may do so according to their internal data management policies during or after the Upcast project's termination.

The **amount of Performance and Operation of Systems** to be collected is expected to be **medium** (potentially large in post project phase deployment).

N.B. Description of specific datasets for operational aspects and for the development of each UPGAST plugin are described in sections 6, 7 and 8.

### 2.3.2 Operational Communication and Collaboration Data

Data related to communication and collaboration among project stakeholders during the pilot case.

This can include meeting minutes, communication logs, or collaboration platforms' data.

- A. Public administration data
- B. Healthcare and healthcare data
- C. Sports data of individuals training
- D. Public web search data regarding the Media sector

All **Operational Data** regarding **Communication and Collaboration** is expected to be **retired** and **deleted** when it is no longer needed in the strict sense of the execution of the Description of Action.

The **amount of Communication and Collaboration Data** to be collected is expected to be **medium** as some video recordings, webinars and manuals are expected to be produced.

## 2.4 User Data

Data related to user behavior, preferences, or feedback is essentially collected during the pilot cases. Information on user interactions, satisfaction levels, or user experience metrics. It is useful internally to measure KPIs and the effectiveness of our approaches and useful externally after proper anonymisation to research community. In the Upcast project and according to this DMP, the management of user data from the pilots is meticulously designed to ensure compliance with the General Data Protection Regulation (GDPR) while preserving user privacy, data anonymization, safety, and adhering to the FAIR principles (Findable, Accessible, Interoperable, Reusable) described in chapter 3.

In terms of **GDPR Compliance and User Privacy**, User data coming from the four pilots or even perhaps used in development phases of the plugins is handled with the utmost respect for GDPR guidelines. This involves obtaining explicit and informed consent from users for data collection and processing during development and testing phases even if no external use of the system is planned. In any commercial use or after the end of the project any personal data will be retired and removed or where deemed essential for the continuing execution of the system the same rules and procedures would apply. Users are provided with clear information about the purpose of data collection, the types of data involved, and their rights regarding their data.

In terms of **Data Anonymization and Safety** the upcast project and this DMP have a key focus on protecting user privacy, any personally identifiable information (PII) is removed or irreversibly anonymized. Identifiers such as names, contact details, and other sensitive information are eliminated. This ensures that individual users cannot be directly identified from the data.

In terms of **Data Encryption and Access Controls**, User Data is stored and transferred using encryption protocols to prevent unauthorized access. IPsec is expected to be used as a minimum to establish secure connections between networks or between devices used in upcast and the plugins or other network resources. Where information is collected via webforms (or similar) Transport Layer Security (TLS) is expected to be used to secure data during transmission over upcast networks or between plugins and the user interfaces to ensure data integrity and confidentiality by encrypting the communication between the sender and the plugins or other Upcast services. Access to the data is restricted to authorized personnel only. Strong access controls are being studied and will be implemented to ensure that only individuals with legitimate reasons and proper permissions can access the data.

All **user data** is expected to be **retired** and **deleted** when it is no longer needed in the strict sense of the execution of the Description of Action.

The **amount of User Data** to be collected is expected to be **small** or **very small**.

## 2.5 Economic Data

Data on economic factors and indicators relevant to the pilot case. This can include cost data, financial metrics, or economic impact assessments. This data is collected primarily in the pilot scenario implementations as the organisations involved in the pilots are already conscious of the costs for performing the data operations that are expected

to be performed during the pilot executions. It is clear that the pilot users are interested in a comparison of the data obtained during the execution using Upcast platform and plugins to ascertain the value of using upcast. The project itself will be interested in this data as it will lead to the impact assessment and sustainability planning. It will be internally in the pilot organisations for measuring KPIs and designing exploitation plans and this data management will be performed according to their customary practices and in line with existing legislation (GDPR et. al.). However it will be useful in the wider consortium for drawing up joint exploitation plans and externally for organisations interested in replicating results. In the case that data regarding economic aspects tied to the project is in fact transferred outside the single participating organisations this DMP will apply and where anonymous metadata is not sufficient all data is expected to anonymized.

All **Economic data** held in project repositories is expected to be **retired** and **deleted** when it is no longer needed in the strict sense of the execution of the Description of Action.

The expected **amount of Economic Data** to be collected is expected to be **small**.

## 2.6 Social Data

Being a European project, much data is expected to be collected to allow the dissemination and divulgation of research results. Not only community information regarding interested individuals and organisations is expected to be collected but also data related to social aspects of the pilot cases, such as social acceptance, social behavior, or community engagement. Information on public perception, stakeholder involvement, or social impact assessments. This will be useful for inviting stakeholders to engage with the project, attend events and webinars receive updates and information but also internally for measuring KPIs and designing exploitation plans. It will be useful externally for organisations interested in replicating results. Data will be collected and maintained in encrypted format on the coordinators internal protected information systems and all data collection will be pursuant to data subjects informed consent.

All **Social data** held in project repositories is expected to be **retired** and **deleted** when it is no longer needed in the strict sense of the execution of the Description of Action.

The expected **amount of Social Data** to be collected is expected to be **small**.

## 2.7 Performance Data

Collecting Performance Data is a clear objective of the project. To understand and demonstrate that the Upcast approach can drive improved use and generate additional income or positive effects like energy efficiency is the central focus of several assessment tasks. Task **T5.4 Evaluation and assessment** will directly measure data against baseline information for each of the plugins and Data on the performance of specific improvements of the current processes in the pilot cases. Task **T7.4 Exploitation and Sustainability** will use the performance data collected during development and integration and piloting to assess the real market facing viability of the upcast approach. Measurements, benchmarks, and comparative analysis with baseline and competing systems will rely heavily on performance metrics and data collected. All data will be used internally to measure KPIs and the effectiveness of our approaches. Data is expected to be stored local and only metadata containing no personal information is expected to be shared beyond or outside the pilot case premises and where data is exchanged it is expected to be summary data or metadata. Where used

externally proper cleaning and anonymisation will be performed before sharing it with the research and practitioner community.

**Performance Data** is expected to be **retired** and **deleted** where it is shared across several partners and when it is no longer needed in the strict sense of the execution of the Description of Action. However, the primary data owners that collected the data may choose to retain the data if it was generated in their organisations and may do so according to their internal data management policies during or after the Upcast project's termination.

The Performance **Data** to be collected is expected to be **medium**.

## 2.8 Regulatory and Compliance Data

Some data related to regulatory requirements and compliance will be collected in the context of the automated negotiation plugin developed in task **T3.1 Contract Negotiation Module** to internally to measure the advance of the tool in question and to detect any problems that could be corrected. It is also expected to be used in the demonstration of all of the pilot case in task **T5.2 Implementation**. Information on contracts, licenses, agreements legal compliance or regulatory standards that have been considered or that need to be met may be collected. This information will be used internally for measuring KPIs but also in the sense of task **T7.4 Exploitation and Sustainability** and designing exploitation plans. It may be used externally in publications for organisations interested in replicating results otherwise it is not meant to be shared externally.

Regulatory and Compliance Data is expected to be **retired** and **deleted** where it is shared across several partners and when it is no longer needed in the strict sense of the execution of the Description of Action. However, the primary data owners that collected the data may choose to retain the data if it was generated in their organisations and may do so according to their internal data management policies during or after the Upcast project's termination.

The **Regulatory** and **Compliance Data** to be collected is expected to be **small**.

## 3 FAIR Data

FAIR data principles ensure digital assets are Findable, Accessible, Interoperable, and Reusable<sup>5</sup>[1]. This framework enhances data sharing, integration, and lasting value across research contexts. In the Upcast Data Management Plan, the FAIR principles play a pivotal role in ensuring the effective management, sharing, and utilization of data. Data generated from the pilots and plugins will adhere to these principles to maximize their value and impact. Data will be made Findable through metadata tagging, indexing, and storage in searchable repositories. Access to the data will be controlled, and authentication mechanisms will be established to ensure Accessibility while respecting data ownership and privacy concerns. Interoperability will be ensured through adherence to standardized formats and use of established ontologies and vocabularies, promoting seamless integration with other datasets and supporting effective data exchange. Additionally, data will be structured and described consistently to facilitate its broad

---

<sup>5</sup> FAIR data principles for Horizon Europe is described in the Open Research Data and Data Management Plans document produced by the European Research Council retrieved at: [https://erc.europa.eu/sites/default/files/document/file/ERC\\_info\\_document-Open\\_Research\\_Data\\_and\\_Data\\_Management\\_Plans.pdf](https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf)

Reusability across different research contexts. These goals are defined in the following paragraphs:

FAIR data principles ensure digital assets are Findable, Accessible, Interoperable, and Reusable<sup>6</sup>. This framework enhances data sharing, integration, and lasting value across research contexts. In the Upcast Data Management Plan, the FAIR principles play a pivotal role in ensuring the effective management, sharing, and utilization of data. Data generated from the pilots and plugins will adhere to these principles to maximize their value and impact. Data will be made Findable through metadata tagging, indexing, and storage in searchable repositories. Access to the data will be controlled, and authentication mechanisms will be established to ensure Accessibility while respecting data ownership and privacy concerns. Interoperability will be ensured through adherence to standardized formats and use of established ontologies and vocabularies, promoting seamless integration with other datasets and supporting effective data exchange. Additionally, data will be structured and described consistently to facilitate its broad Reusability across different research contexts. These goals are defined in the following paragraphs:

### **3.1 Making data findable, including provisions for metadata**

All research data meant to be published will be identified by a persistent identifier. Rich metadata will be provided to allow discovery using the Data Catalog Vocabulary (DCAT)<sup>7</sup> which is a metadata standard used to describe datasets, data catalogs, and data portals in a structured way, making it easier to discover, understand, and manage data resources. DCAT provides a standardized way to represent essential information about datasets, such as their title, description, keywords, distribution formats, access rights, and more. By using a common vocabulary and format, different data catalogs and portals can share and exchange information about their datasets in a consistent manner. UPCAST advances state of the art on data management and annotation and not in the specific research themes of the pilots, hence, no need to use domain specific vocabularies for data generated for research.

Search keywords will be provided, metadata will be encoded in RDF using DCAT to ensure it can be harvested and indexed.

### **3.2 Making data accessible**

All research data meant to be published will be openly available deposited in Zenodo, that ensures data is assigned an identifier that resolves to a digital object. Metadata will be made openly available and licensed under a public domain dedication CC0. It is guaranteed to remain available and findable after the project for the lifetime of the Zenodo repository (>20years according to current policy)

Data that will not be made openly available is described in chapters 7 and 8 and is essentially operational data and performance data from development phases. Refer to specific sections for further details.

---

<sup>6</sup> FAIR data principles for Horizon Europe is described in the Open Research Data and Data Management Plans document produced by the European Research Council retrieved at: [https://erc.europa.eu/sites/default/files/document/file/ERC\\_info\\_document-Open\\_Research\\_Data\\_and\\_Data\\_Management\\_Plans.pdf](https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf)

<sup>7</sup> <https://www.w3.org/TR/vocab-dcat-3/>

Data generated during the project that is linked to existing proprietary data by any of the partners and/or relevant for further commercial exploitation of the project. This data will be shared among consortium partners for the duration of the project. Future access will be governed by exploitation considerations, this document will be updated accordingly to reflect consortium decisions in that matter.

Some data used in the context of the pilots for demonstration of the UPGAST plugins will be published where scientific or operational results may be useful for the wider scientific and research communities as benchmarks for ensuing research.

### 3.3 Making data interoperable

We will use DCAT for describing research outputs. DCAT is particularly relevant in the context of interoperability for data and processes because it helps improve the findability and accessibility of data across different platforms and organizations. This promotes interoperability between data systems and enhances the ability to search for and reuse datasets. We will follow the EOSC interoperability framework<sup>8</sup> whenever sensible.

We also note the fact that the UPGAST project intends to be an enabler of interoperability for the purpose of data exchanges in data spaces, as such, it will contribute with a vocabulary to generate resource descriptions. The UPGAST vocabulary will re-use and extend DCAT and the IDSA information model, keeping close communication with Data Spaces support center and related standardisation initiatives. UPGAST vocabulary will be released with an open license.

### 3.4 Increasing data reusability

For each research output meant to be published. Documentation will be provided to facilitate re-use. At a minimum, a readme and a data dictionary. Additional documentation will depend on the type of output dataset. Licenses of these outputs will permit as wide-reuse as possible considering background knowledge licensing and exploitation intentions of the intellectual property owners, however the intention is to widely publish data and results on public repositories as described in section 3.2.

In terms of outputs not meant for publication, i.e., linked to existing intellectual property or susceptible of exploitation by consortium partners, this document will be updated to refer to the appropriate sections of the exploitation plan (D7.4)

## 4 Other research outputs

Each UPGAST plugin produces software and model outputs. As with datasets, some of them are research outputs meant to be released openly, while others are tied to existing intellectual property or may be subject to future exploitation plans. Refer to chapter 6 for details on each plugin. It is evident at this stage of development that the consortium does not completely understand the content

UPGAST Plugin	Non-dataset outputs (License)
Resource Specification	<ul style="list-style-type: none"> <li>Implementation of resource specification to be released with an Open-Source license (TBD after exploitation plan).</li> </ul>

<sup>8</sup> <https://op.europa.eu/en/publication-detail/-/publication/d787ea54-6a87-11eb-aeb5-01aa75ed71a1/language-en/format-PDF/source-190308283>

	<ul style="list-style-type: none"> <li>Library of algorithms for data profiling (Open source, precise license TBD at after exploitation plan)</li> </ul>
<b>Resource Discovery</b>	<ul style="list-style-type: none"> <li>Implementation of dataset search algorithms, to be released with an Open-Source license (TBD after exploitation plan).</li> <li>Implementation of dataset discovery algorithms, to be released with Open-Source license (TBD after exploitation plan)</li> </ul>
<b>Data Processing Workflow</b>	<ul style="list-style-type: none"> <li>Implementation of Data Processing Workflow execution to be released under an Open-Source license (TBD after exploitation plan).</li> <li>Data Processing Workflow modelling is based on commercially licensed software, precise license TBD after exploitation plan</li> </ul>
<b>Data Pricing</b>	<ul style="list-style-type: none"> <li>Web Service to make price suggestions of datasets (Commercial license, TBD after exploitation plan)</li> </ul>
<b>Environmental Impact</b>	<ul style="list-style-type: none"> <li>Energy profiler software and associated ML models (Commercial licence TBD after exploitation plan)</li> </ul>
<b>Privacy and Usage</b>	<ul style="list-style-type: none"> <li>Evaluation engine for privacy and usage constraints. (License TBD after exploitation plan due to combination of Open Source and Commercial software)</li> </ul>
<b>Negotiation</b>	<ul style="list-style-type: none"> <li>Negotiation engine (Commercial licence TBD after exploitation plan)</li> </ul>
<b>Data Integration</b>	<ul style="list-style-type: none"> <li>Source to target mappings (Open Source licence TBD after exploitation plan)</li> </ul>
<b>Federated Machine Learning</b>	<ul style="list-style-type: none"> <li>TBD after amendment</li> </ul>
<b>Safe and Secure Exchange</b>	<ul style="list-style-type: none"> <li>TBD after amendment</li> </ul>
<b>Monitoring</b>	<ul style="list-style-type: none"> <li>Monitoring engine based on proprietary MIRA platform (Commercial license TBD after exploitation plan)</li> </ul>

## 5 Data Governance and Roles

This section states the roles and responsibilities of the project partners in terms of data governance and management. The upcast project Identify the lead partner responsible for overall data coordination and specify the roles of other partners involved.

In Upcast these roles are defined:

### 5.1 Data Steward

The Project Manager is ultimately responsible for overseeing the data governance strategy and implementation within the Upcast project and is named the Data Steward.

The Data Steward will monitor all data use through continuous dialog with the project Technical Coordinator, Scientific Coordinator and Workpackage Leaders. The coordinator in his role of Data Steward will be responsible for informing the Data Board and chairing that body. All decisions regarding the use of data, ethical and legal implications that affect the project will be logged and documented by the data steward in the project board by the Steward and documented in the meeting minutes. The data Steward will act as a point of contact for data-related queries. All follow-up activities that need to be performed thereafter will be in the remit of the Data Steward until the issue or decision has been retired. The Data Steward will avail of the coordinator's (IDC) organisational **Data Protection Officer** (DPO) that has a standing position and is responsible ensuring compliance with data protection regulations, such as the General Data Protection Regulation (GDPR) or other applicable privacy laws. IDC's Data Privacy Officer will provide guidance on privacy requirements and address privacy concerns raised during the project report to the coordinator and attend project data board meetings where legal issues have arisen.

## 5.2 Data Board

The Data Board is made up of the project coordinator, the technical coordinator, the scientific coordinator and each of the workpackage leaders. The project board will meet at each of the project consortium meetings and have a predefined slot in the agenda. The topics covered and the duration of the session will vary depending on the evolution of data related issues to be discussed or decided. All consortium members are invited to follow the proceedings of the meetings and the session may be limited to approval of continuing in the current course of action if no data related issues have been raised in the period since the last meeting. The board's role is first to set data governance policies, confirm this data management plan and the roles defined herewith. Additionally, the board will proactively communicate with workpackage leaders to ascertain if data issues have arisen during the task level development or assessment activities. The resolve data-related issues and ensure compliance with relevant regulations. The principal remit of this board is to ensure that Upcast has an effective data management practice and ensures that the project and its partners are maintaining the quality, integrity, and security of the project's data. The Data Board will define data standards, enforce data governance policies and facilitate data sharing. The board will propose and validate the project's data architecture. Describing the data models, data flows, integration points across various modules and tools within the Upcast project and ensure data interoperability, scalability, and adherence to industry best practices.

## 5.3 Data Custodians

The Workpackage leaders will identify a Data Custodians in each of organizations involved in development of plugins or tools that access or use data or during the deployment phase in work package 5. These individuals that currently exist in each of the organisations will be confirmed and informed of their obligations in relation to the projects collective activities and will be responsible for the safekeeping and storage of the project's data during development of plugins described in chapter 7 or deployment described in chapter 8. The workpackage leader will ensure that appropriate data security is in place in these organisations and that local teams implement access controls, perform backups, and maintain data availability and have appropriate mechanisms safeguarding data against unauthorized access or loss.

## 5.4 Data Users

Data users are individuals or teams who access and utilize project data for development, analysis, deployment, assessment and decision-making purposes. They are potentially every partner in the project and thus the Data Board will inform the project members as a collective of this data management plan, ensure that all partners are well informed of the policies and data management practices. Data Users will be assigned appropriate access levels, acknowledge data usage guidelines, and be individually responsible for compliance with these data management practices.

## 5.5 Data Ethics Officer

The project Scientific Coordinator is named the Data Ethics Officer and will review ethical considerations associated with data collection, storage, and usage within the Upcast project. The Data Ethics Officer in Upcast will ensure the project adheres to best of breed ethical data management including safeguarding privacy, ensuring fairness, and transparency. The Data Ethics Officer monitors issues including bias, consent, ownership, and compliance. This role promotes data integrity, user protection, and responsible practices throughout the project lifecycle. The Data Ethics Officer will intercept any issues that arise and bring them to the attention of the Data Board. Where decisions are taken the Data Ethics Officer will collaborate with the Data Steward to ensure appropriate measures are undertaken and follow any process until the issue is retired.

# 6 Data Collection and Processing for Operational Purposes

In this section we describe the data collection and processing of data required for non-technical tasks: Dissemination, transferability and exploitation.

## 6.1 Data Processing for dissemination purposes

Data Processing for dissemination purposes is expected in task **T6.1 Dissemination and communication**. In particular we will collect information and data regarding individuals that are interested in being kept informed of the activities in the Upcast project and expect to communicate with them through a periodic newsletter. The data collected however will not be managed directly by Upcast partners but through a third party application that is certified and has strong security and access control and is deemed completely GDPR compliant.

We use the MailChimp<sup>9</sup> platform for our newsletter distribution. MailChimp enables us to effectively communicate project updates, milestones, and valuable insights to our subscribers, fostering a strong and engaged community around our initiative. The newsletter serves as a crucial channel for sharing our progress, research findings, and upcoming events, allowing us to reach a broader audience and keep them informed about the project's developments.

However, it is important to note that as part of this process, we collect certain personal data from our subscribers, such as names and email addresses in order to facilitate newsletter delivery and ensure relevant content reaches our audience. This information is voluntarily provided by subscribers when they sign up to receive our newsletters, and we use it solely for the purpose of newsletter distribution and related communications.

---

<sup>9</sup> <https://mailchimp.com/>

**Subscribe to our newsletter**

\* indicates required

Email Address \*

Full Name \*

Company \*

UPCAST Project is GDPR compliant  
Please read our [Privacy Policy](#) before accepting the form.

I have read and agree to the [Privacy Policy](#) of the UPGAST Project.  
You can unsubscribe at any time by clicking the [link](#) .

We use Mailchimp as our marketing platform. By clicking below to subscribe, you acknowledge that your information will be transferred to Mailchimp for processing. [Learn more about Mailchimp's privacy practices here.](#)

[Subscribe](#)


made with  **mailchimp**

Figure 1 Newsletter subscription form

Mailchimp's security measures, such as encryption during data transmission and secure data storage, give us confidence that our subscribers' information is well-protected against unauthorized access and breaches.

Additionally, Mailchimp allows us to provide our subscribers with choices and control over their data. We ensure transparency by outlining our data handling procedures in our privacy policy<sup>10</sup>, which is easily accessible on our website.). Our data handling practices are in full compliance with applicable data protection regulations, such as the General Data Protection Regulation (GDPR). Subscribers can easily manage their preferences and unsubscribe from our newsletter whenever they wish, giving them autonomy over their data.

The latest 6 months after the end of the project all personal data will be deleted from the MailChimp servers.

## 6.2 Data Management in Transferability and Training

Task T6.3 Transferability and Training a GDPR compliant platform Moodle will be used to host online training. Moodle offers privacy and data protection features, such as user consent management, data retention policies, and privacy settings. Upcast will ensure that it configures and uses Moodle in alignment with GDPR requirements, including obtaining proper user consent, providing clear privacy notices, enabling user data access and deletion, and maintaining data security. Upcast Data Steward will ensure that Work package 6 participants constantly review and update the Moodle settings to meet this data management plans needs and legal obligations regarding data protection and privacy. Further details about detailed collected data will be provided in an update of this plan upon start of the task and configuration of the Moodle platform.

## 6.3 Surveys

<sup>10</sup> <https://www.upcast-project.eu/privacy-policy/>

During the execution of task **T7.4 Exploitation and Sustainability** some data from the pilots' deployments will be collected to inform business analysis and sustainability planning. This is covered and described in chapter 2. However, as part of this process IDC will conduct at least one market survey to assess the landscape. IDC performs over 400 000 surveys with individuals every year and the surveys that are expected to be conducted by IDC during the Upcast project will necessarily be conducted by the specialized unit that performs market surveys for the rest of IDC's clients. The Responsible unit declares that to ensure data access, security, and GDPR compliance IDC follows the following procedures and guidelines:

### **6.3.1 Survey Data Management Processes**

Data Collection: IDC's surveys collect various types of data from respondents, which may include personal, demographic, and business-related information.

Data Processing: Third-party data processors handle data collection, storage, and analysis, following agreed-upon protocols.

Data Storage: Data is stored securely on servers or cloud platforms, with appropriate access controls and encryption measures.

Data Access: Access to collected data is restricted to authorized personnel only, and roles and permissions are defined to limit access based on need.

Data Sharing: Data may be shared with clients or research projects as per agreements, with a focus on data minimization and only sharing necessary information.

Data Retention: Data retention policies are implemented to ensure that data is kept only for the necessary period and then securely disposed of.

### **6.3.2 Survey Data Security Measures:**

Encryption: Data in transit and data at rest are encrypted to prevent unauthorized access.

Access Controls: Strict access controls are implemented, ensuring that only authorized individuals have access to the data.

Authentication: Multi-factor authentication (MFA) may be used to ensure that only authorized users can access the data.

Regular Auditing: Regular audits are conducted to monitor data access and ensure compliance with security protocols.

Data Minimization: Only necessary data is collected and processed to minimize risk.

Incident Response: Protocols are in place to handle data breaches or security incidents promptly and effectively.

### **6.3.3 Survey GDPR Compliance:**

Consent: Respondents' consent is obtained before collecting personal data, clearly explaining the purpose and scope of data usage.

Data Subject Rights: Respondents have the right to access, correct, and erase their data. IDC should have mechanisms in place to accommodate these requests.

Privacy Notices: Privacy notices are provided to respondents, outlining how their data will be processed.

Data Transfer: If data is transferred outside the EU, appropriate safeguards (like Standard Contractual Clauses) are in place.

Data Processing Agreements: Agreements with third-party data processors include GDPR-compliant clauses.

Data Protection Officer (DPO): IDC has an appointed DPO to ensure GDPR compliance and oversee data protection activities during survey execution.

It's important to note that IDC, like any organization, periodically reviews and updates its data management processes to align with the evolving regulatory landscape and best practices in data security and privacy. And this data management plan should be updated before the execution of the survey in question as these processes may have changed.

## 7 Data Management for UPGCAST Plugins

For each of the UPGCAST plugin we describe re-used data (if any) and generated research data. We also describe data processed by the plugin, and for what purpose, from a hypothetical user of the envisioned final products. We also describe software and machine learning models.

### 7.1 Resource Specification

Re-used Research and Development data: Does not use any.

Generated Research and Development data: Measures of efficiency and effectiveness in the task of resource description by resource owners.

The reason for processing and the specific data needed from a user to achieve it:

For generating resource description in interoperable format: Values of metadata fields according to UPGCAST vocabulary. For generating profiles of datasets to improve discoverability, the dataset to be profiled.

Software outputs:

- Implementation of resource specification to be released with an Open Source licence (TBD after exploitation plan).
- Library of algorithms for data profiling (Open source, precise licence TBD at after exploitation plan)

ML Model output: No output

### 7.2 Resource Discovery

Re-used Research and Development data: Benchmarks of dataset discovery algorithms such as Valentine<sup>11</sup> and SANTOS<sup>12</sup>. Open licenced available benchmarks on tabular data to knowledge graph matching SemTab<sup>13</sup>.

---

<sup>11</sup> <https://delftdata.github.io/valentine/>

<sup>12</sup> <https://github.com/northeastern-datalab/santos>

<sup>13</sup> <https://www.cs.ox.ac.uk/isg/challenges/sem-tab/>

Generated Research and Development data: Comparison of our approach with state-of the art using the aforementioned benchmarks.

The reason for processing and the specific data needed from a user to achieve it:

For the purpose of enabling potential business partners to search relevant resources in a catalog, full access to the catalog of resource specifications.

For the purpose of discovering relevant datasets in a catalog: access to abstract resource specification. Optionally, the dataset and its contents.

Software output:

- Implementation of dataset search algorithms, to be released with an Open Source licence (TBD after exploitation plan).
- Implementation of dataset discovery algorithms, to be released with Open Source licence (TBD after exploitation plan)

ML Model output: No output

### **7.3 Data Processing Workflow**

Re-used Research and Development data:

The Data Processing Workflow uses data that comes from the other UPCAST plugins during execution of the workflow specification. These data signal the progress of the various phases of the workflow execution.

Generated Research and Development data:

The Data Processing Workflow plugin generates data dynamically during execution of the workflow. The generated data represents events and states of interest of the Data Processing Workflow plugin, like the successful completion of a workflow phase, and are fed to the monitoring plugin.

The reason for processing and the specific data needed from a user to achieve it

The purpose of the Data Processing Workflow plugin is to apply the UPCAST workflow specification, i.e., the phases that involve data preparation, annotation, advertisement, discovery, negotiation, and execution.

Software output:

The Data Processing Workflow software is a standalone plugin that orchestrates the execution of the workflow specification. The plugin will be built upon existing commercial technology. Licensing details will be available after the exploitation plan is complete.

ML Model output: Not applicable.

### **7.4 Data Pricing**

Re-used Research and Development data:

Reused open research data available in the following link:

<https://gitlab.com/sandresazcoitia1/data-pricing-tool>. This dataset was generated during the elaboration of the following research works, and was made open for

research and industry exploitation by solely citing two papers authored by LSTech team.<sup>1415</sup>

The dataset contains information about:

- 1 **Products.** Contains a list of data products scraped from the different data marketplaces.
- 2 **DMs.** Contains a list of data marketplaces scraped and the number of products and different vendors found there.
- 3 **Labels.** Contains a list of labels used by data marketplaces to tag data products.
- 4 **Product - Label.** Contains an association of labels to data products in our sample.
- 5 **Data providers in Data Marketplaces.** Contains an association of data providers or sellers to the data marketplaces where they were found in this study.
- 6 **Provider-Marketplaces Matrix.** Contains a matrix of data providers vs. data marketplaces.
- 7 **All prices classified.** Contains a list of data products with prices, with their classification according to the labels of AWS Marketplace.

For more information about the scope of the scraping exercise please refer to Sect. II of the paper and to the documentation shared in the Gitlab shared above.

#### Generated Research and Development data:

During the course of the project, additional information may be collected about data products, data providers and data marketplaces in order to update part of this information or get complementary information to feed the specific use cases of the project. This information will be obtained from public Internet websites.

Regarding the license of any additional data generated during the project, this is to be defined after the exploitation plan.

#### The reason for processing and the specific data needed from a user to achieve it

The information will be processed for the following purposes:

- Inferring the category or categories of a data product
- Clustering and detecting similar data products to a given data product or to a description and metadata provided by users
- Inferring the price of a given data product or to a description and metadata provided by users based on the available information about market prices and the metadata of different data products.
- Generating explanations about the inference process followed to arrive to a certain prediction for users to understand which the most important features are affecting market prices.

#### Software output:

---

<sup>14</sup> S. Andres Azcoitia, C. Iordanou, N. Laoutaris, "Understanding the Price of Data in Commercial Data Marketplaces," International Conference on Data Engineering **ICDE'23**.

<sup>15</sup> S. Andres Azcoitia, C. Iordanou, N. Laoutaris, "Measuring the Price of Data in Commercial Data Marketplaces," **ACM Data Economy Workshop**, 2022.

The software provided will include a server exposing a REST API with endpoints allowing users to:

- Find similar data products found in marketplaces to a description and metadata values passed by users as parameters
- Get a price range for data products in the market based on a description and metadata values passed by users as parameters
- Get explanations that rank the relevance of features affecting the price returned in this prediction.

The licensing of this product is still to be defined after exploitation plan.

#### ML Model output:

The pricing plugin will develop data product classifiers and price regressors that will be used in their operation, and will be actively used to generate the answers by the pricing tool. The licensing of these models is still to be defined after the exploitation plan.

## **7.5 Environmental Impact**

#### Re-used Research and Development data:

The MIT Supercloud Dataset<sup>16</sup>, this is a collection of parsed datacenter logs and time series data of hardware utilization from the MIT Supercloud system. It is released with a CCby-nc-nd license.

#### Generated Research and Development data:

The following types of data may be collected from users:

- Hardware information of a server or data centre characteristics where datasets are being created, stored and processed.
- Geographical location of the server/data centre
- Dataset metadata according to the UPCAST resource specification vocabulary

There may be intermediate data generated during the development of this plugin, such as energy or power metrics, which will be aggregated and processed to deliver the final intended output of the plugin.

The licenses are to be defined after the exploitation plan.

#### The reason for processing and the specific data needed from a user to achieve it

The data being collected for this plugin will be processed for the following purposes:

- To estimate the energy consumption and intensity of a dataset, based on the hardware information and dataset metadata
- To assign an energy profile value (such as, a rating) for a dataset in the marketplace

---

<sup>16</sup> <https://registry.opendata.aws/dcc/>

- To generate and aggregate energy consumption metrics of processes being applied throughout the data processing workflow

Software output:

The software being developed as a part of this plugin is the energy profiler, which will assign an energy profile or rating to a dataset. Using explainable AI, the profiler will also explain why a dataset has been given a particular profile or rating. The licensing of this product is still to be defined after exploitation plan.

ML Model output:

The environmental impact optimizer plugin will actively use machine learning models for the development of the energy profiler. The licenses of these models are to be defined after the exploitation plan.

## 7.6 Privacy and Usage constraints

Re-used Research and Development data:

TPC-H<sup>17</sup> , MIMIC-III<sup>18</sup> datasets to be reused as benchmarks for the performance of the privacy and usage queries.

Generated Research and Development data:

Privacy and usage constraint rule data will be generated as output. The rules and constraints files will be open license. Further details about data license will be added to and update DMP after discussion on exploitation.

The reason for processing and the specific data needed from a user to achieve it

For the purpose of creating privacy and constraint rules, access to the dataset and its schema.

Software output:

Implementation of evaluation engine for privacy and usage constraints. As there is some dependency on existing IP by some of the partners, licence TBD after exploitation plan

ML Model output:

No output

## 7.7 Negotiation

Re-used Research and Development data:

*None*

Generated Research and Development data:

The “negotiation result” and “negotiation process” data will be generated by the negotiation plugin. The “negotiation result” data will represent the outcome of the

---

<sup>17</sup> TPC Benchmark H.<http://www.tpc.org/tpch/>, 2018

<sup>18</sup> Medical Information Mart for Intensive Care III <https://mimic.mit.edu/docs/iii/>,2018.

negotiation process while the “negotiation process” will represent the intermediary data form negotiation that will be taking place in ping-pong manner.

As the negotiation contains private information, the generated data will have a closed licence.

As the negotiation contains private information, the generated data will have a closed licence.

#### The reason for processing and the specific data needed from a user to achieve it

For the purpose of recommending the best negotiation counter-offer. Users must provide negotiation terms that include the ranges of values different fields (price, carbon consumption, etc.) can take, as well as relationships between fields (e.g., lower price for research purposes only).

#### Software output:

The software being developed as a part of the plugin will provide an option to negotiate terms (such as pricing, privacy and usage, etc.) between the consumer and provider. Additionally, the software will also provide recommendations, e.g., the best price options.

The licence of the software will be defined after the exploitation plan.

#### ML Model output:

*Any Machine-Learning model developed during the project, and its licence. Same licence considerations as with software*

An ML model will be developed for tasks such as recommendation and possibly for the generation of contracts in natural language after negotiation has concluded.

## **7.8 Data Integration**

### Re-used Research and Development data:

*ForBackBench*<sup>19</sup> is a benchmark for data integration/exchange systems and scenarios from the database and semantic web areas. ForBackBench extends the Chasebench<sup>20</sup> benchmark, which contains the following related systems<sup>21</sup>:

---

<sup>19</sup> <https://github.com/georgeKon/ForBackBench>

<sup>20</sup> <https://github.com/dbunibas/chasebench>

<sup>21</sup> A. Bonifati, I. Ileana, and M. Linardi. Functional Dependencies Unleashed for Scalable Data Exchange. In SSDBM, 2016

ChaseFun, DEMo<sup>22</sup>, LLunatic<sup>23</sup>, PDQ<sup>24</sup>, Pegasus<sup>25</sup>, Graal<sup>26</sup>, DLV<sup>27,28</sup>, E<sup>29</sup>, RDFox<sup>30</sup>.

#### Generated Research and Development data:

We generate new data that can be used to evaluate the performance of data integration/exchange systems and algorithms. Examples are runtimes, what kinds of structured data are integrated.

Evidently, because of the nature of data integration/exchange, every call to this plugin produces either new materialised views or answers to queries. Licenses to be defined after exploitation plan.

#### The reason for processing and the specific data needed from a user to achieve it

To integrate datasets, the plugin needs the esource specifications from the sources and targets that are part of the integration so the system can semi-automatically compute a source to target mapping. Metadata of the structure or domain of the sources may be used to refine the suggestion of mappings.

#### Software output:

A system that allows users to easily and intuitively design source to target mappings. Furthermore, it serves as the integration/exchange step of a DPW. We are likely to use existing implementations of the existing algorithms for query rewriting and view materialisation as solutions to the data integration/exchange problem. This will be released as Open Source, as far as the existing implementations will allow us (to be defined later).

ML Model output: Not applicable.

## **7.9 Federated Machine Learning**

This plugin is currently being re-specified due to the main responsible partner reducing effort as part of an amendment, with a new partner taking its place. Information will be provided in the next update of the plan. However currently Nokia's Data Marketplace platform uses a federated learning module, which is a method of creating artificial intelligence (AI) models using data from multiple sources without actually sharing the

---

<sup>22</sup> (Data Exchange Modelling) R. Pichler and V. Savenkov. *DEMO: Data Exchange Modeling Tool*. *PVLDB*, 2(2):1606–1609, 2009

<sup>23</sup> F. Geerts, G. Mecca, P. Papotti, and D. Santoro. The LLUNATIC Data-Cleaning Framework. *PVLDB*, 6(9):625-636, 2013 <http://db.unibas.it/projects/llunatic/>

<sup>24</sup> M. Benedikt, J. Leblay, and E. Tsamoura. Querying with access patterns and integrity constraints. In *VLDB*, 2015 <http://www.cs.ox.ac.uk/projects/pdq/home.html>

<sup>25</sup> M. Benedikt, J. Leblay, and E. Tsamoura. Querying with access patterns and integrity constraints. In *VLDB*, 2015 <http://www.cs.ox.ac.uk/projects/pdq/home.html>

<sup>26</sup> J.-F. Baget, M. Leclère, M.-L. Mugnier, S. Rocher, and C. Sipieter. Graal: A toolkit for query answering with existential rules. In *RuleML*, 2015 <https://graphik-team.github.io/graal/>

<sup>27</sup> N. Leone, G. Pfeifer, W. Faber, T. Eiter, G. Gottlob, S. Perri, and F. Scarcello. The DLV system for knowledge representation and reasoning. *TOCL*, 7(3):499–562, 2006 <http://www.dlvsystem.com/dlv/>

<sup>28</sup> M. Meier. The backchase revisited. *VLDB J.*, 23(3):495–516, 2014

<sup>29</sup> S. Schulz. System Description: E 1.8. In *LPAR*, 2013 <http://www.lehre.dhbw-stuttgart.de/~sschulz/E/E.html>

<sup>30</sup> B. Motik, Y. Nenov, R. Piro, I. Horrocks, and D. Olteanu. Parallel Materialisation of Datalog Programs in Centralised, Main-Memory RDF Systems. In *AAAI*, 2014 <https://www.cs.ox.ac.uk/isg/tools/RDFox/>

data. This way, data owners can keep their data private and still benefit from the collective intelligence of the network. Federated learning also reduces the need for transferring large amounts of data, which can save bandwidth and energy. Nokia's Data Marketplace platform enables data owners and consumers to collaborate on building AI models using federated learning and blockchain technology.

## 7.10 Safe and Secure Exchange

This plugin is currently being re-specified due to the main responsible partner reducing effort as part of an amendment, with a new partner taking its place. Information will be provided in the next update of the plan.

## 7.11 Monitoring

### Re-used Research and Development data:

No existing research data are expected to be used by the monitoring plugin. The only data the plugin uses are the ones that are generated by the other plugins of UPGAST.

### Generated Research and Development data:

No new data is generated by the monitoring plugin. The plugin consumes monitoring data that is emitted by other plugins, persists them, and provides them for further analysis, for example for checking compliance of the dataflow.

### The reason for processing and the specific data needed from a user to achieve it

The purpose of the monitoring plugin is the collection, persistence and provision of further analysis of the monitoring data coming from other plugins. User needs to provide consent for the processing of their activities with each of the other plugins. Monitoring data includes any type of data that can be produced from different sources and represents events or states of interest of those sources. They are free format, i.e., they have no specific structure.

### Software output:

The monitoring plugin will be an extension of the proprietary MIRA platform that implements digital twin functions including analysis and simulation. The monitoring plugin will be released under proprietary license, similar to the license of MIRA. The software will be made available for use in the project.

### ML Model output:

Not applicable.

## 8 Data Management in the context of pilots

In this section we describe our management plan for the data that will be used in the pilot to demonstrate the utility of the UPGAST plugins in supporting data exchange scenarios. As pilot cases are drawn from very different sector and use different categories of data, we describe our plan for each pilot in a question/answer format.

### 8.1 Pilot Case 1.1: Digital Marketing Data and Resources (JOT)

**1) What types of data will be used or exchanged within the pilot? Is any personal data included?**

JOT data used in this pilot will be based on \*csv data sets including performance indicators of the marketing campaigns. It contains numbers and descriptors used to determine if a specific campaign has achieved the expected impact results in terms of number of clicks and impressions, which is related to the interests of the users in that specific category (topic)

No personal or sensitive data is collected during the campaign monitoring.

**2) Is consent of data owners required in the pilot? If so, how is consent collected?**

JOT is the owner of the data, so no specific consent is needed.

**3) What formats are data collected in?**

In this pilot the following data formats will be used:

\*csv to share the data sets requested by the final users (data consumers)

\*pdf when an advanced analysis is demanded, including conclusions and insights obtained based on the data sets

**4) Is data transformed or processed once collected? How? If so how is it transformed and what is the output format?**

Raw data collected from the ad platform is cleaned before stored in the data basis (input and output format are \*csv)

Data sets are segmented and aggregated based on the user demands (location, categories, timespan...)

When demanded, data sets are processed by analytical algorithms to extract the market knowledge from the marketing performance indicators.

**5) Where is any data stored? If on premises please describe? What are the security measures in place?**

Data are stored in Google Cloud infrastructure managed by JOT, which provides own security measures to avoid fraudulent access as well as duplicates and periodic backups

**6) Has an individual been identified that is clearly responsible for data management in your organisation?**

The company has a CTO which is the main responsible of monitoring data management, both in terms of processing, quality and costs

**7) How will the exchanged marketing data comply with the FAIR principles, ensuring they are Findable, Accessible, Interoperable, and Reusable?**

JOT will exchange the marketing data on business basis, which includes a contract signature. Once approved, the data consumer will have free use rights over the data requested.

Data consumer will have access only to the requested data sets, which will follow the data model defined in the service request

**8) What licensing is expected? If any data be released with an open (open Access) please describe the intended license?**

For private data exploitation, data will be shared on a contract basis

JOT will generate and publish free-to-use several data samples as examples for data consumers and researches

Upon specific agreement, JOT will share data sets for free with researchers for AI models test, train and validation.

## **8.2 Pilot Case 1.2: Digital Marketing Data and Resources (Cactus)**

**1) What types of data will be used or exchanged within the pilot? Is any personal data included?**

Cactus will use client data from Google Analytics, Google Ads, and Meta, along with some economic data for digital performance analysis. No personal data will be included.

**Is consent of data owners required in the pilot? If so, how is consent collected?**

Yes, consent of data owners is required in the pilot. A signed contract clarifies data access and actions between Cactus and its clients, ensuring clients retain ownership of their data

**2) What formats are data collected in?**

Some of them we collected via API and some of them via excel files.

**3) Is data transformed or processed once collected? How? If so how is it transformed and what is the output format?**

Neither transformation nor process takes place after the data are collected.

**4) Where is any data stored? If on premises please describe? What are the security measures in place?**

- Google Cloud
- AWS (Meta Analytics)
- Economic Data (Data stored in our custom-made web based system.)

**5) Has an individual been identified that is clearly responsible for data management in your organisation?**

Not yet. But we schedule to assign DPO responsibilities to one of our employees starting from the end of September 2023

**6) How will the exchanged marketing data comply with the FAIR principles, ensuring they are Findable, Accessible, Interoperable , and Reusable?**

We will implement technical and organizational measures to ensure compliance with FAIR principles during data exchange.

**7) What licensing is expected? If any data be released with an open (open Access) please describe the intended license?**

No open data release is being planned

**8) How long is the data required for the Purpose?**

The detailed data sets of the client's status should stay in our system for the same time that the client is a user of our system. Some data we use to calculate the Averages of the market. These data we should remain in our system until the life time of our system

**9) What happens to data after the Purpose is no longer required?**

After a user requests to terminate his appearance in our system, we start a process called off boarding. In the end of this process we hard delete the majority of data.

### **8.3 Pilot Case 2: Biomedical and Genomic Data Sharing (NHRF-ICTAbovo)**

**1) What types of data will be used or exchanged within the pilot? Is any personal data is managed?**

The types of data that will be used are described in D1.1.and can be distinguished in three types: 1) Genomic -DNA derived- data from clinical samples, 2) Transcriptomic data derived from in vitro models, 3) Clinical and Demographic data

- Clinical/demographic data are pseudonymised (name has been replaced by a code) but could contain data like: date of birth, date of diagnosis, hospital name, profession and other personal data. Data like address, telephone number and e-mail are not included.
- Genomic data from clinical samples is by definition sensitive data

**2) Is consent of data owners required in the pilot? If so, how is consent collected?**

Consent is required in the case we analyse biologic material from clinical samples and the relevant clinical data. Consent is collected from the clinical partners. Detailed information is given to the patients from a specialised physician and the whole procedure has been approved by the Ethics Committee and Scientific Committee of the Hospital.

**3) What formats are data collected in?**

Data formats are described in D1.1. In particular: a) Genomic data are collected as structured text files (indicative formats MAF, VCF). For in-house generated genomic data, raw data are in FastQ format. b) Transcriptomic data are in structured text file format. c) Clinical data are collected in xl format.

**4) Is data transformed or processed once collected? How? If so how is it transformed and what is the output format? (described in D1.1)**

Currently raw data (FastQ) are transformed following pipelines exploiting state of the art tools (aligners, annotators) to generate structured text files. In the frame of Upcast we will try to exploit new tools [VRS (<https://www.ga4gh.org/product/variation-representation/>), phenopackets (<https://www.ga4gh.org/product/phenopackets/>)] in order to adapt our analytical workflows to new standards and transform our datasets to updated standardised-formats that could facilitate and improve sharing of genetic and clinical information.

**5) Where is any data stored? If on premises please describe? What are the security measures in place?**

In-house data is stored encrypted on local servers and external disks (total availability ~30TB) or academic cloud infrastructures (Hypatia/Elixir-GR). Servers can be only accessed locally or via private network ensuring security.

**6) Has an individual been identified that is clearly responsible for data management in your organisation?**

In NHRF there is a person responsible for data management but for research projects the principle investigator is considered responsible for data management of the data related to the specific research project. The principal investigator is supported by the administrative and technical services of NHRF.

**7) How will the exchanged marketing data comply with the FAIR principles, ensuring they are Findable, Accessible, Interoperable, and Reusable?**

The genomic and transcriptomic data will be made findable through appropriate tagging, indexing, and storage in a searchable repository, using unique identifiers for each dataset. Accessibility will be ensured with the appropriate authentication and authorization system and implementation of programmable accession protocols (APIs). To ensure data is interoperable, it will be prepared in standard, machine-readable formats, and it will use biomedical ontologies and vocabularies that follow FAIR principles. To ensure reusability, there should be comprehensive and accurate metadata provided, detailing how the data was collected, processed, and what it represents. This includes information about the methodology used, the context of data collection, the type of genomic sequencing or analyses performed, any data cleaning or preprocessing steps, and potential biases or limitations in the data. Additionally, the data will comply with specific standard formats and terminologies, to ensure that it can be easily integrated and used in different research contexts. Standard formats in genomics include file types like FASTQ for raw sequencing data, VCF for genetic variants, or BAM for aligned sequencing data.

**8) What licensing is expected? If any data be released with an open (open Access) please describe the intended license?**

For open genomic data, Open Data Commons and Creative Commons licenses are applicable. Open data licenses could apply to NHRF-processed data originating from open public resources.

For commercial use, various custom licenses could be applied, e.g. an upfront perpetual license or a subscription license (single-user or enterprise).

### **9) How long is the data required for the Purpose?**

Regarding the experimental process the data are required for different periods that could range from a few weeks up to several months.

### **10) What happens to data after the Purpose is no longer required?**

Raw and processed genomic data as well as clinical data are stored locally. Raw genomic data could be uploaded to public repositories like ArrayExpress (<https://www.ebi.ac.uk/biostudies/arrayexpress>) and European Genome Phenome Archive(<https://ega-archive.org/>).

## **8.4 Pilot Case 3: Sharing Public Administration for Climate across Thessaloniki Cities (MDAT+OKFGR)**

### **1) What types of data will be used or exchanged within the pilot? Is any personal data is managed?**

The types of data include the following: general demographics for the Thessaloniki metropolitan area, urban statistics, households' living conditions related to environmental indicators, transport statistics and urban traffic conditions, environmental statistics and air pollution measurements.

There is no personal data used for this pilot case.

### **2) Is consent of data owners required in the pilot? If so, how is consent collected?**

Consent of the data owners is not required for most data used in this pilot case. Only a few data where the data owner is the Hellenic Statistical Authority (ELSTAT) may require consent. For this purpose, bilateral contracts will be signed between MDAT and ELSTAT.

### **3) What formats are data collected in?**

Excel files, csv and dat files

### **4) Is data transformed or processed once collected? How? If so, how is it transformed and what is the output format?**

Yes, the data is transformed and processed once collected. The processing concerns the data integration and the production of new environmental indicators. The output formats are in excel files also.

### **5) Where is any data stored? If on premises please describe? What are the security measures in place?**

Datasets may be stored in MDAT's local server which is offline and secured. A certain file can be locked and only authorized people can have access using personal passwords. Additionally, datasets that will be uploaded in the pilot implementation platform that uses the UPCASt plugins, will be stored most likely in a cloud server, which will be managed by MDAT, with the technical support of OFKGR.

**6) Has an individual been identified that is clearly responsible for data management in your organisation?**

Responsible for the data management of the pilot will be the project managers of both pilot partners, MDAT and OKFGR, Paraskevi Tarani and Charalampos Bratsas respectively.

**7) How will the exchanged marketing data comply with the FAIR principles, ensuring they are Findable, Accessible, Interoperable, and Reusable?**

All the datasets of the pilot will be published as open data (unless stated otherwise by the contracts signed), while unique identifiers will be used to describe them. Furthermore, the DCAT vocabulary will be used to enrich the datasets with metadata, so they can be discoverable (Findable).

With regards to data Accessibility all dataset metadata will be open licensed, even if the data themselves may not be open (this will apply only in cases the contracts prohibits publishing the data as open).

To ensure data interoperability, datasets will be semantically described using vocabularies and will be also stored in RDF enabling them to be both syntactically parseable and semantically understandable and allowing them to be exchanged and reused between researchers, institutions, municipalities, organizations etc..

Finally, regarding data reusability, all pilot datasets and the outcomes of the pilots, along with their respective metadata will be available with an open license, and the least amount of restrictions in order to encourage their widest reuse and easier integration with other data sources. To reinforce this point, relevant documentation may also be available.

**8) What licensing is expected for? If any data be released with an open (open Access) please describe the intended license?**

All open datasets will be released with Creative Commons or Open Data Commons licenses. Other datasets may have per case licensing according to the data owner and the contract that will be signed.

**9) How long is the data required for the Purpose?**

Ideally, pilot data will be available in the Platform beyond the project's lifetime, since MDAT would provide the distribution of these datasets as a free service to allow and support policy making. In the case where some data owner demands it, data will be deleted.

**10) What happens to data after the Purpose is no longer required?**

Same answer as previous question.

## **8.5 Pilot Case 4: Health and Fitness Data Trading (Nissatech)**

**1) What types of data will be used or exchanged within the pilot? Is any personal data is managed?**

Health-related data collected in the fitness/sport training

**2) Is consent of data owners required in the pilot? If so, how is consent collected?**

Every user signs consent through the Nissatech application.

**3) What formats are data collected in?**

json / csv

**4) Is data transformed or processed once collected? How? If so how is it transformed and what is the output format? (described in D1.1)**

Data is cleaned and prepared for analysis. Output format is json/csv as well

**5) Where is any data stored? If on premises please describe? What are the security measures in place?**

Data is stored on servers (AWS) after which it is collected onto our internal server with no public access. AWS is AZURE secured.

**6) Has an individual been identified that is clearly responsible for data management in your organisation?**

Yes

**7) How will the exchanged marketing data comply with the FAIR principles, ensuring they are Findable, Accessible, Interoperable, and Reusable?**

We will make FAIR compliant middleware which will communicate with our cloned private repository with suggested authentication and authorization.

**8) What licensing is expected? If any data be released with an open (open Access) please describe the intended license?**

There is no open access planned

**9) How long is the data required for the Purpose?**

As long as possible, however, user retains right to be forgotten.

**10) What happens to data after the Purpose is no longer required?**

Data can be deleted if the user isn't the member of the club anymore

## **9 Data Sharing and Access**

This chapter describes Data Sharing and Access among the partners during the course of the project. Data Sharing, Access controls and Dissemination are described in the following paragraphs.

### **9.1 Data Sharing Agreements between Partners**

The Upcast project recognizes the importance of data sharing among consortium partners to achieve the project's objectives effectively. Data sharing is tacitly accepted through the Consortium Agreement and all conditions cited there are applicable and reiterated in this document where data is needed by another partner to perform the work expected in the Description of Action. Where particular issues arise and the case may require specific attention, Data Sharing agreements will be established between all partners involved in the data sharing. These agreements will outline the terms and conditions governing the sharing of data, including the types of data shared, the purpose of sharing, and the responsibilities of each partner. The agreements will also address issues related to data ownership, confidentiality, and data usage restrictions.

## **9.2 Access Controls and Authentication Mechanisms**

To ensure secure and controlled access to project data, access controls and authentication mechanisms have been implemented. Each consortium partner has designated access rights based on their role and responsibilities within the project. Role-based access controls have been applied to , granting access only to authorized personnel. Access to sensitive or confidential data will require multi-factor authentication to prevent unauthorized access. Regular access reviews will be conducted to ensure that access permissions remain up-to-date and aligned with project needs.

## **9.3 Data Dissemination and Publication Policies**

Data dissemination and publication policies have been established in the Consortium Agreement to guide the sharing of project findings and outcomes with external stakeholders, including the research community and the public. Non-sensitive and anonymized data may be published in public repositories to promote transparency and enable data reuse. However, data dissemination will be subject to compliance with ethical guidelines, privacy regulations, and intellectual property rights defined in the Consortium Agreement. Dissemination will be accompanied by clear metadata and documentation to enhance data findability and usability.

## **9.4 Intellectual Property Rights and Data Ownership Considerations**

Intellectual property rights and data ownership considerations are addressed through the Consortium Agreement which is a legally binding contracts among consortium partners. Data generated, collected, or processed as part of the Upcast project will be subject to that or ensuing agreements that define ownership, usage rights, and potential restrictions. Clear provisions are established there to specify the ownership of derived insights, analysis results, and any intellectual property generated from the data. If the conditions change and partners wish to change their Intellectual property rights in relation to the other parties that document will need to be amended.

# **10 Data Security and Privacy**

The UPCAST project is committed to ensuring the highest level of data security and privacy. Robust encryption, access controls, GDPR compliance, and incident response procedures will collectively safeguard project data and maintain the trust of both consortium partners and external stakeholders. This approach guarantees that the data exchanged and managed within the project is protected against unauthorized access,

breaches, and potential privacy violations. The following paragraphs describe the Upcast approach.

## **10.1 Security Measures**

The UPCASt project places a paramount emphasis on data security and privacy to protect the confidentiality and integrity of the exchanged data. The following measures will be implemented:

### **10.1.1 Data Encryption during Storage and Transfer**

All project data, whether in transit or at rest, will be encrypted using industry-standard encryption protocols. Secure Socket Layer (SSL) and Transport Layer Security (TLS) will be employed to ensure the encryption of data during transfer. Data stored within repositories or databases will be encrypted using strong encryption algorithms to prevent unauthorized access.

### **10.1.2 Access Controls and User Authentication**

Access controls will be enforced to restrict data access based on roles and responsibilities. Role-based access controls will ensure that only authorized individuals can access specific datasets. User authentication will require strong passwords, and where appropriate, multi-factor authentication mechanisms will be employed to prevent unauthorized access.

## **10.2 Privacy**

Privacy is particularly important to the project partners and the following paragraphs commit the consortium to adherence to GDPR and outline procedures in the case of a data breach.

### **10.2.1 Compliance with Relevant Data Protection Regulations**

The UPCASt project will rigorously adhere to relevant data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union. Data collection, processing, and sharing will be conducted in full compliance with GDPR principles, including obtaining informed consent, ensuring data subject rights, and implementing safeguards for cross-border data transfers.

### **10.2.2 Data Breach Response and Incident Management Procedures**

In the event of a data breach or security incident, the UPCASt project will immediately contact the IT Department of IDC which has an incident response procedures in place and longstanding experience in data breach management. IDC's Incident Response Team will be responsible for assessing and containing the breach, notifying affected parties, and coordinating necessary actions. The breach will be reported to the relevant authorities as required by applicable regulations.

IDC as coordinator and Data Steward maintains comprehensive documentation of security measures and incident response procedures as per company . Regular security audits and vulnerability assessments will be conducted to proactively identify and mitigate potential threats.

## **11 Data Retention and Disposal**

The UPCASt project has a clear data retention and disposal policy, specifying retention of data will only persist as long as that data is essential for the accomplishment of the tasks and development envisioned in the Description of Action. At the end of the need

for the data or at the termination project, data will be anonymized or securely disposed of, following established procedures to ensure compliance with European legislation regarding privacy regulations and ethical standards. Where the need for data is demonstrated for commercial or deployment activities after that period the conditions will be made known to all affected partners and an agreement and policies will be laid down otherwise this DMP will be deemed to be effective and data will be retired.

## 12 Data Management Plan Review

This DMP is deemed at the present state to reflect the needs of the project, be in line with existing legislation and common commercial and research practice. However the conditions may change during the course of the project and the data management board may at any point determine that this document does not fully address any issue affecting data and data management and that Board may choose to update this set of policies and practice. This deliverable is therefore deemed to be a living document. Already at the time of writing this document the project has determined that a new partner is required to provide additional technical competencies and an amendment is underway to add an additional Data Marketplace for the testing and deployment of the plugins that are being developed. These modifications are expected to spurn a review and amendment of this document in the early fall of 2023. An additional review of this document is planned to occur as a result of the Data Board meetings. Although no issues are foreseen that would currently merit a revision of this document it is clear that for example when exploitation plans are drawn up (or before) licensing of some components may be identified and it is expected where they affect data management that these

## 13 Conclusion

The Upcast Data Management Plan (DMP) outlines a comprehensive framework to ensure efficient and responsible handling of data throughout the project's lifecycle. This plan encompasses various aspects, from data collection and processing to sharing, security, and compliance with regulations. The DMP sets the context for the document's significance in achieving project objectives. It then delves into the Data Summary section, classifying the types of data involved, including reuse, operational, user, economic, social, and regulatory data. It highlights the purpose and contributions of each data type, emphasizing their utility beyond the project's scope. The document introduces the principle of FAIR data (Findable, Accessible, Interoperable, and Reusable) and shows how it is core to the Upcast plan. The FAIR Data section elucidates how data will be tagged with metadata for findability, access controls will be enforced for accessibility, standard formats will ensure interoperability, and comprehensive documentation will enhance reusability. Apart from data, the DMP addresses other research outputs and the governance structure. Roles such as Data Steward, Data Board, Data Custodians, Data Users, and Data Ethics Officer are defined, establishing a clear hierarchy for data management and accountability. The DMP provides guidelines on Data Collection and Processing for Operational Purposes encompassing various contexts, including data processing for dissemination, transferability, training, and surveys. Data security, integrity, and compliance with regulations are at the forefront of these processes.

Data Management for the Upcast Plugins is a pivotal part of the DMP, outlining the technicalities involved in resource specification, discovery, pricing, and privacy concerns. It highlights the integration of federated machine learning and secure exchange mechanisms, ensuring the utmost confidentiality and reliability during data exchange.

Pilots also play a crucial role in the project, and the Data Management in the Context of Pilots section provides detailed strategies for each pilot case. These strategies address data types, consent, formats, security, and licensing, ensuring responsible and compliant data handling.

Data Sharing and Access delve into data sharing agreements, access controls, dissemination, and intellectual property rights. This section ensures that data exchange is governed by explicit agreements that address ownership, confidentiality, and usage restrictions.

Data Security and Privacy are also clearly described in this document. The DMP outlines robust measures, including encryption during storage and transfer, role-based access controls, and adherence to data protection regulations such as GDPR. A comprehensive incident management procedure is established to address any potential breaches promptly.

THE DMP also outlines how Data Retention and Disposal policies are in place to ensure data is retained only for essential purposes and anonymized or securely disposed of when no longer needed. The evolving nature of the DMP is acknowledged, with provisions for review and adaptation based on changing conditions and requirements.

In essence, the Upcast Data Management Plan is a structured, usable framework that ensures responsible, secure, and compliant handling of various types of data. From the initial collection to sharing, security, and eventual disposal, the plan instills confidence in the project's commitment to effective and ethical data management.

# ANNEX I ACRONYMS & ABBREVIATIONS

## Acronyms

Acronyms List	
CP	Consortium Plenary
DoA	Description of Action
PC	Project Coordinator
PMB	Project Management Board
PPR	Project Periodic Report
QM	Quality Management
RM	Risk Management
TM	Technical Manager
WPL	Work Packages Leaders
DMP	Data Management Plan
GDPR	General Data Protection Regulation

Table 1 Acronyms

## Abbreviations

Abbreviation List	
API	Application Programming Interface - A set of rules and protocols that allows different software applications to communicate and interact with each other.
DCAT	Data Catalog Vocabulary is a metadata standard used to describe datasets, data catalogs, and data portals.
DMP	A document that outlines how data will be handled throughout the research project, including data collection, storage, sharing, and preservation.
DOI	Digital Object Identifier - A unique alphanumeric identifier assigned to a digital object, such as a dataset or publication, to provide a persistent link to its location on the internet.
DRM	Digital Rights Management - Technologies and measures used to protect and manage the rights associated with digital content, including data access and usage restrictions.
ETL	Extract, Transform, Load - The process of extracting data from various sources, transforming it into a consistent format, and loading it into a target system or database.

FAIR	Findable, Accessible, Interoperable, and Reusable - A set of principles that aim to make data discoverable, accessible, and usable by both humans and machines.
GDPR	The European Union regulation that governs the protection and privacy of personal data. It sets guidelines for data processing, consent, and individuals' rights.
NDA	Non-Disclosure Agreement - A legal contract that establishes confidentiality obligations between parties involved in sharing sensitive or proprietary information.
NFR	Non-Functional Requirements - Requirements that specify the characteristics and qualities of a system, such as security, performance, scalability, and usability.
PII	Personally Identifiable Information - Information that can be used to identify an individual, such as name, address, contact details, or unique identifiers.
Pseudonymization	The process of replacing identifiable data with artificial identifiers, called pseudonyms, to protect individual privacy while still allowing data analysis and processing.
RDM	Research Data Management - The practice of managing research data throughout its lifecycle, including data organization, documentation, storage, sharing, and preservation.
TLS	Transport Layer Security is one of the most common encryption protocols used to secure data during transmission over a network.

*Table 2 Abbreviations*