2024

Draft Document

DELIVERABLE 1.1

THIS DOCUMENT IS IN DRAFT FORM AND PENDING OFFICIAL APPROVAL. IT IS SUBJECT TO REVIEW AND MAY BE UPDATED.

D1.1: PROJECT CONCEPT REQUIREMENTS SETUP

Title:	Document version:
D1.1 Project Concept Requirements Setup	0.5

Project number:	Project Acronym	Project Tittle
101093216	UPCAST	Universal Platform Components for Safe Fair Interoperable Data Exchange, Monetisation and Trading

Contractual Delivery Date:	Actual Delivery Date:	Deliverable Type*-Security*:
M6 (June 2023)	M6 (June 2023)	R-PU

*Type: P: Prototype; R: Report; D: Demonstrator; O: Other; ORDP: Open Research Data Pilot; E: Ethics.
**Security Class: PU: Public; PP: Restricted to other program participants (including the Commission); RE: Restricted to a group defined by the consortium (including the Commission); CO: Confidential, only for members of the consortium (including the Commission).

Responsible:	Organization:	Contributing WP
Audun Vennesland	SINTEF	WP1
Shanshan Jiang	SINTEF	WP1

Contributing Authors (organization):

Audun Vennesland (SINTEF), Shanshan Jiang (SINTEF), Dumitru Roman (SINTEF), Luis-Daniel Ibáñez (University of Southampton), George Konstantinidis (University of Southampton), Semih Yumusak (University of Southampton), Jaime Osvaldo Salas (University of Southampton), Ricardo Simon Carbajo (CeADAR), Aditya Grover (CeADAR), Julia Palma (CeADAR), Kostas Kalaboukas (Maggioli), Sofoklis Efremidis (Maggioli), Mariza Koukovini (ICT ABOVO), Eugenia Papagiannakopoulou (ICT ABOVO), Georgios Lioudakis (ICT ABOVO), Santiago Andres Azcoitia (LS TECH), Evangelos Kotsifakos (LS TECH), Paraskevi Tarani (Major Development Agency of Thessaloniki), Anthi Tsakiropoulou (Major Development Agency of Thessaloniki), Charalampos Bratsas (Open Knowledge Foundation), Lorenzo Gugliotta (KU Leuven), Orian Dheu (KU Leuven), Fernando Perales (JOT INTERNET MEDIA), Francisco Rodriguez (JOT INTERNET MEDIA), Miguel Lopez (JOT INTERNET MEDIA), Amit Kumar (ALCATEL-LUCENT (Nokia)), Olga Papadodima (NATIONAL HELLENIC RESEARCH FOUNDATION), Alexandros Lemperos (CACTUS), Anestis Stamatis (CACTUS), Nenad Stojanovich (NISSATECH), Milan Vuckovic (NISSATECH)

Abstract:

This document is deliverable D1.1 Project Concept Requirements Setup for the UPCAST project. The document describes the background and challenges related to the software plugins to be developed in the project, the pilots that will implement and demonstrate use of these plugins, and a set of initial user and system requirements that will scope and steer the technical architecture and system developments as the project continues. Furthermore, the document presents an initial version of a legal framework relevant to the UPCAST project.

Keywords:

User Requirements, Technical Requirements, Legal Requirements, Legal Framework

REVISION HISTORY

Revision:	Date:	Description:	Author (Organization)
V0.1	17.04.2023	First version of the document, adding structure and initial content.	Audun Vennesland (SINTEF)
V0.2	16.05.2023	Second version, draft content in most sections	Audun Vennesland (SINTEF)
V0.3	16.06.2023	Version sent to internal review	Audun Vennesland (SINTEF)
V0.4	19.06.2023	Review and editing	Giorgio Micheletti (IDC)
V0.5	22.06.2023	Formatting	Nevena Raczko (IDC)
V0.6	26.06.2023	Review and comments	Richard Stevens (IDC)
V0.7	29.06.2023	Version addressing comments and proposed changes from internal review + proper formatting	Audun Vennesland (SINTEF)
V0.8	30.06.2023	Final version for submission	Audun Vennesland (SINTEF)

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

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 UPCAST Project

UPCAST, Universal Platform Components for Safe Fair Interoperable Data Exchange, Monetisation and Trading provides a set of universal, trustworthy, transparent and user-friendly data market plugins for the automation of data sharing and processing agreements between businesses, public administrations and citizens. Our plugins will enable actors in the common European data spaces to design and deploy data exchange and trading operations guaranteeing:

- automatic negotiation of agreement terms;
- dynamic fair pricing;
- improved data-asset discovery;
- privacy, commercial and administrative confidentiality requirements;
- low environmental footprint;
- compliance with relevant legislation;
- ethical and responsibility guidelines.

UPCAST will support the deployment of Common European data spaces by consolidating and acting upon mature research in the areas of data management, privacy, monetisation, exchange and automated negotiation, considering efficiency for the environment as well as compliance with EU and national initiatives, AI regulations and ethical procedures. Four real-world pilots across Europe will operationalise a set of working platform plugins for data sharing, monetisation and trading, deployable across a variety of different data marketplaces and platforms, ensuring digital autonomy of data providers, brokers, users and data subjects, and enabling interoperability within European data spaces. UPCAST aims at engaging SMEs, administrations and citizens by providing a transferability framework, best practices and training to endow users in order to deploy the new technologies and maximise impact of the project.

Work package 1, UPCAST concept and MVP definition, addresses the following project objectives:

- Objective 1: Apply models and standards to easily specify data processing requirements in the context of common European data spaces,
- Objective 4: Enable interoperability of data sharing across different entities, platforms and marketplaces,
- Objective 5: Provide a legal and ethical framework for automated contracts, and,
- Objective 8: Pilot and evaluate the platform in Real Market Dataspaces.

These project objectives will be achieved by WP1 through the following sub-objectives:

- Sub-objective 1.1 Establish the vision and direction for the project by defining a Minimum Viable Product (MVP) and agreeing the technical and pilot requirements and usage scenarios to achieve sustainability of the UPCAST set of tools. A methodology to define the requirements and the MVP will be used to stir up the process.
- Sub-objective 1.2 Define the data model and vocabularies for expressing the UPCAST workflows, preferences and other features based on extending existing efforts in GAIA-X and IDS.
- Sub-objective 1.3 Provide a legal framework based on European and National regulations and best practices and ethics guidelines for UPCAST.

- Sub-objective 1.4 Define the UPCAST system architecture using best practices on architecture specification and compliant with the data spaces and with legal and ethical aspects.

1.2 Purpose of the document

This document describes initial requirements that will be further addressed by the MVP definition, the UPCAST system architecture, all software development, and pilot demonstrations in UPCAST. These initial set of requirements, which include functional and non-functional requirements, are classified into pilot requirements which represent the end-user requirements, plugin requirements which represent the technical requirements, system-wide requirements which represent more transversal and generic requirements, and legal requirements.

1.3 Scope of the document

Deliverable D1.1 is the first deliverable in WP1 and collects input from Tasks 1.1 (MVP Definition and Requirements for the Data Value Chain), 1.2 (Pilot Design and Functionalities) and 1.4 (Legal Framework and Requirements). D1.1 provides important input to all other deliverables in WP1 as well as developments taking place in WP2 and WP3, integration and deployment in WP4 and demonstrations in WP5.

1.4 Structure of the document

The remainder of the report is structured as follows: Chapter 2 provides a short background description of the research topics covered in UPCAST along with a reference to the core challenges to be addressed by the UPCAST Plugins. Chapter 3 presents the methodology used for defining the requirements and how these requirements are further addressed in the project. Chapter 4 describes the pilot requirements, that is, requirements related to the realisation of the UPCAST Plugins in the pilots, including user stories that provide context related to the identified requirements. Chapter 5 describes detailed technical requirements relevant for each UPCAST Plugin. Chapter 6 presents system-wide requirements associated with the realization of the UPCAST Platform as a whole. Chapter 7 describes requirements related to the legal framework in UPCAST. Finally, Chapter 8 concludes the document.

2 BACKGROUND AND UPCAST SOLUTIONS

This chapter introduces the background and state of the art of the research topics covered in UPCAST along with a description of the main challenges that still exist within each topic.

Figure 1 provides an overview of the three phases of a dataset undergoes following the UPCAST process. In the Preparation phase the dataset is prepared for processing, and annotated by metadata, pricing information and environmental information. In the Agreement phase the specifications of a dataset that are set by the dataset provider and the requirements for the dataset that are expressed by the consumer are matched. In the Transfer and Monitoring phase the execution of the dataset is performed in a safe environment. Figure 1 also shows how existing software solutions, either brought in by partners in the project or other existing software and specifications, align with the research topics covered in UPCAST. On the basis of the initial requirements described in chapters 4, 5, 6 and 7 of this report, these software solutions will be extended to address the requirements and ultimately the objectives of UPCAST.

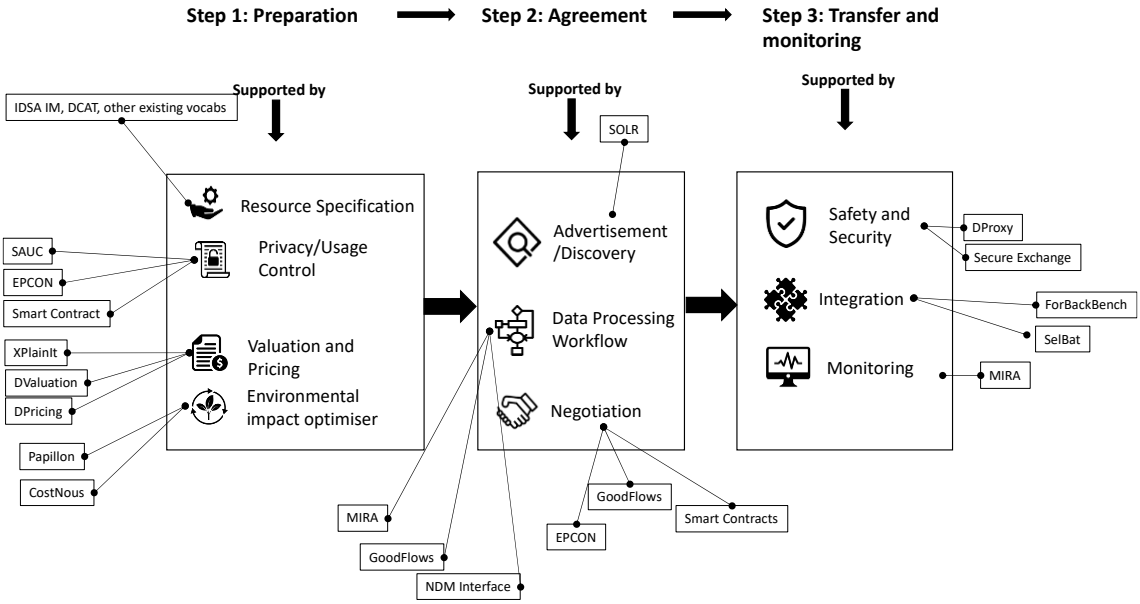


Figure 1. Research topics in UPCAST and existing software solutions.

2.1 Resource Specification

The problem of specifying an UPCAST resource is the same as the problem of generating descriptive metadata about datasets and data apps, services or operations. The technical challenges are (1) Choosing the appropriate schema for the description, balancing expressive power with interoperability and reuse requirements across different systems and platforms; (2) Quality and consistency, that is, ensuring the values of the properties of the chosen schema are complete, accurate, consistent and up-to-date, while balancing the amount of effort required to do so (3) Annotation tools that support the metadata generation process. These tools can be classified as “manual”, that support a user in creating the annotation but may not scale to large numbers of resources or generate inconsistent annotations depending on the user that creates them (Corcho, 2006) and “(semi-)automatic” where an algorithm is tasked with generating the annotations. Algorithms are more scalable in the sense that they reduce the manual effort in annotating large datasets or multiple datasets. However, the development of

such algorithms requires access to valid annotations for training or space search exploration before being useful. They also need to be customised for different output schema properties (Park & Brenza, 2015).

One of UPCAST's main objectives is *Universality*, that is, outputs should be useful to as many stakeholders and data spaces as possible. Our initial requirements elicitation from the pilots reveal they all share the need for metadata generation but at different granularities and often requiring domain-specific vocabularies. We also identified that for relatively small companies and public sector departments there is not so much need of scale in number of datasets, but rather to enable the exchange or sale of a few key datasets as fast and efficiently as possible, either with already known business partners or through a data marketplace. This leads to two additional technical challenges (1) universality and interoperability, (2) Ease of use and speed-up the annotation process.

Our solution is comprised of two components. First, the UPCAST vocabulary, that we will release as a schema to enable general description of datasets and data processing operations. We will re-use as many classes and properties as possible from the established DCAT[1] and IDSA core vocabularies (Mader et al., 2022) (so to ensure interoperability with ongoing efforts) while identifying extensions aimed at facilitating its use for as many organisations as possible, based on the requirements collected. The UPCAST vocabulary will also include properties that link a resource to their privacy/usage constraints, pricing, and environmental impact estimation that can be set following the output of the corresponding UPCAST plugins. We reviewed the state of the art in vocabularies for the following UPCAST plugins:

- Privacy and Usage Constraints: Efforts have been made to model GDPR specific constraints, SPECIAL (P. Bonatti et al., 2022) allows the expression of both the data subjects' consent and the data usage policies of data controllers in formal terms. IDSA suggestions (MyDATA, LUCON, and D0) stem from the Open Digital Rights Language (Iannella & Villata, 2018), a W3C recommendation aimed at representing permitted and prohibited actions over assets, as well as the obligations required to be met by their users. In addition, policies may be limited by constraints (e.g., temporal or spatial constraints) and duties (e.g. payments) may be imposed on permissions. Consortium's existing technology uses the BPR4GDPR vocabulary (G. V. Lioudakis et al., 2020). Refer to section 2.4 for further information on proposed extensions to these approaches.
- Pricing and Valuation: The GoodRelations ontology (Hepp, 2008) covers products, offers, points of sale, prices, terms and conditions. It has been used extensively for e-commerce and is supported by Google and other search engines. IDS Information model (Mader et al., 2022) includes the most basic pricing models (under the name "Payment Modalities" We expect some of the properties to be re-usable for the UPCAST vocabulary, but extensions will be required to account for pricing models exclusive to datasets. These extensions will be informed by ongoing work in the consortium on analysis and characterisation of existing data marketplaces (Azcoitia & Laoutaris, 2022). In terms of pricing of data processing operation, there are related vocabularies for Cloud Services (R. Greenwell et al., 2016) and general Web Services (Lamparter & Schnizler, 2006).
- Energy consumption: Formal ontologies have been developed in the context of general Cloud Computing (L. Youseff et al., 2008). Other works have provided ontologies to support the deployment of low carbon networks of ICT resources

[1] https://www.w3.org/TR/vocab-dcat-2/

(A. Daouadji et al., 2010). Our initial analysis from pilot requirements suggests that in the context of UPCAST there is no need for an extensive ontology, but to translate the terminology and units specific to the energy consumption of data processing operations and the service and transfer of datasets across a network. Refer to section 2.6 for more details.

The second component is the Resource Descriptor, a manual annotation tool specifically designed to help owners to produce metadata for datasets and data operations. The tool will enable annotation using the UPCAST vocabulary. To make the tool as universal as possible, it will enable annotation with user-defined domain-specific vocabularies to allow the support of multiple data exchange contexts. Examples of such contexts are the use of a recommendation or standards specific to the sector of the data space or marketplace where the transaction will take place, and the use of pre-established vocabularies between already existing partners. Data annotation tools for text, image and video data have been developed for the purpose of collecting ground truth to train machine learning algorithms (Simon et al., 2017)(B. Pande et al., 2022). Hinze et al. (2012) performed usability studies of a semantic annotation tool and provided recommendations to designers, while Khalili & Auer (2013) surveyed user interfaces in the related field of Semantic Content Authoring.

The Resource Descriptor and the domain-specific vocabulary support will be demonstrated in each pilot. We already surveyed existing efforts for each sector in the context of each pilot:

- Genomics: Several established efforts like OBI (Bandrowski et al., 2016) and the Gene Ontology (Carbon et al., 2019). The Gene Ontology participates in the Open Biological and Biomedical Ontology Foundry (OBO Foundry) (Smith et al., 2007) providing a suite of interoperable, free and open-source tools for sharing scientific knowledge in the domain of biology. The OBO Foundry puts in place various ontologies, starting from a set of upper ontologies: a) Basic Formal Ontology (BFO), being the upper-level ontology upon which OBO Foundry ontologies are built; b) Core Ontology for Biology and Biomedicine (COB), that brings together key terms from a wide range of OBO projects to improve interoperability; and c) Relation Ontology (RO), defining the relationship types shared across multiple ontologies. Upon these fundamental structures, a large number of ontologies have been based, spanning various domains of biology, including biological systems, microbiology, biochemistry, phenotype, etc. A category of ontologies under the OBO Foundry is headlined as “Information” and provides vocabularies for enriching data with multiple information types. An ontology in this family with particular interest for UPCAST is the Data Use Ontology (DUO) (Lawson et al., 2021), maintained by the Global Alliance for Genomics and Health (GA4GH)[2]. DUO allows the semantical tagging of datasets with restriction on their usage, making them automatically discoverable based on the authorisation level of users, or intended usage. In this context, it comprises hierarchical vocabulary of data use terms most often used to denote secondary usage conditions for controlled access datasets. Structurally, DUO contains 25 terms representing two types of data use terms, notably permissions and modifiers.
- Environmental public sector data: For observation and sensors, the W3C recommended, Semantic Sensor Network Ontology[3] could be used to represent

[2] https://www.ga4gh.org/
[3] https://www.w3.org/TR/vocab-ssn

devices and their measurements of various observations such as speed, car locations, the number of cars at specific locations, congestion, movement orientation and travel time duration (floating car data). Sensor Model Language (SensorML)[4] and Observations and Measurements[5] are also widely used for sensor observations, thus they could be also considered, as complementary. Additionally, for all measurements that do not refer to environmental data and do not include any sensor information, such as demographics, urban or other transport statistics, the RDF Data Cube Vocabulary[6] can be used as a starting point in order to create a more complete ontology that represents the data viewpoints. For administrative subdivision naming, it is desirable the use of the NUTS thesaurus[7].

- Digital Marketing: To the best of our knowledge, there are no formal vocabularies for this. We plan to create a lightweight one for the use cases.
- Fitness: No specific vocabulary has been identified to be close to the requirements, some classes and properties can be reused from existing ontologies developed for medical wearables (Kim et al., 2014).

Regarding the description of data operations, i.e., transformation, aggregation analytics, cleaning or integration, the IDSA core vocabulary (Mader et al., 2022) includes the concept of "Data App", a containerised piece of software that processes a dataset, and can be deployed in a Cloud provider. Our initial requirements elicitation revealed our pilots do not have their data processes containerised, meaning we cannot mandate the use of "Data App". We plan to introduce a super-class to DataApp to enable the description of data operations that are not containerised. There is related work on Semantic Web Services. OWL-S (Martin et al., 2005) provides foundations for the description of Semantic Web Services, answering the question "What this does?". Fensel et al. (Fensel et al., 2011) provides further technological insight on definition and composition of Web Services. Contrary to Web Services, in UPCAST's context the discovery of data operations is intimately connected to the data they process; hence we will also define attributes to facilitate the match with datasets.

Another data operation attribute we identified as important is the type of processing operation it executes with respect to regulatory frameworks. We will explore the re-use of ontologies developed for GDPR annotation (SPECIAL and BDPR4GDPR) to enable data operation owners to annotate processing types, that then can be matched with usage policies at the dataset level.

The Resource Specification plugin will enable the generation of data profiles as an additional layer of metadata. Data profiling refers to techniques used across various data models to extract metadata and gain insights from the data, usually focusing on statistics and hidden relationships. The process produces a comprehensive summary that facilitates the identification of data quality concerns, risks, and overall trends.

Data profiling techniques may use ad-hoc methods like visually inspecting random data subsets or create aggregation queries, or apply more systematic approaches that infer structural information and statistics from a dataset (Abedjan et al., 2017). Methods for profiling semi-structured data, specifically JSON data, have been presented in (Loetpipatwanich & Vichitthamaros, 2020) and (Möller et al., 2019). Challenges for

[4] https://www.ogc.org/standard/sensorml/
[55] https://www.ogc.org/standard/om/
[6] https://www.w3.org/TR/vocab-data-cube/
[7] https://ec.europa.eu/eurostat/web/nuts/background

structured data profiling are for example related to dependency discovery and profiling for dynamic data. In the Big Data context, sampling techniques are utilized in big data profiling to reduce the computational burden and accelerate the data profiling process. (Z. Liu & A. Zhang, 2020) provides a survey of various sampling techniques for such big data profiling.

Most existing dedicated data profiling applications cover NoSQL data sources and do for example not explicitly address graph-based data. This should be considered by the data profiling task, as the UPCAST Vocabulary will re-use metadata concepts from the IDSA Information Model which uses a graph representation based on RDFS/OWL. Understanding *knowledge graphs and RDF data* is challenging as they may have very complex structures and use different vocabularies. Available approaches/techniques include:

- *Structural summarization* methods that generate a concise summary that facilitates data comprehension and visualization of complex graphs (Konrath et al., 2012), (S. Campinas et al., 2012) (Consens et al., 2015).
- *Pattern mining* methods that extract representative patterns from an RDF graph, producing informative and diverse summaries (Louati et al., 2011) (Q. Song et al., 2016) (Riondato et al., 2017).
- *Statistical* methods that generate quantitative summaries of an RDF graph to help users evaluating the dataset's usefulness such as LODSight (Dudáš et al., 2015) and SPADE (Diao et al., 2019).
- Hybrid approaches that combine methods from the structural, pattern-mining and statistical methods (Mihindukulasooriya et al., 2015) (Čebirić et al., 2015).

A number of tools and applications for data profiling are available, such as:

- Data profiling features are offered in some commercial graph-based data management applications, such as Neo4J[8], Stardog[9], TopQuadrant TopBraid Enterprise Data Government[10], PoolParty Semantic Suite[11].
- Data profiling tools for RDF data on the web like Linked open Vocabularies (LOV),
- JSON schema generators such as Python libraries[12], NodeJS libraries[13].
- Tools support data profiling to improve data quality or support data cleaning: Trifacta[14], Atlan[15], Kylo[16], Open Source Data Quality and Profiling[17].

Few marketplaces provide profiling analysis for datasets. The UPCAST Resource Specification plugin will explore existing data profiling tools and solutions to provide dataset profiling which can be utilized in e.g., the pricing plugin. Providing dataset preview/samples may however be technically challenging.

[8] https://neo4j.com/
[9] https://www.stardog.com/
[10] https://www.topquadrant.com/
[11] https://www.poolparty.biz/metadata-management
[12] E.g., *https://github.com/gonvaled/jskemator*, *https://github.com/perenecabuto/json_schema_generator*
[13] E.g., *https://github.com/easy-json-schema/easy-json-schema*, *https://github.com/aspecto-io/genson-js*
[14] https://www.trifacta.com/
[15] https://atlan.com/
[16] https://kylo.io/
[17] https://sourceforge.net/projects/dataquality/

2.2 Resource Discovery

Dataset Search has emerged as a separate sub-field from databases and Information Retrieval aimed at closing the gap between what datasets are available, what dataset a user needs, and what datasets a user can find, trust and is able to use (Chapman et al., 2020). The advent of "Data Lakes", which are massive collections of datasets that grow over time and that are consumed on demand instead of being extracted-transformed–loaded into curated schemas have motivated the research and development of dataset discovery approaches (Nargesian et al., 2019). As data lakes, Data Spaces have the same challenges in terms of number and heterogeneity of datasets with the additional difficulty that datasets are owned and controlled by different organisations and their location is often decentralised.

The UPCAST approach to dataset search and discovery is to facilitate the task of generating metadata for resource owners. This is achieved through the UPCAST vocabulary and the resource description component described in section 2.1.

Regarding data operations, we will build upon the descriptions defined in section 2.2 and provide a system that supports different search styles on top of the union of dataset and data operation descriptions, i.e. resource catalogs. First, the classical keyword and filter interface on top of the metadata fields, with particular emphasis on the main UPCAST innovations: privacy/usage constraint, and pricing and energy consumption. Second, the search by example paradigm (Mottin et al., 2016) in which the user inputs an example of the resource they want in the form of an incomplete description with the UPCAST vocabulary. The combination of traditional search and search by example enables a third paradigm: Search by a Data Processing Workflow (see section 2.3), where the system can recommend to the user alternatives to resources that have been previously bought or used.

We plan to demonstrate resource discovery in two ways: first, on top of data catalogs prepared by the pilot partners using the resource description tool, first within a single data marketplace; and second, allowing for discovering resources across marketplaces. Regarding the latter, we cannot expect data marketplaces to adopt UPCAST vocabularies and resource descriptions in the project's lifetime. Therefore, we will resort to the classical technique of web scraping to collect information from already existing commercial data marketplaces to create a snapshot of a crawl of the datasets from multiple data marketplaces conducted by LSTech for the Pricing module. In this context, the idea is to demonstrate the plugin's usefulness for a data broker that collects a combined data catalog across multiple data spaces and wants to offer it as a service.

Web crawlers have been developed to render and download web pages presenting data products publicly available in commercial data marketplaces, following common crawling best practices. Such crawlers scrape available metadata from data products, such as the product id, title, description, source, seller and, when available, its geographic scope, volume, category, use cases, update rate, historic time span, format, etc. (Hils et al., 2020). A preliminary exercise allowed the collection of information about more than 200k data products from more than 2k data providers in 10 different marketplaces (Azcoitia, Iordanou, et al., 2022).

A limitation of this approach is that the amount of information available will be limited by the information that the data marketplaces make publicly available in their websites. Data management and processing will be done to upload all the information to a common cross-marketplace database in the pricing plugin, and this information will be

exported to the Resource Discovery plugin so that data products from those marketplaces are made discoverable through this tool.

2.3 Data Processing Workflow

A data processing workflow is a structured approach to the management of data, including collection, annotation, discovery, availability, and usage. A workflow comprises a series of actions that need to be performed to meet the objectives of a task for which the workflow is defined.

In UPCAST data processing workflows comprise actions that relate to pre-processing and processing of datasets. For example, all preparatory tasks of annotating a dataset with pricing information and its environmental footprint, negotiating its access and usage policies, are parts of a data processing workflow. Access to a dataset as well as its processing though enforcement of negotiated usage policies are also actions of a data processing workflow.

Workflows are modelled with the use of (workflow) diagrams. These models may contain sequential or parallel executions of actions, decision points, iterations, timing constraints and so on. Typically, a workflow has a start point and one or more end points. Actions are atomic operations or transactions that may be performed on a dataset, like reading, searching, calculating statistics, but otherwise their internal structure or behavior is not further detailed.

A workflow model, once created, can be used for carrying out the steps of the workflow. The execution of a workflow is the application of the workflow model to a real scenario, by which, the actions of the workflow model are executed, either sequentially or in parallel. In this real scenario different execution paths are selected at decision points based on the outcomes of previous executions of actions, until the end of the workflow model is reached.

The execution of a workflow can be done either manually or automatically with a workflow engine, which is a processor that is responsible for automating the execution of workflows. Given a representation of a workflow the processor executes it, starting the execution from its start point, following the flow of execution until it reaches an end point. There are various workflow engines available, both proprietary and free or open source. Examples include: ProcessMaker[18], Bonita BPM[19], Camunda[20].

UPCAST supports both the modelling and the execution of data processing workflows. Tasks that can be modeled and executed through the UPCAST data workflow engine are the following:

- Dataset annotation with metadata that allows efficient discovery, e.g., “financial data”, “stock exchange open and close prices”, etc.
- Dataset annotation with metadata that allows efficient execution.
- Attaching pricing information to a dataset.
- Attaching environmental footprint information to a dataset.
- Attaching access policies to a dataset
- Attaching execution policies to a dataset

[18] https://www.processmaker.com/
[19] https://www.bonitasoft.com/
[20] https://camunda.com/

- Negotiating usage parameters of a dataset
- Attaching contract parameters to a dataset

The UPCAST support for data workflow modeling and execution will be provided by MIRA, a technology for Digital Twins Maggioli brings to the project. MIRA is a Digital Twin enabler platform that provides core features allowing the user to *define its environment* in terms of:

- Digital assets (with related properties and telemetries)
- Relationships between individual assets
- Networks of related assets

Assets are any entities of interest, which are subject to monitoring. Physical assets are modelled by their digital counterparts, which are processed by MIRA. Assets have related properties, i.e., specifications of parameters of interest that can be monitored. The actual values of properties that are obtained through the monitoring process of physical assets are called telemetries. The monitored values of parameters (telemetries) are associated with the digital representations of assets for further visualization and analysis by MIRA.

The digital representations of assets that are modeled and processed by MIRA can be used to:

- Track, store, manage and monitor the physical asset's data in a secure and structured manner.
- Automate and optimize the asset's operations.
- Simulate different scenarios by using analytics and machine learning methods to predict future behavior and thus support improved decision making.
- Create digital identity for physical assets, enabling secure and seamless access of data from multiple devices.

MIRA can naturally model datasets and data workflows as assets and networks of assets, respectively, whereby assets model datasets and networks model workflows. Moreover, MIRA will be extended to support the execution of data processing workflows of UPCAST as well.

Alternatively, GoodFlows provided by Abovo can be used for data processing workflow modelling. GoodFlows provides a generic compliance-oriented solution to modelling different kinds of processes with an easy-to-use no-code design and editing tool. Its foundations lie in a customisable Information Model, able to formally codify the data model of an organisation in sufficient detail, and in a Compliance Metamodel, allowing to accordingly formalise workflow models taking a variety of execution and compliance aspects into consideration; both the Information Model and the Compliance Metamodel are implemented as ontologies. More specifically, GoodFlows workflow editor features can more specifically be summarised as follows:

- It enables the comprehensive specification of workflow elements, providing extensive coverage of all three core workflow perspectives, i.e., control, data and resource; to this end, beside *actors, operations* and *information*, it also introduces the novel concept of *assets*, as a means for explicitly representing the entities being subject to the execution of workflow tasks.
- It allows the explicit modelling of both control and data flows, thus being suitable for applications based on either of them or both of them combined.
- Workflows are defined as ontologies; this, apart from the inherent benefits regarding formal and machine-interpretable semantics, offers the additional

advantage of their direct and transparent integration with the ontological Information Model, but also any other ontologically defined vocabulary or knowledge base in general. This means that the solution can be tailored to different domains in a standardised way, supporting semantic consistency as well as inference of knowledge not explicitly contained in workflows themselves.
- Its high expressiveness provides for the *in-design* expression of sophisticated security and privacy constraints.

The above may be used in UPCAST through defining DPWs by means of the Compliance Metamodel and the respective Editor functionality, including of course the necessary alignment of the Information Model to the needs of UPCAST and its plugins. Such environment would allow a potential data consumer to define intended data-centric processes alongside specific usage requirements, forming a suitable user interface as the basis for further negotiation.

2.4 Privacy and Usage Control

Traditionally, agreements between data providers and consumers were written in natural language, leaving virtually no possibility of automating this process. In recent years, there have been several efforts to introduce frameworks and technologies to allow users (both providers and consumers) to define their data access and usage policies, among other things, in standardised formats that are machine-readable, and thus are able to be processed automatically. For instance, the W3C has presented the Data Privacy Vocabulary (DPV)[21], which defines an ontology that allows for the definition of the use and processing of data under specific legislation (for example, the GDPR).

Efforts to protect the privacy of data subjects can be mostly classified in two approaches: statistical and logic based. Statistical approaches are those that protect users' identities by use of statistical noise to obfuscate individual data that is part of larger aggregations (Dwork, 2008) (Machanavajjhala et al., 2007) (Sweeney, 2002). On the other hand, logic-based approaches are those that protect individual data by determining if certain (sets of) queries may compromise the identity of users, and then filtering them (Bertossi & Li, 2013) (P. A. Bonatti & Sauro, 2013) (Chirkova & Yu, 2017) (Grau & Kostylev, 2016) (Rizvi et al., 2004) (Rosenthal & Sciore, 2000).

Automated contract verification usually assumes an adversarial relationship between provider and consumer. This approach leads to coarse-grained agreements that establish "all or nothing" policies wherein agreements are not possible unless the restrictions and requirements of each side aligned perfectly. Evidently, this leaves no room for a compromise such that the requirements of both parties are partially fulfilled while respecting their imposed restrictions. This lack of flexibility and dynamic agreement may be useful in certain domains and circumstances, but not in cases where parties are collaborating. In addition, these agreements are usually defined with top-down policies that protect the interests of organisations rather than the individual data subjects.

[21] https://w3c.github.io/dpv/dpv/

Effective access and usage control are deemed critical for data protection. As part of UPCAST, we wish to follow the example of the study found at (Konstantinidis et al., 2021), where instead of outright rejecting agreements that do not match entirely, we compute the "overlap" between requirements and restrictions, and then return this overlap where we may remove tuples of data that have no consent.

In this direction, UPCAST will provide a semantic access and usage control framework, and incorporate the BPR4GPPR (G. Lioudakis et al., 2021) and EPCON systems for handling privacy constraints in relational databases[22] (Konstantinidis et al., 2021). The primary objective of the plugin is to address access, usage and privacy requirements for distributed environments in a holistic and comprehensive manner. This will be achieved by a bilateral agreement process: The data owner will have the possibility to state their consent, privacy preferences, usage and access control policies.

The data consumer will be able to specify intentions, processing scenarios, purposes, and business processes about the use of the data. The plugin will reconcile the two and offer a best effort reconciled privacy, access and usage agreement.
Managing consent in a data workflow involves intricate algorithms to ensure compliance with privacy constraints. Although this poses a challenge as it is an NP-hard problem, the privacy plugin can leverage previous research (Konstantinidis et al., 2021) (Filipczuk et al., 2023) to explore heuristic approaches and algorithmic solutions for addressing privacy requirements effectively.

Incorporating the BPR4GDPR framework as a constraint generation tool, the privacy plugin within the UPCAST system will introduce innovative approaches to resolving privacy conflicts that may arise between data consumers and data providers. It is important to consider scenarios where privacy concerns may need to be adjusted, either enhanced or scaled down. To address these conflicts, the negotiation plugin will play a pivotal role, ensuring that conflicts are appropriately represented and solution scenarios are provided. The framework is built upon a comprehensive Semantic Information Model or Knowledge Graph that offers an abstract representation of the basic entities of distributed systems and the relations between them. This model is firmly grounded on the requirements derived from the elaboration of legal and regulatory provisions regarding data protection.

The UPCAST semantic policy-based access and usage control framework offers a flexible means of specifying rules that govern the entities within the information model. This framework provides a high degree of expressiveness and allows for the specification of constraints that are not accommodated by existing models. In addition to the typical pattern of specifying which user holding which role performs which action on which object, the framework incorporates a variety of aspects, including purpose, attributes, context, events, actions that must have taken place before or actions that must take place after the enforcement of a rule, obligations, etc.

The key features of the semantic policy-based access and usage control framework include:

- Multi-aspect access and usage rights definition
- Purpose-based rules and privacy awareness
- Extended attribute-based access control with sophisticated constraints
- Flexibility in rule expression and abstraction level

[22] https://github.com/georgeKon/enabling-personal-consent

- Context- and history- awareness
- Complex dependency and duty constraints
- Two-stage knowledge extraction for improved performance
- XACML compliance and interoperability
- Decision export capabilities for reporting

The framework will also provide reasoning with purposes and intentions state on the information/data model of the plugin. The knowledge extracted by the rules through reasoning can be leveraged in two ways. Firstly, for serving the role of a typical Policy Decision Point (PDP) by making real-time decisions on access and usage authorizations. Secondly, for providing appropriate guidance on the verification of processes and workflows as regards their compliance to the underlying policies, as well as their transformation in order to become compliant.

UPCAST's advanced decision making offered by the framework will allow the policy decision point to:

- Accept the request as is, possibly prescribing/forbidding the execution of other actions in the future;
- Accept a transformed version of the original request, by means of selection, projection and change of state of fields, possibly prescribing/forbidding the execution of other actions in the future;
- Recognise the conflicts in the request, present them in an intuitive and user-friendly way and proceed to the negotiation phase.

The privacy plugin will introduce innovative approaches to safeguard the privacy of data owners while also addressing the needs of data requestors. It aims to provide effective solutions that protect the sensitive information of data owners, while also satisfying the data requirements of requestors. By combining privacy-enhancing techniques and thoughtful data handling practices, the plugin will enable a balanced and privacy-conscious approach to data sharing and access.

2.5 Valuation and Pricing

Data marketplaces (DM) are becoming a key enabler of the data economy, fostering innovation, collaboration and competition among data stakeholders. They also enable the monetization of data assets to help data providers reach a wider audience of potential customers. The key challenges associated with pricing datasets are based on designing pricing policies and strategies that can capture the multi-dimensional information in the marketplace, which can include information from buyers, sellers, other data assets, intended use, the complexity and the heterogeneity of the data market (Andres et al., 2023).

To sort out the challenges faced by sellers and buyers at the time of setting the price for a piece of data, a market-based pricing tool will be developed in order to give a hint of the market price of a dataset or data product based on those of similar products in the market. A phased approach to development of the pricing and valuation plugin will be followed. Initially, the plugin aims to generate a static price of a dataset based on its metadata and similar products in the market.

Figure 2 summarises the high-level diagram of the envisaged tool for static pricing:

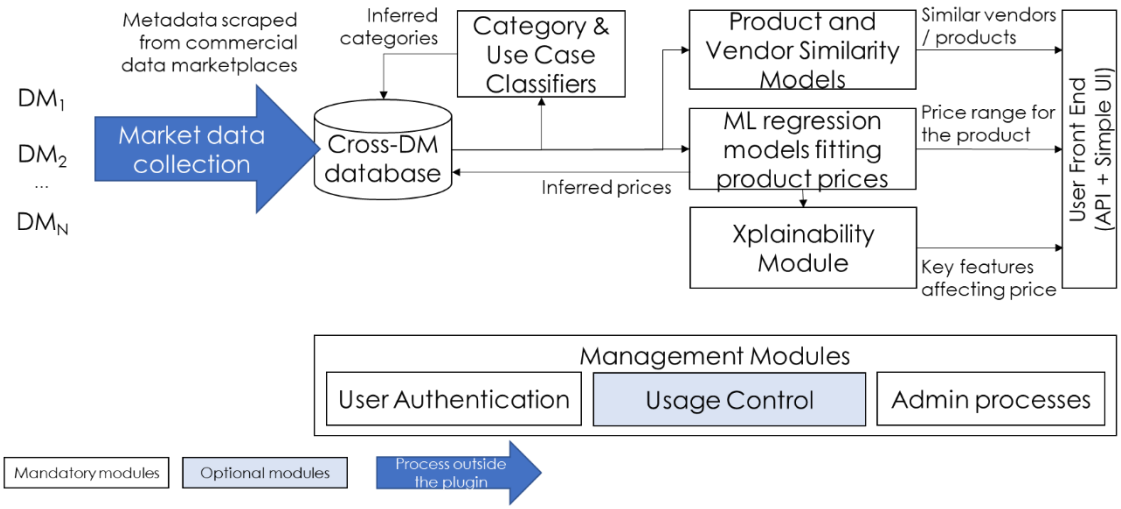


Figure 2. The envisaged tool for static pricing

There are already a number of data marketplaces and data providers offering products in the market. Based on information scraped from those public data marketplaces, a cross-DM database will be created containing metadata of those products and their prices. A preliminary capture of this information has produced information about 200k data products from more than 2,1k vendors and contains more than 10k price references. These datasets will be used to train the ML models used in the pricing plugin. Since DMs use different categories and criteria to label data products, Machine Learning classifiers will be built to learn the criteria used by a source DM and be able to apply them to data products in other DMs, thereby homogenizing the labelling and enriching the whole sample.

Basically, the tool will provide three key functionalities to other external modules and use cases of UPCAST:

1. Given a product specification, it returns similar products found in other data marketplaces that can be used by buyers or sellers to price their data or to learn how other providers are structuring and monetizing their data product offer.
2. In addition, it will produce a range of price estimations based on those of different regression models fitted to the prices and descriptions of real products in the market.
3. Finally, it will also provide some explanations to the price range produced, identifying the features that are responsible for that prediction.

This tool provides a static price reference to be considered by buyers and sellers when bidding for or setting a price on data products. It produces static starting prices to "break the ice" and circumvent the problems of companies that do not know how to fill in the "price" of their data products in marketplaces. However, as a market-based tool it has some limitations:

- It depends on market information, and therefore is limited to the type of products being offered in the market,
- It does not take into account the cost of producing the data or the different utility of buyers in using it, only metadata features as specified by data providers,
- It produces a static price that will be refreshed once the cross-DM database is updated.

After the static price reference is provided, it will be used as an input to create dynamic pricing models, which is the next phase towards the development of the plugin. This will incorporate market conditions and other marketplace interactions to produce a dynamic price range of datasets. Finally, as more datasets and historical transaction data are available and fed into the models, the final pricing plugin will be developed, which will support the data selection and negotiation strategies.

The pricing component will use explainable AI (XAI) multimodal tools (based upon CDR's XplainIT solution) to provide transparent feedback about the algorithm's decision to price a dataset or combination of datasets, e.g., what are the features that contribute to increasing or decreasing the price of a certain data product, or why some resources are more valuable than others for certain workflows. CeADAR's XPlainIT is a web-based tool which offers a series of functionalities for explaining the process of machine learning and deep learning models in a user-friendly manner. It is about opening the 'black box' decision making of machine learning algorithms so that decisions are transparent and understandable.

Once a transaction is closed, some use cases are in need of distributing (part of) data payoffs among the data sources that contributed to the data transactions. Commercial marketplaces resort to simple heuristics to carry out this distribution, such as the volume of data, or share payoffs equally among data sources. There is a wide research literature that studies how to carry out this distribution process according to the "value" that each data source brings to the transaction to incentivise the provisioning of better-quality data. The problem is that this value depends heavily on the task the data will be used for, given by a machine learning (ML) model M and an accuracy / valuation function v to measure how good the resulting model is.

Most papers resort to the Shapley value (Shapley, 1952), a well-known concept in game theory that calculates the average marginal contribution to the value function v of a data source when combined with any possible permutation of the rest of data sources in the transaction. The Shapley value is widely acknowledged as a "fair" method to distribute the value of a game between the players of a coalition in the game theory and ML literature, due to its remarkable "fairness" properties (efficiency, symmetry, linearity, null-player) (Rozemberczki et al., 2022).

However, the exact calculation of Shapley value is complex for N data sources - O(N!) or $O(2^N)$. Therefore, a number of approximation methods are available in the literature (Ghorbani & Zou, 2019) (Jia, Dao, Wang, Hubis, Gurel, et al., 2019) (Jia, Dao, Wang, Hubis, Hynes, et al., 2019). That efficiency comes at a cost. First, they lose precision in the calculation. Second, they lose generality. Even though the Shapley value is a general concept, some of its approximations work well for certain problems and not so well for other problems depending on the behaviour of v with data sources (Azcoitia, Paraschiv, et al., 2022).

As part of the work on data valuation and pricing plugin, we will test existing data valuation techniques on specific problems specified by the use cases of UPCAST to find out the value of sources of data contributing data to a transaction.

2.6 Environmental Impact Optimiser

High-performance computing (HPC) generates a large amount of carbon emissions due to its high energy consumption. As the digital economy relies on data centres to process and store massive amounts of data, improving their energy efficiency is crucial for environmental sustainability.

In relation to green or environmentally efficient computing, UPCAST builds on existing mathematical analytical models using features of resources such as energy consumption, carbon footprint, hardware, computational, properties, location, network characteristics, and others, to improve data operations, give datasets an energy profile and optimise entire workflows. UPCAST will provide a module to estimate the environmental impact of a data processing workflow, beginning with the dataset itself and how it is processed throughout the workflow. With this aim, the environmental optimization plugin will leverage a system that CeADAR has access to, which is a unique and innovative, comprehensive energy management system for data centres, measuring energy consumption at rack, server and application-level in real time.

For data and its processes running on cloud platforms, the plugin will leverage know-how from LST's CloudNous. CloudNous has defined data collection and processing tasks that allow to learn and correlate the usage of cloud resources to the services and to the users that make use of those resources. CostNous retrieves cloud service usage from APIs and consoles of cloud service providers and implements an innovative mechanism to track the usage of the platform by the underlined systems and its users. For example, in the case of web services this is done by means of placing a cookie in users' browsers and detecting these cookies in connecting to the activity taking place in the system. This allows for causal cost analytics and, with appropriate data modelling energy consumption of cloud platforms, energy consumption analytics for a number of public cloud platforms.

The plugin will include efficient AI models and visualization tools of the environmental cost and the energy requirements of data processing workflows based on hardware, location, computational needs and other features. The plugin also aims to assess the energy efficiency of datasets, through energy profiling, and provide the end user with useful information about the possible environmental impact of using a particular dataset. Further, measurements related to network transfer will be investigated, although this is anticipated to be a difficult task.

Access to High-Performance Computers from the partners (e.g., CDR's HPC) will be key to testing, deploying, and validating the energy footprint modelling in various processes dealing with processing, storing, updating, and serving data.

2.7 Integration and Exchange

In the last decade, the amount of data collected and then published by various entities has grown in an unprecedented manner. Similarly, the processing of data has evolved and become more complex and more of it is required. However, entities that need data cannot always produce the very data they need on their own. Therefore, they must look for other sources of data to fill this need.

Data integration is the process through which data from different, heterogenous, and distributed sources are successfully included in a global schema. On the other hand, data exchange is the process in which data is transformed from a source schema into a target schema describing the same information.

One of the main problems of data integration and exchange, is the computation of queries over said data. The two main approaches to this are the materialization of sources and query rewriting. Materialization of sources (or forward-chaining) consists in populating our target dataset by generating new tuples with data from a source view. This is usually achieved through a family of algorithms known as *chase* that infer new data based on predefined constraints (such as functional dependencies). Although these algorithms will often produce the intended result, they present some limitations that must be considered: the chase may not terminate for some constraints and does not scale well with large databases. Query rewriting (or backward-chaining) consists in reformulating queries according to certain constraints to extract the data directly from its source(s) without materializing new views. Because query rewriting does not generate additional tuples, and scales better with the size of the data, it is often seen as the more practical of the two approaches. Nevertheless, in practice many common scenarios have structures that allow for efficient and scalable runs of the chase algorithm (such as the *frugal chase (Benedikt et al., 2017)*).

UPCAST will address the users' needs to compile data from various sources found in the marketplace and provide data-warehousing to users to move data while respecting all privacy and pricing negotiated conditions. The *ForBackBench* benchmark will serve as a starting point for this plugin, since it contains code for several common Data Integration/Exchange scenarios such as the generation of mappings, visualisation of graphs, and query translators, among others (Alhazmi et al., 2022) (Alhazmi & Konstantinidis, 2022).

In the case data is sold on a per-query basis, or for when administrative, geographical or other constraints do not allow for "forward" movement of data we will deploy a virtual integration approach where SPARQL or SQL queries are posed over the central vocabulary and a query rewriting algorithm uses the mappings to rewrite them as queries over the source schemas/endpoints/APIs. All query rewritings will be achieved using GQR, a state-of-the-art algorithm for the query rewriting problem following a graph-like representation of the materialised sources (Konstantinidis & Ambite, 2011).

In addition, the aforementioned benchmark will be complemented with a graphical interface to aid users that have only domain specific knowledge. This will allow for the definition of schema mappings in an intuitive manner that requires little effort in designing schema mappings. The tools found in *ForBackBench (Alhazmi et al., 2022)* will serve as the backbone of this plugin while UPCAST will provide a user-friendly interface that seamlessly integrates the various functions of this plugin.

To obtain a "global schema" as described above, the schema (or ontology) of the external data sources (target) must be mapped to the source schema, a task that requires semantic alignment of the source and target(s) schemas. The alignment process can be supported using different semi-automated techniques depending on the source and target representation (schema matching, ontology matching, instance matching/link discovery) and will typically result in machine-readable alignment[23] that contains either basic or complex mappings that can be used by either forward-chaining or backward-chaining operations. Often, contextual information related to the source and target schemas/ontologies to be aligned is sparse, making the process of automatically inferring mappings difficult. To remedy this, we will use techniques from

[23] The Alignment API is a defacto standard representation of ontology mappings: https://moex.gitlabpages.inria.fr/alignapi/

the field of Semantic Table Interpretation (Liu et al., 2023) to support the alignment process. STI aims to automatically associate elements in a semi-structured data source (typically in a tabular format) to entities in knowledge graphs and can be applied to enrich context of elements in the schemas/ontologies that must be aligned. This work will build on the STI tool s-elBat (Cremaschi et al., 2022).

2.8 Negotiation and Contracting

In recent years there has been a rising interest in data sharing. Often, there may be differences in what a consumer interested in a specific dataset wishes to do and what the provider of said data is willing to offer. This may range from the scale of data the provider is willing to share to an interval of time the provider is willing to share data during. Nevertheless, these differences are not necessarily unreconcilable, so both parties may still reach an agreement through negotiation.

There has been an increase in technologies that aid in various stages of the negotiation between data producers and consumers with the aim to reduce the human input needed. One such stage is the contracting, a process that historically needed humans to put in the effort and write contracts in natural language. Recently, there have been numerous efforts to model contracts with different approaches such as semantic-based models (O. Perrin & C. Godart, 2004) and ontologies (Kabilan & Johannesson, 2003), (de Cesare & Geerts, 2012), (Petrova et al., 2017). Some of these approaches are more restricted to a specific domain such as business and health, but the results may naturally be adapted to other domains.

In addition, systems have been developed to automatically generate contracts in both machine-readable formats, under standards such as the Open Digital Rights Language (ODRL[24]), and natural language using boilerplate text or even Large Language Models (LLMs). Furthermore, some of these systems have incorporated technologies such as Blockchain to enhance their privacy and security when managing electronic contracts (L. Guo et al., 2021), (Simić et al., 2021).

To address the verification of compliance, (Tauqeer et al., 2022) have worked in automated contract compliance using Knowledge Graphs. GDPR introduced some conditions that impacted the verifications process (Gangl, 2019), and subsequently (Doe, 2018) have proposed guidelines for the compliance verification from the perspective of law firm sectors.

The stage of the negotiation is where a system must administer the life cycle of a policy, detect conflicts between contracts, resolve said conflicts, etc. As part of a project by the Fraunhofer Cluster of Excellence Cognitive Internet Technologies (CCIS) negotiations that are both automated and autonomous were adapted specifically for data usage scenarios. In addition, GeniusWeb is an open architecture for negotiation via the internet and provides many reusable components, e.g., utility functions and negotiation strategies for participants.

An example of a negotiation process is presented in Figure 18 of the IDSA Position Paper (IDSA, 2021).

[24] https://www.w3.org/TR/odrl-model/

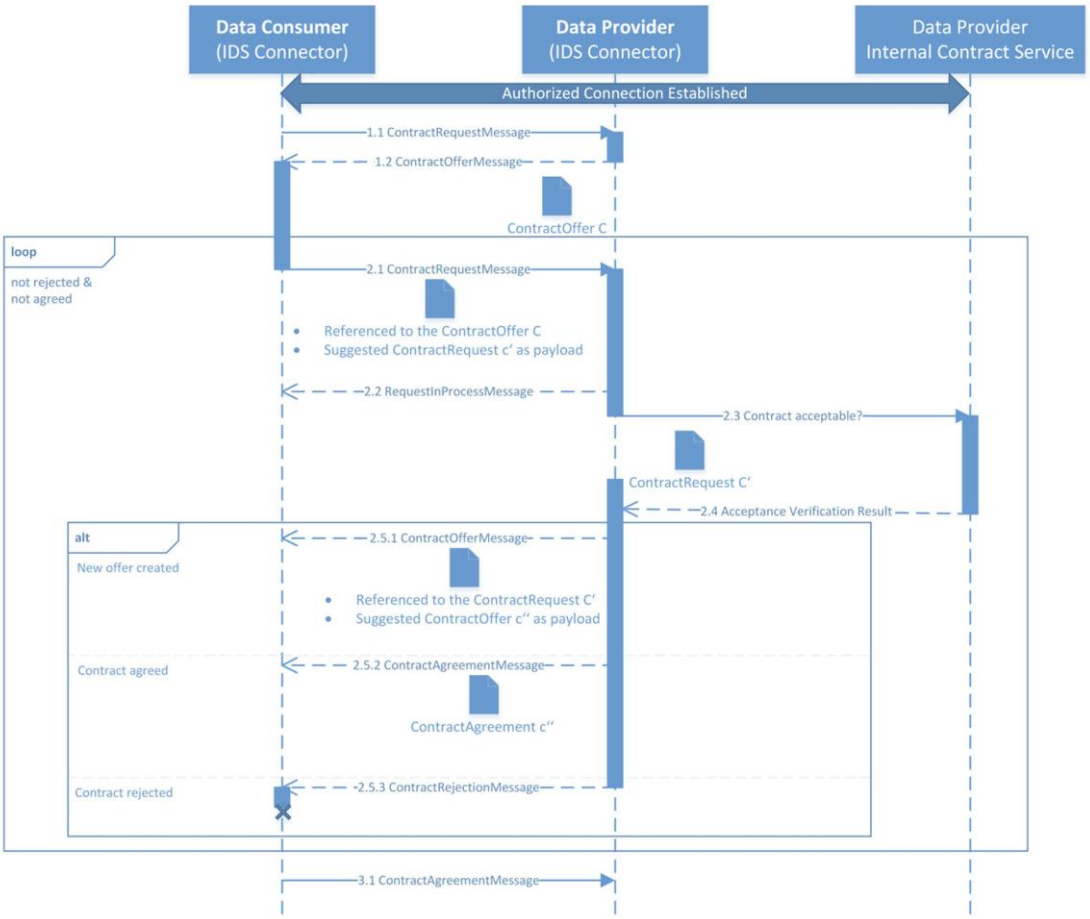


Figure 3. Example of a contract negotiation process from IDSA (IDSA, 2021)

Figure 3 illustrates that a data consumer requests use of a dataset from its provider. The data provider then responds to the request by sending an offer stating the conditions under which the data may be used (in the example, the data may only be used for a limited time). If there is a discrepancy between the provider and the consumer the process may be repeated until the data provider accepts or rejects the terms of the request contract. In the former case, an agreement contract is created by the system and sent to both parties for review.

GoodFlows (Carvalho & Lioudakis, 2020; G. Lioudakis et al., 2021; Papagiannakopoulou, 2020) has been developed with the aim to aid companies of various sizes and government agencies to ensure and prove their compliance to the GDPR. It essentially constitutes a process planning tool, which automatically assesses defined models in terms of GDPR compliance; it can also automatically re-engineer a non-compliant process model into a compliant one (if possible) by introducing appropriate transformations. GoodFlows may also be used for the negotiation tasks between data providers and data consumers.

UPCAST will allow users to create, offer, request and negotiate contracts. These contracts will mostly conform to the specification of usage control defined by IDSA (IDSA, 2021). This specification represents an extension of the Open Digital Rights Language (ODRL) that makes more descriptive and technology-independent contracts.

This plugin will serve as a Policy Management Point (PMP), which administers usage restrictions. It will read the machine-readable contracts from the privacy and usage, environment impact, and pricing plugins, and automatically reach an agreement in case there are no conflicts between contracts. Otherwise, a negotiation will be initiated where the data owner or the data user may present a counter-offer that may or may not be accepted by the other party. Ultimately, the owner has the final say on whether the negotiation goes through by either accepting, rejecting or sending another counter-offer. In addition, we will provide a Policy Administration Point with a user-friendly graphical interface so users can edit policies. This will allow users to define restriction, privacy, usage policies, etc.

UPCAST will combine the strengths of two previously demonstrated technologies - Southampton's toolkit for enabling personal consent (EPCON) and the BPR4GDPR framework from ICT-ABOVO, extending them to support business rules (for EPCON), and general compliance (for BPR4GDPR).

Although the overall approach is currently oriented towards privacy and the GDPR, it can be transparently extended to other types of compliance as well. In UPCAST specifically this may be exploited in order to compromise and enforce already *at design time* various different types of "policies" and constraints that may arise: i) first and foremost, the usage preferences of the data consumer versus the usage constraints imposed by the one or more data owners that may be involved in a DPW, and the subsequent negotiation results; ii) potential marketplace rules that may apply could also be of relevance, stemming, for instance, from regulatory sources, like the GDPR, the upcoming Data Act, the European Data Governance Act, etc.; iii) apart from data owner or marketplace constraints, it could be that also intra-organisational policies from the data consumer's side need to be respected, either regulation- or business-driven. Besides, it should be noted that, next to privacy and usage policies, also energy and pricing constraints could be addressed in a similar fashion. The above may indeed be achieved by appropriately adjusting and extending GoodFlows, in order to be able to assess and re-engineer DPWs for compliance with underlying policies by means of process verification against the latter, while incorporating negotiation over data consumers' and data providers/subjects' preferences/policies as part of the verification procedure.

The UPCAST negotiation plugin, in coordination with the privacy plugin, will be leveraged for the definition of machine-readable access and usage constraints, while the developed reasoning mechanisms will be employed for negotiating access and usage constraints between data providers and data consumers. In fact, access and usage constraints defined by different stakeholders prescribe negotiation, conflict resolution, combination/merging; the provided reasoning mechanism can offer negotiation over requestors' and data subjects'/providers policies, by employing a variety of prevalence schemes (most recent rule prevails, deny overrides, more strict rules prevail, Inclusion-Exclusion principle for comparing all kind of constraints, pre-actions and contextual conditions). The plugin will also negotiate based on pricing and environmental preferences of the two sides. It is anticipated that data providers' constraints will be reflected in the respective licenses (potentially in ODRL), which will be translated into the underlying semantic policy language in order for the negotiation to take place (and vice versa, for annotating datasets with the underlying access and usage policies).

2.9 Safety and Security

Safety and security in a Data Marketplace refer to the measures and practices implemented to protect the integrity, confidentiality, availability, and privacy of the data being exchanged within the marketplace. It involves ensuring that the data is handled in a secure manner and that appropriate safeguards are in place to prevent unauthorized access, data breaches, misuse, or manipulation.

Data marketplaces present several challenges in terms of safety and security as mentioned below.

- **Data Integrity**: Maintaining data integrity is critical in data marketplaces. Ensuring that the data has not been tampered with or modified in transit or storage is important for maintaining trust.
- **Trust and Authentication**: Establishing trust between data providers and consumers is crucial. Authenticating data sources and verifying the credibility and reliability of data are essential.
- **Data Quality:** Data quality is a significant challenge in data marketplaces. Ensuring that the data is accurate, reliable, and up to date is essential for making informed decisions.
- **Malicious Activities**: Data marketplaces are vulnerable to various malicious activities, including data breaches, unauthorized data access, and data poisoning attacks.
- **Legal and Ethical Compliance:** Data marketplaces must comply with legal and ethical standards. Adhering to regulations like the General Data Protection Regulation (GDPR) and ensuring ethical data use and sharing practices pose significant challenges.
- **Data Governance:** Managing data ownership, usage rights, and permissions in a multi-party marketplace can be complex.
- **Cybersecurity Risks**: Data marketplaces are potential targets for cyberattacks. Protecting against threats like data breaches, denial-of-service attacks, and malware is crucial. Implementing robust cybersecurity measures, regular vulnerability assessments, and incident response plans are necessary to mitigate cybersecurity risks.

2.9.1 Safe and Secure Execution

Safe and secure execution involves ensuring that data transactions between buyers and sellers occur in a way that minimizes the risk of data breaches, unauthorized access, or other malicious activities.

Some Important features are as mentioned below:

- Authentication and access controls
    - Buyers and sellers in the data marketplace should be required to authenticate themselves before participating in any transactions.
    - Smart Contract based Access controls can be implemented to ensure that only authorized users can get access to any sensitive data stream.
- Encryption
    - All data should be encrypted during storage and transmission to protect against unauthorized access or interception.
- Smart contracts
    - Smart contracts can be used to enforce the terms of data stream transactions and ensure that data access is only given to buyers who have met the specified conditions.
- Data privacy

    - Data seller should have control on how their data steam is shared.
- Auditability
    - All transactions should be logged and auditable to ensure transparency and accountability.
- Secure Computation
    - Computation in a data marketplace should be secure and resilient against attacks, such as tampering or denial-of-service. This can be achieved by using secure computing techniques, such as trusted execution environments, homomorphic encryption, and secure multi-party computation.

2.9.2 Secure Exchange

Secure exchange in a data marketplace refers to the secure and trustworthy exchange of data between buyers and sellers on the marketplace. This involves ensuring that the data is protected from unauthorized access and that the transaction is conducted in a way that is fair and transparent for all parties involved.
Secure exchange in a data marketplace involves protecting the confidentiality, integrity, and availability of data being exchanged between parties.
The solutions mentioned below can be implemented to ensure secure data exchange in data marketplace.

- Encryption
    - One way to secure data exchange is by using encryption. This involves converting the data into a coded form that can only be read by someone with the key to decrypt it. This ensures that even if the data is intercepted by unauthorized parties, they won't be able to read it.
- Access controls
    - Access controls can be used to ensure that only authorized parties have access to the data. This can include implementing strong authentication methods, such as two-factor authentication or biometric authentication, and limiting access to specific individuals or groups.
- Secure data transfer protocols
    - Using secure data transfer protocols such as HTTPS, SFTP, or FTPS can ensure that data is transferred securely between parties. These protocols encrypt data during transmission and help prevent unauthorized access.
- Data masking and tokenization
    - These methods can be used to protect sensitive data by substituting it with tokens or masked data that cannot be used to identify the original data.
- Audit logs and monitoring
    - Implementing an audit logging and monitoring system can help detect and respond to any unauthorized access or attempts to access data. It can also help identify potential security risks and improve security measures.
- Compliance with regulations
    - It is important to ensure that the data exchange process complies with applicable regulations, such as GDPR, CCPA, and HIPAA, among others.

Overall, implementing a combination of these solutions can help ensure that data exchange in a data marketplace is secure and protected from unauthorized access.

To implement Safety, Security and Secure Exchange of data, **Nokia Open Analytics Exchange (NOAX, earlier known as NDM)** will be used in UPCAST. Detailed information about NOAX is given in the following subsection.

2.9.3 Nokia Open Analytics Exchange (NOAX)

The Nokia Open Analytics Exchange is a decentralized – powered by Blockchain – data marketplace for the safe and automated exchange of digital assets in the form of data streams. The exchanged data streams are monetized and supported by technical and policy-based data verification. The unique function of the Nokia Open Analytics Exchange, both for generating revenue and for the exchange of data streams between interested parties in a business ecosystem is a private and permissioned Blockchain-technology that ensures network security, data integrity, and the use of smart contracts for fast and automated transactions (data streams). The Nokia Open Analytics Exchange can be used to stream any type of data from any source type such as: administrative, IoT devices, physical assets, autonomous cars, drones, and many more. It furthermore enables the business ecosystem to integrate third party data and to monetize it through the same marketplace. Analyzes and reporting are provided based on the Nokia Data Marketplace computing capabilities, about the continuous transformation in both the business ecosystem and the data assets. Figure 4 shows the core components of the NOAX system architecture.

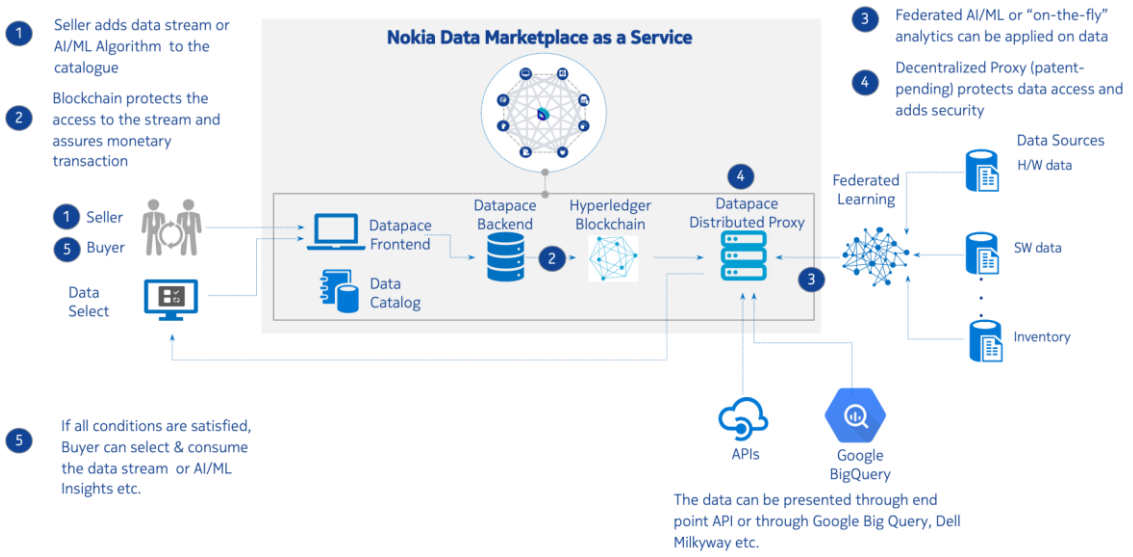


Figure 4. Illustration of the NOAX System Architecture.

The initial solution for secure data exchange may look like as shown below.

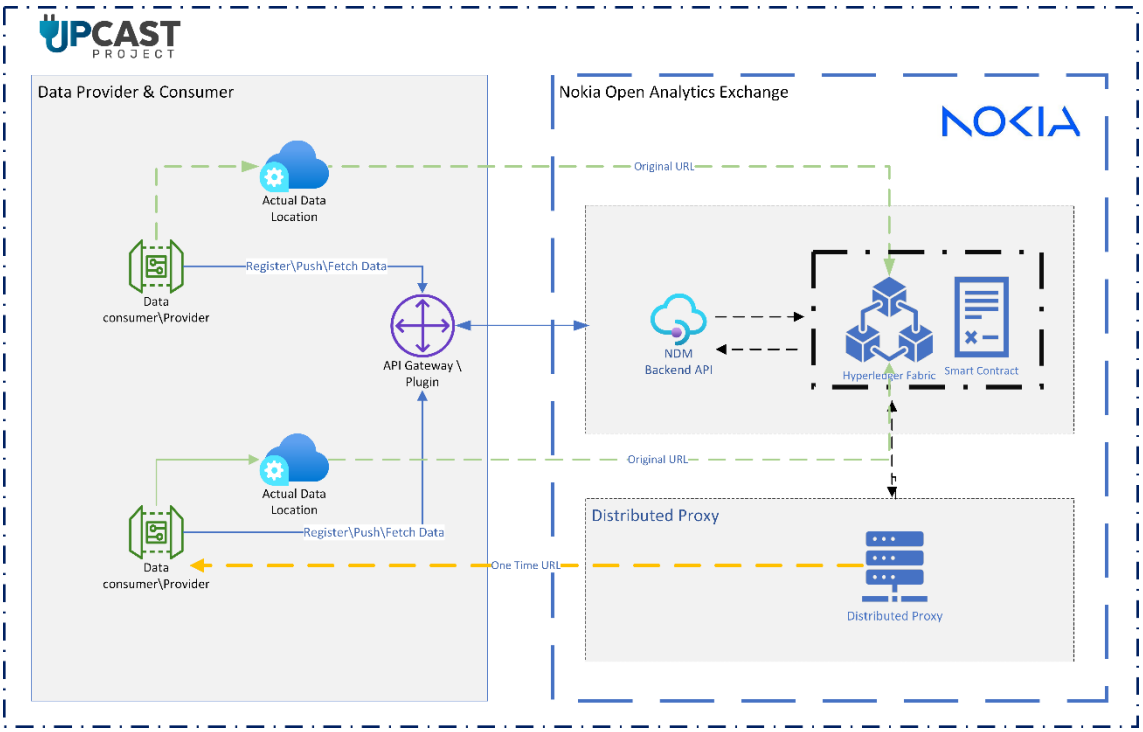


Figure 5. Initial solution for secure data exchange.

As shown in the design illustrated in Figure 5, data buyer\seller will communicate with each other using "Nokia Open Analytics Exchange". User will register itself in Nokia Open Analytics Exchange through API Gateway\Plugin.

- Data provider offering data that is been published in the data catalogue of Nokia Open Analytics Exchange.
- Data Consumer collecting and processing data from data provider

Smart Contract Design

- Nokia Open Analytics Exchange deliver several smart contracts for different function
- Token, access-control, contract revenue sharing, system fee, terms
- All smart contracts are Hyperledger fabric chain code written in Go lang
- Each of the smart contract an abstraction layer with exposed API's is available as microservice.

In this solution a blockchain-based controller manages identity and access control policies and acts as a tamper-proof log of access events and uses one-time URLs to secure data exchange only once. A one-time URL has the following characteristics: (1) Once a URL is used, it cannot be used again. (2) The URL will expire after a certain period. (3) The administrator can revoke a valid URL, and if the device tries to use this URL again, an error message will be seen.

DProxy

DProxy is one of the most important components of Nokia Open Analytics platform that provides scalability and distribution of access control. It is a software component that can be deployed like a microservice in the cloud or can be deployed on-premises.
This means that, rather than providing data to the platform, it is possible to "bring the platform" to the data by integrating DProxy on the edge device. DProxy is a simple software component that is used to mask the real data location and replace it with the

cryptographically secured URL. DProxy is a simple component developed in the Go programming language. It is written in such a way to be lightweight and simple to deploy so it can be in the cloud or on the edge. Also, code-level abstractions provide simple extensibility for different storage solutions in case that database is not acceptable due to hardware constraints. Currently, support for MongoDB and PostgreSQL is provided. DProxy service is rather simple with only two features: transforming the real URL to the secured URL and decrypting secured URL and getting the real asset URL. The secured URL consists of host location of the DProxy service followed by JWT. JWT is used as a convenient way to encode data to the string and provide signature to guarantee integrity of the URL.

2.10 Federated Machine Learning

2.11 Monitoring

The objective of the monitoring process is the collection of data from a subject system, subsystem, component, or infrastructure for further processing and analysis. Monitoring of a target system typically relates to the observation of the system's behavior. Following the collection of the monitoring data, processing and analysis is performed to determine the state of the monitored system, the compliance of its behavior to specifications and its overall health. Processing of collected monitoring data may be done either in real time or offline.

A monitoring service entails three items:

- The way by which monitoring data is collected, i.e., if a push or a pull model is used to extract data from the monitored system.
- The interface specifications for collecting the monitored data.
- The data models used for the monitored data.

Monitoring can take place for different reasons, for example:

- Monitoring the health of a system. Monitoring data are used to determine if the subject system, or some of its components, is or is not functioning. The objective in this case is determining the functioning or lack of functioning of a monitored system.
- Monitoring to determine if a system has reached a desirable state. In this case the monitoring data carry enough information to allow the processing part to determine if the desired state has been reached.
- Monitoring for performance analysis. In this case the performance of the target system is the objective of its monitoring.
- Monitoring access to a system and auditing: monitoring is done for determining the actors that access a system and the type of actions they perform on it.
- Monitoring for the compliance of the functioning of a system with specifications and requirements. Compliance with some specifications is generally a difficult type of problem.

Two models may be used for collecting monitoring data, the push and the pull model:

- The pull model works by sending a request from the monitoring service to the target system and waiting for a response. A deadline is typically specified to avoid the monitoring service blocking forever.

- The push model works by having the monitored target system send monitored data to the monitoring service when they become available. Typically, in the push model the monitoring service registers an interface with the target monitored system for monitoring data to be sent.

Monitoring of the execution of data processing workflows is a core functionality of UPCAST. Monitoring is a support functionality, which may not appear in users' requirements but is needed to support other functionalities of UPCAST, for example data processing workflow. The technology that will be used for monitoring is the Maggioli MIRA platform, which implements the digital twin paradigm and provides monitoring capabilities by its design. MIRA allows the modelling of assets and relationships between them and provides hooks for collecting monitoring data from them for further analysis. Data are collected through telemetries, which are the sources for monitoring data. The modelling part allows the association of a telemetry to an asset, which, when activated, sends monitoring data to be further analyzed by MIRA. Analysis can be as simple as generating graphs, but MIRA allows involved analytics services to be defined as separate services that run on top of it for doing more elaborate processing. MIRA will be used in UPCAST for implementing the underlying monitoring services and the data processing workflows as explained above in Section 2.3.

3 METHODOLOGY FOR REQUIREMENTS SPECIFICATION

The methodology for the requirements specification process is illustrated in Figure 6.

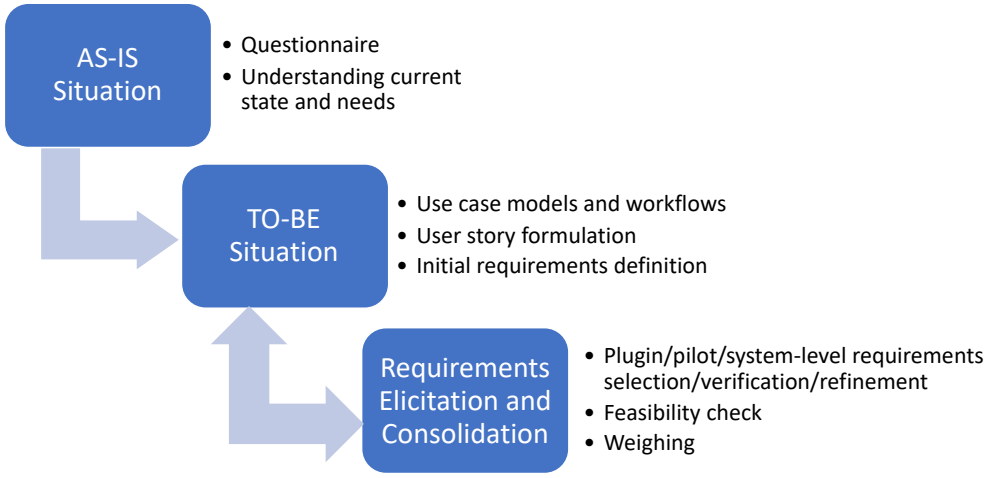


Figure 6. Methodology for the requirements specification process.

The different steps of the methodology are described in detail in the following:

AS-IS Situation

To get an understanding of the current state and needs of the pilots a questionnaire (see Annex 3) as an AS-IS interview guide was developed in the initial stage of the project. The questionnaire was responded by the pilot partners. Once the questionnaires were filled out, online sessions involving the pilot partners and technical & legal partners were conducted for each pilot to reach a common understanding of the current state as well as needs and how technical developments in the project can contribute to address the needs and objectives of the pilots.

TO-BE Situation

Based on information provided in the filled out AS-IS questionnaires and the online sessions, a draft version of the Context View of the ARCADE Framework (see the explanation in the following page) was developed. This included use case models defining relevant roles and needed functionality for each use case and business aspect workflow models defining how the UPCAST Plugins collectively contribute to address the needed functionality. Based on these models, user stories and functional/non-functional requirements were formulated by the pilot partners. The user stories and requirements were based on the templates described in Annex 1. Next, online TO-BE sessions including pilot partners and relevant technical & legal partners were conducted to verify the correctness of the use case and workflow models as well as to discuss and propose refinements of the formulated user stories and requirements. The resulting user stories and requirements for the pilots are presented in Chapter 4 while relevant legal requirements constituting a legal framework are described in Chapter 7.

Requirements Elicitation and Consolidation
There are four types of requirements elicited in this work:

- Pilot requirements: These are the user requirements describing needs for functionality (functional requirements) and system qualitative characteristics (non-functional requirements) as seen from the perspective of the pilot users.
- Plugin requirements: These are the more technical requirements defined by the partners that are aware of the technical capabilities and constraints of the plugins to be developed in the project.
- System-wide requirements: These are the requirements that define which functionality and system characteristics are needed for the UPCAST ecosystem, i.e., transversal requirements that relate to multiple plugins.
- Legal requirements: These are requirements that specify what is needed in order to be compliant with EU laws related to data management.

Once user stories and requirements had been defined by all pilots, the technical & legal partners reviewed them all to determine how they related to the different UPCAST Plugins. To support this process a requirements matrix mapping requirements defined by the pilots to the different UPCAST Plugins was developed. This process included an assessment of the technical feasibility of addressing the requirements as well as weighing them to determine an appropriate scope of the further developments. Based on this process, more detailed functional and non-functional requirements were defined for each UPCAST Plugin1. Once these more detailed requirements had been defined, the general requirements related to the system-wide dimension were elicited. These requirements, which focus on what is needed in terms of overall technical infrastructure to allow the UPCAST ecosystem of plugins to operate and interact properly to address the pilot needs, are described in Chapter 6.

Use of ARCADE Framework for overall systems architecture and requirements tracing
The ARCADE Framework[25] is an architecture description framework that is used to create a holistic system architecture covering the software engineering lifecycle from a requirements specification to a complete software component specification. The framework is based on IEEE-standardised specifications for system architecture specifications which suggest that a complete system architecture is best specified through a set of interconnected viewpoints. The viewpoints that are relevant for the requirements collection, MVP specification and technical architecture in UPCAST are illustrated in Figure 7.

[25] http://arcade-framework.org/

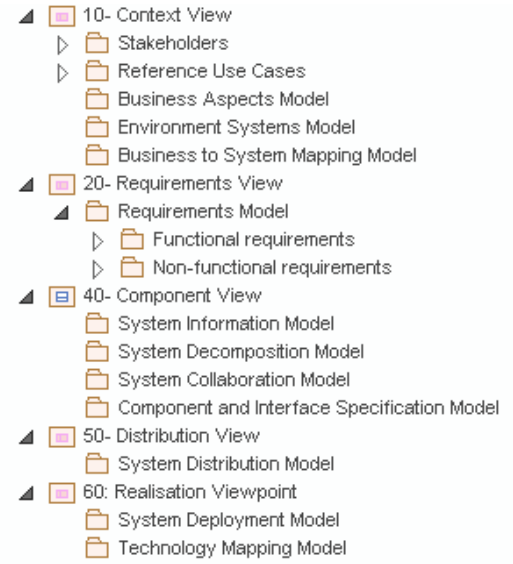


Figure 7. ARCADE Framework and views to be developed in UPCAST.

In D1.1, the viewpoints Context Viewpoint, Requirements Viewpoint and Component Viewpoint[26] are used for providing a contextual foundation for defining relevant requirements and represent a starting point for further MVP and architecture definitions. In the MVP specification (D1.2) and technical architecture specification (D1.2 and D1.3) architectural models instantiating and elaborating the Component Viewpoint, Distribution Viewpoint and Realisation Viewpoint will be specified. During the development of these models, it will be possible to trace back to the initial requirements as they will be embedded in the architecture modelling environment used in UPCAST, as illustrated in Figure 8.

[26] A view is what you see when you look at the architecture from a particular viewpoint. In other words, once the viewpoints are instantiated by models, they become views.

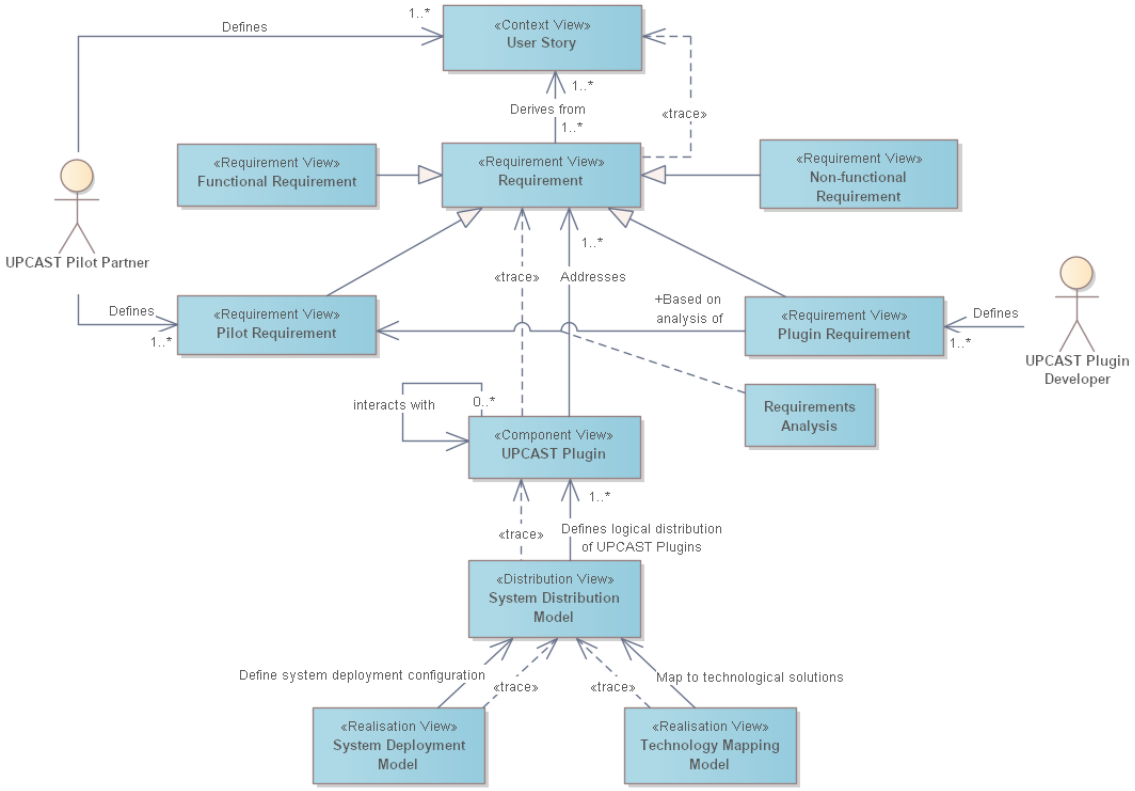


Figure 8. A meta-model illustrating how requirements are traced across architecture views of ARCADE.

4 PILOT REQUIREMENTS

This section gives an overall description of the pilots in UPCAST. This includes a description of the pilot case and goals, the main challenges, the users involved and datasets to be used and exchanged. Each pilot has formulated a set of user stories describing the core functionality required to address the goals and challenges in the pilot. A set of functional and non-functional pilot (user) requirements have been defined based on the user stories and general pilot needs. As some of these requirements are pilot-specific, these requirements have been filtered out and are presented as separate items at the end of each pilot requirements specification.

4.1 Biomedical and Genomic Data Sharing

This pilot concerns genomic and biomedical data sharing. NHRF is working on the field of cancer genomics and exploits experimental and computational approaches, aiming at the deeper understanding of molecular mechanisms implicated in cancer pathophysiology. NHRF utilizes bioinformatics tools for the analysis of in-house generated genomic data and for the exploration of molecular and clinical data from well-known cancer-associated repositories. In order to acquire biological material from cancer patients and clinical data, NHRF is collaborating with clinical partners.

4.1.1 Case description

This use case was formulated with four high-level use cases:

1) Establish collaboration

This use case focuses on creating a framework that enables different parties to establish contractual agreements with NHRF for collaboration, and to allow the parties to define specific clauses, obligations, and timelines, respecting all legal and ethical aspects, also including a review and approval process to ensure all parties involved are in agreement before finalizing the contract.

2) Share data

This use case focuses on developing a secure data sharing framework that allows NHRF researchers to share genomic data with specific individuals or groups. The framework will incorporate data privacy and access controls to ensure that sensitive information is protected.

3) Integrate and harmonize data

This use case focuses on developing tools to integrate and harmonize genomic data from various sources, such as public databases, research repositories, or private datasets. The system will also support data preprocessing and adequate transformations to ensure consistency across different datasets. It will handle differences in data formats, allowing researchers to combine and analyze data effectively for further genomics research.

4) Commercialize data

This use case focuses on developing a framework towards commercialization of NHRF proprietary or curated genomic datasets. The system will provide features for licensing and access control mechanisms to facilitate the commercial use of high-quality genomics data, ensuring that the data is appropriately utilized while protecting intellectual property rights.

4.1.2 Main technical challenges

Genomic and clinical data sharing raises challenges mainly due to the sensitive and heterogenous nature of such data. Ensuring compliance with data protection standards,

implementing encryption, and enforcing secure data transfer and storage are critical challenges. In addition, the analytical pipelines applied to raw genomic data frequently result in considerably diverse data outputs. This variation stems from the use of different reference sequences and annotations within these pipelines. Several initiatives aim to build specific standards and guidelines for data formatting, metadata annotation and analysis workflows. Ensuring compliance with these standards and enabling interoperability with external systems or databases can be technically challenging and requires adherence to community-accepted practices. The Global Alliance for Genomics and Health (GA4GH) (Rehm et al., 2021) is establishing policy frameworks and technical standards to support data sharing, including data models and specifications such as:

- Data Use Ontology (DUO)[27] to semantically tag our datasets with data use restrictions and requirements.
- Variation Representation (VRS), which provides a framework for scalable, federated computable exchange of genomic variation[28]
- Phenopackets: representation of genomic, phenotyping and clinical data[29] .

Ensuring that data is consistent with GA4GH standards for data representation and exchange formats will enable interoperability across different systems and platforms, a technical challenge that requires careful data management and quality control.

4.1.3 Main users/stakeholders

Main stakeholders include genomics researchers, data scientists, clinicians and healthcare practitioners but could also include diagnostic and pharmaceutical industry.

4.1.4 Datasets

Three types of datasets will be used concerning Transcriptomic data, Genomic data and clinical data.

Transcriptomic dataset: Transcript profiling ("Transcriptomics") of human tissue (cancerous or normal) is a widely used technique that obtains information on the abundance (number of sequencing reads) of multiple mRNA transcripts within a biological sample simultaneously. A RNAseq count file is a tabular file that contains the number of reads (sequencing data) that were aligned to specific genomic regions, typically gene regions, for each sample in a RNA-seq experiment. It is used to quantify the expression levels of genes in a sample and is a key input for many downstream analysis methods such as differential gene expression analysis, gene set enrichment analysis, and clustering analysis. The columns in a RNAseq count file typically include gene names, transcript IDs, and the count of reads that were aligned to the corresponding gene or transcript. The size of an RNAseq count file can vary widely depending on several factors, including the number of samples, the sequencing depth, the number of genes being quantified, and the specificity of the read alignment. Among the public genomic repositories that host RNAseq count files, the following databases hold a dominant position: cBioPortal[30], the Gene Expression Omnibus (GEO)[31] and the ArrayExpress[32].

[27] https://www.ga4gh.org/product/data-use-ontology-duo/
[28] https://www.ga4gh.org/product/variation-representation/
[29] https://www.ga4gh.org/product/phenopackets/
[30] https://www.cbioportal.org/
[31] https://www.ncbi.nlm.nih.gov/geo
[32] https://www.ebi.ac.uk/biostudies/arrayexpress

Genomic dataset: Genomic data describe alterations, somatic mutations, identified in DNA from cancer tissue in the form of MAF files. Mutation Annotation Format (MAF) is a tab-delimited text file with aggregated mutation information from VCF files and is generated on a project-level. A MAF file from a cancerous tissue typically contains information about genetic variants that are specific to the tumor sample, as compared to a matched normal sample or a reference genome. The dataset structure includes the following information as columns: Hugo_Symbol (gene identification), Chromosome (variant chromosome location), Start_Position (numeric position on the genomic reference sequence; start coordinate), End_Position (numeric genomic position; end coordinate), NCBI_Build (reference genome used for alignment), Variant_ID (dbSNP ID or a custom ID), Reference allele (nucleotide at the reference position in the genome), Alternate allele (nucleotide(s) that differ from the reference allele in the tumor sample), Transcript_ID (Ensembl ID of the transcript affected by the variant), Variant_Classification (translational effect of variant allele), AF (Allele Frequency of the alternate allele in the tumor sample, as a fraction or percentage of the total reads at the variant position), Annotation (information about the functional impact, predicted pathogenicity, and other features of the variant depending on the used databases). Among the public genomic repositories that host genomic data, the following databases hold a dominant position: CbioPortal[33] and the European Genome-phenome Archive (EGA)[34] (under controlled access).

Clinical dataset: Clinical data are in excel format as a table containing all clinical characteristics per patient (from the genomic dataset). Essential columns include: A unique identifier for every sample (Sample CODE) and a patient Identifier (1 patient can be related to two or more samples). 5-10 columns describing demographics for each patient, including date of birth, age at diagnosis, age group, sex, nationality, race, occupation, eye/hair/skin colour, family history, smoker (age start/quit), alcohol usage, other diagnoses (i.e. autoimmune diseases, cancer history) etc. 5-10 columns for clinical data including Diagnosis, Disease status, Degree/Staging, Metastasis, Therapy, Type of therapeutics received, Relapse, Symptoms, Biopsy, Hospital, date of sample excision, body location, histology, lymph node metastasis, mitosis etc. Specific vocabularies will be used like the Disease Ontology[35], the Oncotree[36] and NCI Thesaurus[37].

A more detailed description of the datasets relevant for this pilot can be found in Annex 3.

[33] https://www.cbioportal.org/
[34] https://ega-archive.org/access/data-access
[35] https://disease-ontology.org/
[36] http://oncotree.mskcc.org/#/home
[37] https://ncithesaurus.nci.nih.gov/ncitbrowser/

4.1.5 User stories and requirements

In this section the user stories (Table 1), functional requirements (Table 2) and non-functional requirements (Table 3) for the Biomedical and Genomic Data Sharing pilot are presented.

Table 1. User stories for the Biomedical and Genomic Data Sharing pilot.

ID	Use Case Name	As <a type of user>	I want to <goal/objective>	So that <benefit/result/some reason>
US_BM_1	Integrate and harmonise data – Discover data sources	NHRF	Manually search for data providers	I am able to locate all cooperating data providers (exploitation of available data is maximised)
US_BM_2	Integrate and harmonise data – Discover datasets	NHRF	Manually search for datasets	I am able to locate all suitable datasets (exploitation of available data is maximised)
US_BM_3	Establish collaboration – A priori negotiation	NHRF	Semi-automatically reach a collaboration agreement before any particular DPW is defined	This can be automatically enforced in all relevant DPWs
US_BM_4	Establish collaboration – Usage preferences	NHRF	Declare usage scope and context for biomaterial or raw data, and patient data, through a user-friendly GUI	It is ensured that informed contract negotiation takes place
US_BM_5	Establish collaboration – Usage constraints	Data provider	Declare usage constraints for offered data through a user-friendly GUI	Applicable legal and ethical requirements are imposed, intellectual property is safeguarded
US_BM_6	Establish collaboration – Consent assurance	NHRF	Obtain formal and traceable assurance of patient consent (where applicable)	Applicable legal and ethical requirements are guaranteed
US_BM_7	Establish collaboration – Consent enforcement	NHRF	Include patient consent terms into negotiation procedure (where applicable)	Data subject constraints are taken into account
US_BM_8	Establish collaboration – Negotiation outcome notification	NHRF, Data provider	Be visually informed of negotiation result	Data processing terms are transparent to all stakeholders

US_BM_9	Establish collaboration – Negotiation outcome approval/denial	NHRF, Data provider	Approve/deny negotiation result	Stakeholders retain control over negotiation procedure
US_BM_10	Integrate and harmonise data – Define DPW model	NHRF	Define a DPW through a user-friendly GUI (abstract level, without concrete implementation)	Intended data processes can formally and seamlessly be identified and defined
US_BM_11	Integrate and harmonise data – Specify data processing	NHRF	Specify in detail the data processed per DPW step, and the type/context of processing	Intended data processing can be accurately described
US_BM_12	Integrate and harmonise data – Negotiation outcome visualisation	NHRF	Be able to view the effect of negotiation results and other constraints on the DPW	Data processes to be executed are transparent in a user-friendly way
US_BM_13	Integrate and harmonise data – Execute DPW	NHRF	Execute DPWs, automatically enforcing negotiation results and any other privacy and usage control policies	DPWs are materialised, while established collaboration terms and any other provisions are properly enforced
US_BM_14	Share data – Data accessibility	Data provider	Make clinical or raw data or other datasets and data points accessible	NHRF can perform their DPW
US_BM_15	Integrate and harmonise data – Request dataset	NHRF	Describe unknown datasets (with or without data provider) and type of processing per DPW step, including usage scope and context	Desired type of data and intended data processing are accurately included in the DPWs
US_BM_16	Integrate and harmonise data – Query	NHRF	Describe targeted abstract queries to data providers' systems (with or without data provider) and type of processing per DPW step, including usage scope and context	Desired type of data and intended data processing are accurately included in the DPWs

US_BM_17	Integrate and harmonise data – Automatic resource discovery	NHRF	Dynamically discover data sources based on abstract specifications	The correct type of information is automatically included in the DPWs
US_BM_18	Establish collaboration – Ad hoc negotiation	NHRF	Semi-automatically reach a collaboration agreement in the context of a particular DPW	Ad-hoc collaborations are supported
US_BM_19	Integrate and harmonise data – Data resources description	Data provider	Formally describe datasets and data points	Datasets and datapoints can be discovered by NHRF
US_BM_20	Integrate and harmonise data – Data resources policies	Data provider	Link dataset/datapoint descriptions to usage policies	Collaboration can be established on this basis with NHRF according to “Establish collaboration”
US_BM_21	Integrate and harmonise data – Data resources advertisement	Data provider	Properly publish available datasets and datapoints	Datasets and datapoints can be discovered by NHRF
US_BM_22	Commercialise data – Data resources description	NHRF	Describe available datasets including pricing information and other usage policies	Selected datasets can be commercially exploited, based on collaborations established according to “Establish collaboration”
US_BM_23	Commercialise data – Data resources advertisement	NHRF	Properly publish commercially exploitable datasets	Commercially exploitable datasets are discovered by interested parties

Table 2. Functional requirements for the Biomedical and Genomic Data Sharing pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_BM_F_1	A graphical user interface shall be made available for NHRF (Data Consumer) to specify data resource usage preferences.	US_BM_4, US_BM_11, US_BM_15, US_BM_16	Validation with a Business Case	Must have
REQ_BM_F_2	A graphical user interface shall be made available for Clinicians and Data repositories (Data Providers) to specify data resource usage terms and conditions.	US_BM_5, US_BM_20, US_BM_22	Validation with a Business Case	Must have
REQ_BM_F_3	Dynamic negotiation on usage terms and pre-requisites must be provided, with GUI support	US_BM_3, US_BM_6, US_BM_7, US_BM_8, US_BM_9, US_BM_18	Validation with a Business Case	Must have
REQ_BM_F_4	A graphical user interface shall inform Data Consumers and Providers of negotiation results	US_BM_8, US_BM_12	Validation with a Business Case	Must have
REQ_BM_F_5	NHRF should be allowed to define DPWs (abstract level, no concrete implementation) – GUI support	US_BM_10, US_BM_11, US_BM_15, US_BM_16	Validation with a Business Case	Must have
REQ_BM_F_6	Users should be able to accurately define data to be exchanged, accessed and, or processed per DPW step (abstract level, no concrete implementation)		Validation with a Business Case	Must have
REQ_BM_F_7	The type and context of data processing within a DPW should be feasible to accurately specify	US_BM_11, US_BM_15, US_BM_16	Validation with a Business Case	Must have

	(abstract level, no concrete implementation)			
REQ_BM_F_8	Execution of distributed DPWs as modelled shall be supported, through transformation to concrete implementation	US_BM_13	Validation with a Business Case	Must have
REQ_BM_F_9	DPWs shall be executed respecting established collaboration agreements and any other privacy and usage control policies	US_BM_13	Validation with a Business Case	Must have
REQ_BM_F_10	Data Providers' datasets and endpoints shall be discoverable by Data Consumers	US_BM_1, US_BM_2, US_BM_15, US_BM_16, US_BM_17	Validation with a Business Case	Must have
REQ_BM_F_11	Data shared among components of the system must conform to the system data model	US_BM_1, US_BM_2, US_BM_3, US_BM_4, US_BM_5, US_BM_6, US_BM_7, US_BM_8, US_BM_9, US_BM_15, US_BM_16, US_BM_17	Validation with a Business Case	Must have
REQ_BM_F_12	There must be a procedure to integrate data from external sources to conform to the local data model	US_BM_1, US_BM_2, US_BM_3, US_BM_4, US_BM_5, US_BM_6,	Validation with a Business Case	Must have

		US_BM_7, US_BM_8, US_BM_9, US_BM_15, US_BM_16, US_BM_17		
REQ_BM_F_13	Message routing and dispatching and data processing shall take place under security, privacy and trust guarantees		Validation with a Business Case	Must have
REQ_BM_F_14	Users should be able to accurately define data sources to be accessed within a DPW (abstract level, no concrete implementation)	US_BM_11, US_BM_15, US_BM_16	Validation with a Business Case	Must have
REQ_BM_F_15	Data Providers' datasets and endpoints shall be formally described	US_BM_19, US_BM_20, US_BM_22	Validation with a Business Case	Must have
REQ_BM_F_16	Data Providers' datasets and endpoints shall be properly advertised	US_BM_21, US_BM_23	Validation with a Business Case	Must have
REQ_BM_F_17	Data Providers' datasets and endpoints shall be accessible by Data Consumers		Validation with a Business Case	Must have
REQ_BM_F_18	The system must support all required communication patterns, e.g., synchronous and asynchronous messaging, publish, subscribe, batch data delivery etc.	US_BM_13, US_BM_14	Validation with a Business Case	Must have
REQ_BM_F_19	The system must provide for the interaction among heterogeneous data providers and consumers		Validation with a Business Case	Must have

REQ_BM_F_20	It should be possible to become aware of the environmental impact implied by processing a discovered dataset prior to negotiating its acquisition	US_BM_2 US_BM_3 US_BM_15 US_BM_16 US_BM_17 US_BM_19 US_BM_21	Validation with a Business Case	Must have
REQ_BM_F_21	Means should be offered to estimate the environmental impact of processing a generated/offered dataset	US_BM_22, US_BM_23	Validation with a Business Case	Must have

Table 3. Non-functional requirements for the Biomedical and Genomic Data Sharing pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_BM_NF_1	Availability of all data sources involved in UPCAST modules operation must be guaranteed		Collaboration between team members	Must have
REQ_BM_NF_2	The data required as input to UPCAST modules must be provided by the respective data sources		Collaboration between team members	Must have
REQ_BM_NF_3	The data required to measure the proposed KPIs must be provided by the respective data sources		Validation with a Business Case	Must have
REQ_BM_NF_4	Data preserved must be consistent with the state of the system at each point in time and across all storage locations		Validation with a Business Case	Must have
REQ_BM_NF_5	Actions committed during all DPW lifecycle stages should be		Validation with a Business Case	Must have

	guaranteed to be preserved in case of failures			
REQ_BM_NF_6	Components performing some task efficiently (e.g., negotiation) should be reused as much as possible.		Collaboration between team members	Should have
REQ_BM_NF_7	The system shall, to the extent possible, provide means for integrating existing NHRF data processing infrastructure, procedures, standards, etc.		Collaboration between team members	Should have
REQ_BM_NF_8	A user-friendly GUI should be offered guiding the user through the main operations of the system	US_BM_1, US_BM_2, US_BM_14, US_BM_17, US_BM_19, US_BM_20, US_BM_21, US_BM_22, US_BM_23	Validation with a Business Case	Should have
REQ_BM_NF_9	Completeness and accuracy of data resources representations must be guaranteed throughout time		Validation with a Business Case	Must have

Table 4. Pilot-specific Requirements for the Biomedical and Genomic Data Sharing pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_BM_F_17	Data Providers' datasets and endpoints shall be accessible by Data Consumers		Validation with a Business Case	Must have

4.1.6 Key Performance Indicators

Key Performance Indicator	Measure
Number of data processing workflows implemented	≥ 3
Number of integrated data sources	≥ 5
Percentage of privacy flaws identified in data processing workflows:	≥ 90%
Time to assess a data processing workflow	< 1 min

4.2 Public Administration

This pilot concerns the 11 Municipalities that comprise the Metropolitan Area of Thessaloniki and their need to realise data driven environmental policy making. Working under their umbrella organisation, the Major Development Agency Thessaloniki (MDAT) and Open Knowledge Foundation Greece (OKF Greece) this pilot will use of the UPCAST plugins for integration and exchange of all data related to its environmental use case.

4.2.1 Case description

The availability of environmental data has the potential to change the ways in which cities are governed for sustainability and climate change mitigation and adaptation. More than ever, better environmental data is required to address urban challenges to the climate crisis. The availability of such data is critical to improve the monitoring and management of urban systems, as well as enabling robust assessments of policy and planning interventions.

A data-driven public sector is a process that transforms the design and delivery of public policies and services through the strategic management, sharing and use of data (OECD, 2019).

Thessaloniki's local authorities, facing the challenge of 100 Neutral Cities for 2030, have increased their level of understanding and acknowledgement of environmental data as vital resources for good policy making and urban prosperity. The efforts seem to have been directed towards bridging legacy systems, organisational, operational and infrastructure silos to enable the establishment of a data-driven public sector.

The goals of this pilot case are:

1) Characterise the different uses of environmental data, analyse focused interventions and informing operational decision-making, to monitoring progress against policy goals of the Metropolitan Area of Thessaloniki.

2) Define the local ecosystem of various actors for capturing, maintaining and using environmental data, such as the Resilient Thessaloniki municipal office and other intermediaries such as local NGOs and civil initiatives

3) Test the organizational/governance challenges in terms of managing and presenting environmental data delivered to other actors/data consumers such as researchers, practitioners, citizens, entrepreneurs etc.

4.2.2 Main technical challenges

Data required to calculate environmental indicators for the whole metropolitan area is collected in a distributed way by each city, requiring cleaning, integration and aggregation to be usable. Thus, a major challenge is the efficient data integration, in order to automate the methodology of gathering data from the data providers with responsibility, reliability and to be updated with a specific frequency, securely maintained

and shared with an efficient methodology to the data/tools' users.

Before getting to the data integration step, as the processes are still manual, they require ad hoc negotiation and contracts with the data providers. A critical challenge is to automate the contracting processes, without time-consuming negotiation, based on the respective UPCAST plugins.

Moreover, some of the required data may be private or confidential and the cities would like to ensure that any processing needed for computing the indicators respects any privacy conditions. Thus, another challenge is that these stakeholders can ask or define their privacy conditions and requirements regarding their data, in a clear and automated way.

Finally, as the pilot case comprises various heterogeneous datasets, a challenging goal would be to include relevant standards and vocabularies for encoding and representing the respective data.

4.2.3 Main users/stakeholders

An essential element for data-driven environmental policy making is the cross-boundary information integration in between authorised agencies, between research institutions and local authorities or/and not-for-profit organizations and private firms and the public sector. One of the aims of this pilot is to create a local ecosystem of environmental data stakeholders providing a multistakeholder data governance and policy making approach.

There are two generic types of stakeholders related to this pilot:

1) Data providers: authorized organizations, research institutions and/or civil initiatives that have, measure and provide the data

2) Data / tools users: municipalities, authorized organizations, researchers that can exploit the data.

4.2.4 Datasets

Collecting long-term environmental data is crucial when it comes to assessing changes in environmental policy making. Assembling and managing high quality environmental data sets is the main challenge for this pilot case. Thus, thoughtful analysis is needed to interpret available environmental data.

The pilot will use environmental data and datasets for Thessaloniki Metropolitan area that comes from various sources, such as:

1. Extracting data for Thessaloniki metropolitan area from global and EU open-source resources such as Eurostat. Especially for Eurostat the pilot uses available data referred to the EU Metropolitan Regions, i.e., datasets under code MET-. These regions are defined as urban agglomerations NUTS level III. Source data ('building blocks') for the metropolitan data are existing NUTS III indicators in the Eurostat production database for the statistical themes of Area, Demography, Population projections, and Transport. There are no specific data in the theme "Environment". In addition, some pilot datasets are extracting from Eurostat database "cities and greater areas" (Urban Audit) focusing on available data for Functional Urban Area (dataset with the code: urb_luz)
2. Extracting data for Thessaloniki metropolitan area from national resources either open or on demand. The main source for these datasets is the Hellenic Statistical Authority, but also the Ministry of the Environment and Energy and occasionally other national institutions.

3. A third source for data retrieving is the local research institutions where the data is available by case and after negotiations. These are real time data or foreseen environmental data produced by the research laboratories and institutions.

An illustrative summary of the data sources and needs to be solved for the Public Administration pilot is provided in Figure 9.

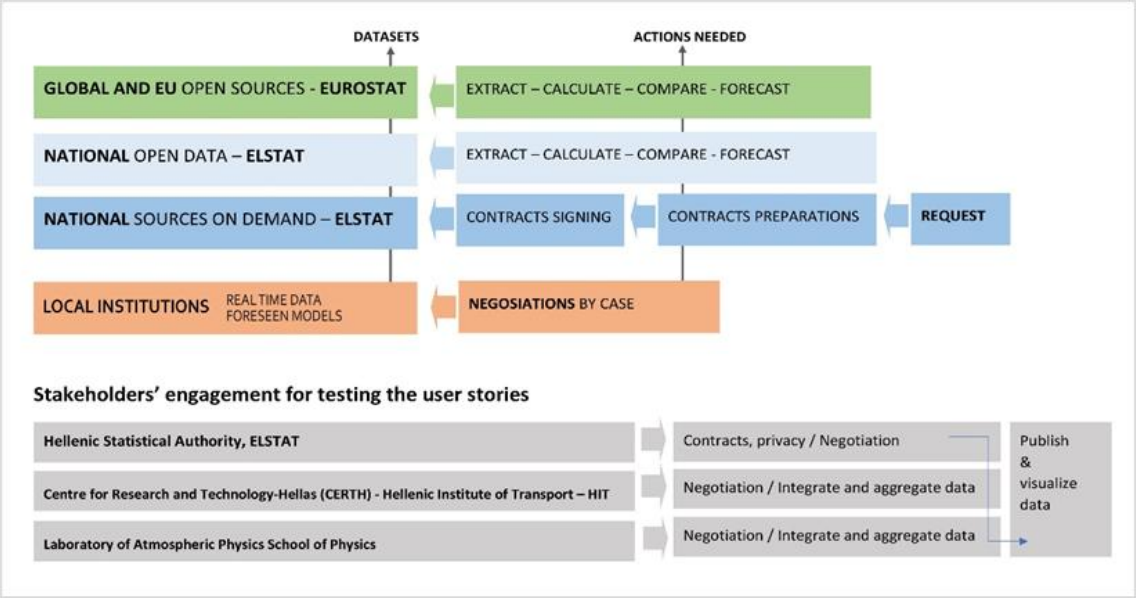


Figure 9. Datasets sources and needs to be solved for pilot case "public administration"

For the business case, the following datasets has been chosen:

- Dataset 01: General Demographics statistics for the metropolitan area such as population and gender distribution, household composition, occupation distribution etc.
- Dataset 02: Urban statistics including referred to the metropolitan area and the functional urban area of Thessaloniki, regarding mainly the land use occupation.
- Dataset 03: Living conditions (related to environmental issues) such as dwellings construction period, heating and insulation availability, heating and hot water source, energy source etc.
- Dataset 04: Households and car and parking availability
- Dataset 05: Transport statistics, including general statistics such as number of cars registered, bicycle lane length etc. As transportation habits have a major impact on environmental issues in Greek cities, pilot give specific focus on data related to transport
- Dataset 06: Urban traffic conditions
- Dataset 07: Road freight transport in Thessaloniki metropolitan area
- Dataset 08: Environmental statistics for Thessaloniki metropolitan area,
- Dataset 09: Air pollution measurements based on measurements located in specific city locations

A more detailed description of the datasets relevant for this pilot can be found in Annex 3.

4.2.5 User Stories and Requirements

ID	Use Case Name	As <a type of user>	I want to <goal/objective>	So that <benefit/result/some reason>
US_PA_1	Determine data providers	Public Administration (MDAT-OKF)	Search for other data providers and browse their datasets	I can identify key stakeholders and contact them in order to acquire/access datasets that are of interest (required data)
US_PA_2	Establish contract for data exchange	Public Administration (MDAT-OKF)	Negotiate with data providers in an automated way	I can ask and acquire/access to datasets of interest with the same responsibility and reliability but more time-efficiently compared to manually negotiating and signing contracts
US_PA_3	Establish contract for data exchange	Actor holding data	Define privacy and access control constraints	The datasets that I will provide will be used under specific terms
US_PA_4	Establish contract for data exchange	Public Administration (MDAT-OKF)	Negotiate for updated data (with data providers)	They provide regular updates to their data
US_PA_5	Publish dataset	Public Administration (MDAT-OKF)	Use domain-specific vocabularies and ontologies	Before publishing the datasets, I semantically represented them and used relevant metadata to describe them. Thus, data users can find them and integrate them more conveniently.
US_PA_6	Publish dataset	Public Administration (MDAT-OKF)	Use a (semantic) repository for publishing the datasets	I publish all the datasets within the same repository (e.g., CKAN, DCKAN) to be used by data consumers.
US_PA_7	Integrate and aggregate data	Public Administration (MDAT-OKF)	Get access to datasets from external actors holding data	I can integrate them into my data processing workflow, aggregate them and perform data analysis and calculations to output indicators of interest
US_PA_8	Define data processing workflow	Public Administration (MDAT-OKF)	Define a data processing workflow	I can determine the data providers who (might) have the required data and specify the operations that will be performed to them abiding on the negotiated terms, targeting specific data

				users (for whom the data will be useful)
US_PA_9	Define data processing workflow	Actor holding data	I can give approval to requests of other parties	they can use my data in data processing workflows

Table 5. Functional requirements for the Public Administration pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_PA_F_1	Data Providers' datasets and endpoints shall be formally described using domain-specific vocabularies and ontologies	US_PA_5 US_PA_6	Validation with a Business Case	Must have
REQ_PA_F_2	Data Providers' datasets and endpoints shall be discoverable by Data Consumers	US_PA_5 US_PA_6 US_PA_1	Validation with a Business Case	Must have
REQ_PA_F_3	Data Providers' datasets and endpoints shall be accessible by Data Consumers	US_PA_5 US_PA_6 US_PA_1	Validation with a Business Case	Must have
REQ_PA_F_4	There must be a procedure to integrate data from external sources to conform to the local data model	US_PA_7	Validation with a Business Case	Must have
REQ_PA_F_5	A GUI shall be made available for Data providers to specify data resource usage preferences.	US_PA_3	Validation with a Business Case	Must have
REQ_PA_F_6	A GUI shall be made available for data providers to specify data resource usage terms and conditions.	US_PA_3	Validation with a Business Case	Must have
REQ_PA_F_7	A GUI shall be made available for data providers and data consumers to negotiate the terms and conditions for the use of data.	US_PA_2 US_PA_4	Validation with a Business Case	Must have
REQ_PA_F_8	A GUI shall be made available for data providers and data consumers to view the negotiation results.	US_PA_2 US_PA_4	Validation with a Business Case	Must have
REQ_PA_F_9	A GUI shall be made available for the MDAT-OKF to model Data Processing Workflows.	US_PA_8	Validation with a Business Case	Must have
REQ_PA_F_10	MDAT-OKF should be able to accurately define data sources to be accessed within a DPW	US_PA_8	Validation with a Business Case	Must have

REQ_PA_F_11	MDAT-OKF should be able to accurately define data to be exchanged, accessed and/or processed per DPW step	US_PA_8	Validation with a Business Case	Must have
REQ_PA_F_12	MDAT-OKF should be able to accurately specify the type and context of data processing within a DPW	US_PA_8	Validation with a Business Case	Must have
REQ_PA_F_13	DPWs shall be executed respecting established collaboration agreements and any other privacy and usage control policies. In the case there are no prior agreements the Data Providers should have the ability to give consent or reject any request from the data consumer who tries to execute a DPW using their data.	US_PA_8 US_PA_9	Validation with a Business Case	Must have
REQ_PA_F_14	A GUI shall be made available for MDAT-OKF (Data Consumer) as Public Administrator in order to search for datasets of interest and identify the key stakeholders/data providers. Search should be conducted both using keywords and by available data providers.	US_PA_1	Validation with a Business Case	Must have
REQ_PA_F_15	The system must provide for the interaction among heterogeneous data providers and consumers	US_PA_7	Validation with a Business Case	Must have
REQ_PA_F_16	The energy profile of a dataset should be provided		Validation with a Business Case	Must have

Table 6. Non-functional requirements for the Public Administration pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_PA_NF_1	All system components have to support the Greek language throughout all their functionalities		Validation with a Business Case	Must have
REQ_PA_NF_2	A user-friendly GUI should be provided in order to guide users through the main operations of the system		Validation with a Business Case	Should have

Table 7. Pilot-specific Requirements for the Public Administration pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_PA_F_3	Data Providers' datasets and endpoints shall be accessible by Data Consumers	US_PA_5 US_PA_6 US_PA_1	Validation with a Business Case	Must have

4.2.6 Key Performance Indicators

Key Performance Indicator	Measure
Datasets shared with other public administration	≥ 8
Datasets anonymised and following legal constraints ready for sharing with external parties	≥ 33
Data Processing Workflows implementing decision making indicators	≥ 3 workflows
Efficiency increases in decision-making process	≥ 20% faster, measured in days to reach a data-sharing agreement and take a decision.

4.3 Health and Fitness

Millions of people are sharing data in various fitness apps with the help of devices like wearables and IOT-enabled fitness equipment, creating very large datasets and streams. Personal fitness/health data, collected during various physical activities is extremely valuable for both the data producer (trainee), service providers (fitness, healthcare, wellbeing) and product vendors (e.g., vendors of the fitness equipment, nutrition supplements). However, wearables and fitness equipment are often used in "isolation", meaning they are tailored to scenarios that benefit a single trainee. In this pilot, the UPCAST plugins will be used to valuate, share and trade data streams related to health and fitness data.

4.3.1 Case description

Smart4Fit[38] (TRL9) is a system for real-time monitoring for fitness based on personal wearables integrated in a bigger IoT environment (fitness club with plenty of connected fitness device) and is used in collaborative scenarios, like group training and collaborative gamification.

On the other hand, personal fitness/health data, collected during various physical activities has an good value not only for the data producer (trainee), but also for many service providers (fitness, healthcare, wellbeing) and product vendors (e.g., vendors of the fitness equipment, different supplements).

The use case is resolving challenges for an efficient and secure monetarization of such data. It explains the need for sharing the data and monetarize its value properly.

4.3.2 Main technical challenges

Data Collection challenges

The most important feature is the data creation process, to ensure that the data is collected in high quality and without loss. This is challenging since the data is created on the extreme edge (wearables) and transferred using wireless methods under challenging conditions (often: weak/instable connection). Additionally, devices/wearables are small electronic devices attached to the body of trainees – intensively moving in the space and risking the interruptions in the data generation / transfer.

[38] https://smart4fit.nissatech.com/

Challenges for sharing the data:
One of the main issues for data sharing is that the data providers (trainees) are not aware of the value of data for "others" (even the value for themselves is not completely clear), meaning that they cannot understand how the data is "valuable" for specific service/vendor providers, interested in that data. By knowing this, data provider will be not only aware of the "price" of data, but more importantly, she/he will understand how that price is formed (that can influence the data production/creation process). The first one is passive and leads to data sovereignty – data provider controls the usage of own data (already produced). The second is active and empowers the data provider to produce data of a particular quality (data that has a higher price will be created) – the data creation process will be controlled (different from controlling the usage of produced data)

4.3.3 Main users/stakeholders

- Data creators – those who are performing a physical activity, the data will be collected from / trainees
- Data owners – those who are transferring data in the digital form /
- Data traders – those who are selling the data
- Data consumers – those who are buying the data

4.3.4 Datasets

HR: Heart rate data refers to the measurements or recordings of a person's heart rate over a period. The heart rate is a measure of the number of times the heart beats per minute (bpm) and is commonly used as an indicator of a person's cardiovascular health and physical exertion. It is measured in beats per minute (BPM). Smart4Fit collects data at a frequency of 0.5 Hz, which means a new heart rate measurement is recorded every 2 seconds. Data is obtained from Bluetooth sensors during training sessions, allowing us to track and monitor the trainees' heart rate throughout their workout. Data is further analyzed in other **smart** analytical services to gain insights into the trainees' physiological response, intensity of the exercise, recovery patterns and overall cardiovascular fitness.

ACC: Acceleration data from an accelerometer refers to the measurements or recordings of the acceleration experienced by an object or body in three-dimensional space. Unlike a gyroscope that measures rotational movement, an accelerometer specifically detects linear acceleration, including both static and dynamic acceleration. The accelerometer provides acceleration data in three axes: x, y, and z. Each axis represents a different direction or dimension of linear movement. The data collected from these axes allows tracking of changes in velocity or speed of the object in those directions. The sensor can be configured to collect data at various frequencies - 5, 52, 208 and 416 Hz. Higher frequencies provide more detailed data and capture rapid changes in acceleration, while lower frequencies may be suitable for capturing slower movements or conserving battery life. By collecting acceleration data from the accelerometer, one can analyze a trainee's movement patterns, assess the intensity of physical activities, detect impacts or sudden changes in velocity, and monitor body dynamics during training sessions. This data can be used to evaluate exercise techniques, quantify physical exertion, identify areas for improvement, and enhance the overall training outcome for the trainees.

A more detailed description of the datasets relevant for this pilot can be found in Annex 3.

4.3.5 User Stories and Requirements

ID	Use Case Name	As a <type of user>	I want to <goal/objective>	So that <benefit/result/some reason>
US_HF_0	Data monetization clarification	Trainee	Clarify the data monetization process	I can decide if I want to do the monetization
US_HF_1	Individual contribution reward	Trainee	Get a revenue for my data contribution	I can decide if give my data in exchange of revenue
US_HF_2	Data product pricing	Data owner	Price of data product that I intend to sell.	I can decide if I want to do the monetization (and how much)
US_HF_3	Valuating contributions from trainees/gyms	Data owner	Valuate data contribution from trainees to a data product	I can decide if I want to do the monetization (and how much) And be motivated to generate as much as possible of such data
US_HF_4	Data price Negotiation	Trainee	Know how I can determine the value of my data	I can decide if I want to do the monetization (and how much) And be motivated to generate as much as possible of such data
US_HF_5	Data usage constraints	Trainee	make/use some constraint on the data usage	I can decide if I want to do the monetization (and how much)
US_HF_6	Data bundles	Trainer	make some bundles of data	I can make a better offering
US_HF_7	Request for Specific data	Data trader	Define a request (quantity, price) for some specific data	I can trade/sell such data

Table 8. Functional requirements for the Health and Fitness pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_HF_F_1	A trainee should be able to ask for the price of particular data		Validation with a Business Case	Must have
REQ_HF_F_2	A trainee should be able to influence (somehow) the price of own data	US_HF_3	Validation with a Business Case	Must have
REQ_HF_F_3	A trainee should be able to define some constraint on the data usage	US_HF_5	Validation with a Business Case	Must have
REQ_HF_F_4	A trainer should be able to create some bundles of data (to be sold as a bundle)	US_HF_6	Validation with a Business Case	Must have
REQ_HF_F_5	A data trader should be able to create a request (quantity, price) for some specific data	US_HF_7	Validation with a Business Case	Must have
REQ_HF_F_6	A data trader should be able to create a semantic query for data	US_HF_7	Validation with a Business Case	Must have
REQ_HF_F_7	A data consumer should be able to create a request (quantity, price) for some specific data	US_HF_7	Validation with a Business Case	Must have

Table 9. Non-functional requirements for the Health and Fitness pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_HF_NF_1	The Smart4Fit platform should with the enhancements made in UPCAST be scalable for 10000 users. 100 clubs with 100 members is one of the (business) KPIs.		Validation with a Business Case. Some objective metrics (like cumulative delay per club per upload) are relevant, as well as subjective metrics such as no complaints from customers (regarding the performances).	Must have

REQ_HF_NF_2	A user-friendly GUI should be provided in order to guide users through the main operations of the system		Validation with a Business Case. Relevant objective metrics are: learning to use the system in N hours), but more important is customer satisfaction with the learning process.	Should have

Table 10. Pilot-specific requirements for the Health and Fitness pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_HF_F_7	A data consumer should be able to create a request (quantity, price) for some specific data	US_HF_7	Validation with a Business Case	Must have
REQ_HF_NF_1	The system should be scalable for 10000 users		Validation with a Business Case. Some objective metrics (like cumulative delay per club per upload) are relevant, as well as subjective metrics such as no complaints from customers (regarding the performances).	Must have

4.3.6 Key Performance Indicators

Key Performance Indicator	Measure
Increase the amount of data shared with external stakeholders	≥ 50%
Increasing the number of users who understand what is the value of data	≥ 50%
Increase the average value/price of data	≥ 30%
Automatize the procedures for sharing data with an external stakeholder incl. ethical and legal issues	≥ 85%

4.4 Digital Marketing Data and Resources 1 (JOT)

Today, data on the performance of digital marketing campaigns is only used to determine the most adequate optimization actions to increase the engagement and ROI of the active campaigns. However, there is high value in the data that is not exploited at all. By analyzing this type of data it is possible to determine the real user interests in a wide variety of business verticals (classified based on the Google categories taxonomy) and locations. For that reason, this pilot is developing a new data-as-service business model where data consumers can decide which data is needed and how the data and insights should be delivered.

4.4.1 Case description

Figure 107 shows the main components of the pilot: (i) the service request represents the interface where the data consumer can check some examples of the data sets, sign and log in, and define the service request based on a pre-defined set of filters and features; (ii) the service signature is the component responsible for calculating the price of the service requested, define and sign the contract and (iii) the service delivery, where the query is automatically generated based on the data consumer needs, and the different services are implemented (data sharing, and reporting).

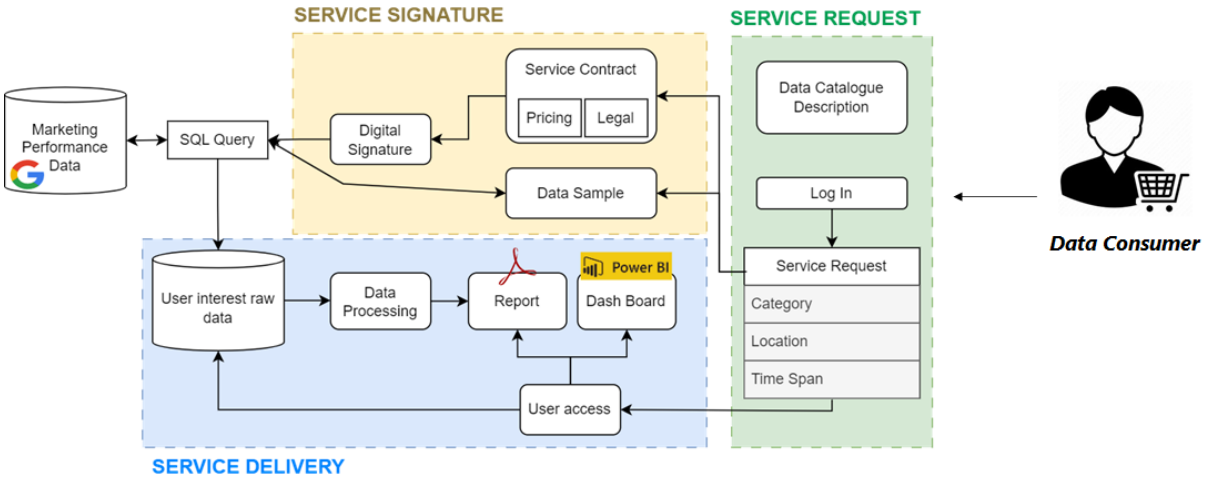


Figure 10. Architecture of Digital Marketing Data monetization business case

In summary, this marketing related data as a service offers the possibility of getting access to new information and insights depending on the data consumer needs, who will be able to know more about real topics and interests in their targeted markets.

4.4.2 Main technical challenges

The development and deployment of this pilot has associated critical challenges to be delivered in productive conditions:

1. Selection of the parameters to define the requested data set: So the data set request can be specific enough to address the data consumer needs but containing a significant amount of data to be able to calculate the insights of the market and domain.
2. Evaluation of the requested data and price calculation: Depending on both the service requested (combination of data set and reporting service) and the pricing competition in the market.
3. Transformation of the service request in a SQL query: To reduce the human coding effort (and errors) and promote the scalability of the service
4. Definition of report templates and KPIs: Enabling the automation of the document generation with the main insights for each data service request as well as the definition of relevant KPIs to support the initial analysis of the data.

There are also additional challenges related to the secure and private access to the service and allocation of dedicated resources for data processing that are also relevant but less specific to this pilot and will be addressed in a more general and common approach for all pilots.

4.4.3 Main Users/Stakeholders

For this business case, four different stakeholders have been defined combining internal and external users.

- Service Provider (SP): Person with analytical profile responsible of the definition and development of the services requested by the data consumer. The general service will be formed by a combination of data set, report with the main insights and interactive dashboard to visualize the main indicators.
- Data Provider (DP): Person responsible for the generation of the data set requested by the data consumer. The DP will define the degree of freedom of the DC when defining the data set request and will translate them into a SQL query to access the data.
- Resource Provider (RP): Technical person responsible of the orchestration of the entire service value chain and deploying the required resources to deliver the data monetization service. This will include both cloud storage and computing resources.
- Data Consumer (DC): Final user that will request the data sets offered by JOT through a web-based interface. Main variables the DC may define are the category, location, time spam as well as the type of service requested.

4.4.4 Datasets

According to the case description, the data sets exploited in this business case are formed by the statistics of the digital marketing campaigns that are related to the end user interests in the search engines like Google.com. As seen in Figure 11 the raw data sets are directly downloaded from Google Ads (the platform where all the marketing campaigns are managed). Different data is collected depending on the aggregation level from campaign to group and keyword level, being more granular as the level in the marketing structure is lower. In most of the cases, this raw data needs to be cleaned and prepared to be processed and used. For that reason, a set of specific ETLs are developed, including the final linking and joins processes to correlate the information and different levels.

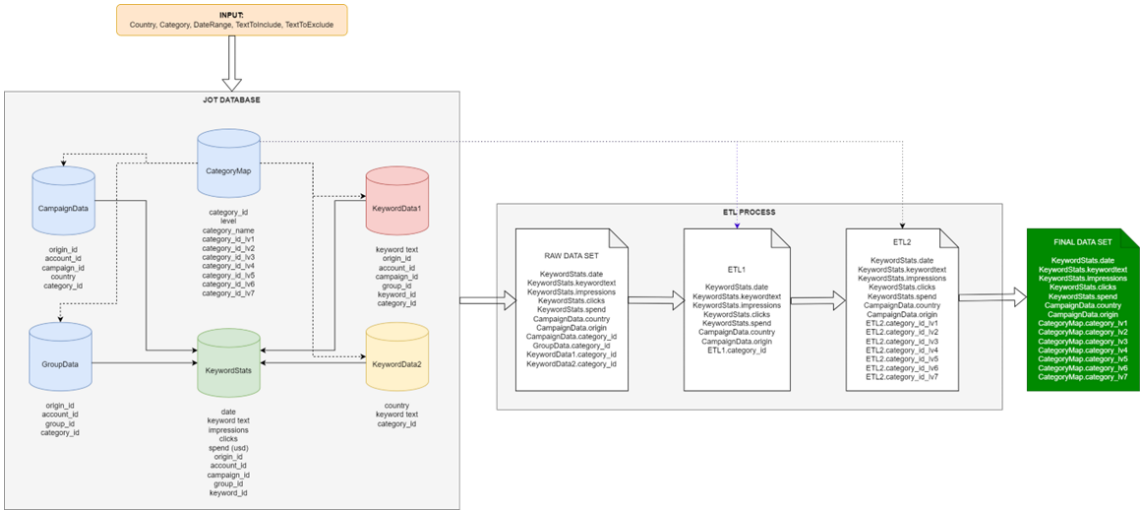


Figure 11. Data set generation for the digital marketing data case

Depending on the service request, different data sets can be generated as a function of the category (topic or business vertical), location (country), time span (from days to years) and aggregation level (day, week, month, campaign, group, etc.).

A more detailed description of the datasets relevant for this pilot can be found in Annex 3.

4.4.5 User Stories and Requirements

Based on the user profiles and the data set offered for data monetization, the following user stories have been defined to cover all the service development and delivery. They have been divided in two major groups:

- Green: Prerequisite functionality that needs to be developed to support the run-time functionality.
- Blue: Run-time functionality

Table 11. User stories for the Digital Marketing (JOT) pilot.

ID	Use case name	As a <type of user>	I want to <goal/objective>	So that <benefit/result/some reason>
US_DM_JOT_9	Generate reporting services	Service provider	Offer interactive and updatable dashboard as an additional service for data consumer	They can access to updated data and insights depending on their initial service definition
US_DM_JOT_7	Define user needs	Service provider	Define a service definition interface combining data set and additional service features	I can know the type of service requested
US_DM_JOT_8	Develop analytical services	Service provider	Generate a general list of KPIs and reports	I can automate the generation of the reports and dashboards depending on the data set properties (categories, time span and locations)

US_DM_JOT_10	Data consumer access to the required services	Service provider	Give access and track the access to the data and services requested by the data consumer	Each of the data consumers can have access to their specific repository and dashboards
US_DM_JOT_3-1	Define framework for automated query generation	Service provider	Have functionality that enables automated generation of the SQL based query to the data base	Transform data needs from the Data consumer (11) to an SQL query.
US_DM_JOT_12	Define service requirements	Data consumer (external)	Define the type of service requested and expected price	I can adjust the service to the budget and info expectations
US_DM_JOT_11	Define data needs for consuming	Data consumer (external)	Provide as much information as possible about data set needs and expected insights	I can get as much value as possible from the data
US_DM_JOT_3	Define the query to the SQL server	Data Provider	Automate the generation of the SQL based query to the data base	All the data set request can be managed with no/minor manual actions
US_DM_JOT_5	Allocate processing resources	Resource provider	Know the properties of the requested data set	I can allocate processing resources to deliver the additional services
US_DM_JOT_6	Allocate storage resources	Resource provider	Know the properties of the requested data set	I can allocate storage resources to deliver final data set
US_DM_JOT_4	Generate price offer depending on service (data) definition	Data Provider	Generate a customized price for the data set	The data consumer can get a personalized offer depending on their specific needs
US_DM_JOT_1	Generate data samples	Data Provider	generate a reduced sample of the data set	The data consumer can confirm the attributes (data model) and formats
US_DM_JOT_2	Generate final data sets	Data Provider	generate the complete data set as requested by the data consumer	The data consumer can get access to it and the service provider can generate the additional services

Table 12. Functional requirements for the Digital Marketing (JOT) pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_DM_JOT_F_1	There should be a UI that allows the User to log in with its unique access	US_DM_JOT_10	Synthetic datasets, Validation with a Business Case	Must have
REQ_DM_JOT_F_2	There should be a UI that allows de Data Consumer to establish the parameters for the service request.	US_DM_JOT_10 , US_DM_JOT_11 , US_DM_JOT_7 , US_DM_JOT_12	Validation with a Business Case	Must have
REQ_DM_JOT_F_3	Based on the parameters established by the Data Consumer, there should be a service that generates automatically a dataset (for example a csv file), a dashboard (for example PowerBI) or a custom report with pre-defined KPIs (in pdf format for example).	US_DM_JOT_8 , US_DM_JOT_9 , US_DM_JOT_1 , US_DM_JOT_2	Synthetic datasets, Validation with a Business Case	Must have
REQ_DM_JOT_F_4	Based on the parameters given by the Data Consumer, an automatic SQL query should be generated in order to retrieve the data	US_DM_JOT_5 , US_DM_JOT_6 , US_DM_JOT_3	Validation with a Business Case	Must have
REQ_DM_JOT_F_5	Based on the parameters given by the Data Consumer, a price for the data set should be calculated and a UI should be enabled so the Data Consumer can agree with it	US_DM_JOT_4	Collaboration between team members	Must have

Table 13. Non-functional requirements for the Digital Marketing (JOT) pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_DM_JOT_NF_1	The service should be adapted to the company and/or platform where is integrated	US_DM_JOT_7, US_DM_JOT_11		

REQ_DM_JOT_NF_2	A user-friendly GUI should be offered guiding the user through the main operations of the system	US_DM_JOT_1 , US_DM_JOT_9 , US_DM_JOT_7, US_DM_JOT_11		
REQ_DM_JOT_NF_3	Each user must have access only to the data, reports and dashboards requested by them.	US_DM_JOT_4, US_DM_JOT_10, US_DM_JOT_2		
REQ_DM_JOT_NF_4	User may have access to data sets, reports and dashboards during the entire period of contract validity. Also, the info should be displayed in laptop, tablet and smart phones	US_DM_JOT_2, US_DM_JOT_6, US_DM_JOT_5 , US_DM_JOT_9 , US_DM_JOT_10		
REQ_DM_JOT_NF_5	The service must ensure that the user experience is not impacted by the lack of processing resources for data and report generation.	US_DM_JOT_5, US_DM_JOT_9 , US_DM_JOT_10		

Table 14. Pilot-specific requirements for the Digital Marketing (JOT) pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_DM_JOT_F_1	There should be a UI that allows the User to log in with its unique access	US_DM_JOT_10	Synthetic datasets, Validation with a Business Case	Must have
REQ_DM_JOT_F_3	Based on the parameters established by the Data Consumer, there should be a service that generates automatically a dataset (for example a csv file), a dashboard (for example PowerBI) or a custom report with pre-defined KPIs (in pdf format for example).	US_DM_JOT_8 , US_DM_JOT_9 , US_DM_JOT_1 , US_DM_JOT_2	Synthetic datasets, Validation with a Business Case	Must have

REQ_DM_JOT_F_4	Based on the parameters given by the Data Consumer, an automatic SQL query should be generated in order to retrieve the data	US_DM_JOT_5 , US_DM_JOT_6 , US_DM_JOT_3	Validation with a Business Case	Must have
REQ_DM_JOT_NF_1	The service should be adapted to the company and/or platform where is integrated			
REQ_DM_JOT_NF_3	Each user must have access only to the data, reports and dashboards requested by them.	US_DM_JOT_4, US_DM_JOT_10, US_DM_JOT_2		
REQ_DM_JOT_NF_4	User may have access to data sets, reports and dashboards during the entire period of contract validity. Also, the info should be displayed in laptop, tablet and smart phones	US_DM_JOT_2, US_DM_JOT_6, US_DM_JOT_5 , US_DM_JOT_9 , US_DM_JOT_10		
REQ_DM_JOT_NF_5	The service must ensure that the user experience is not impacted by the lack of processing resources for data and report generation.	US_DM_JOT_5, US_DM_JOT_9 , US_DM_JOT_10		

4.4.6 Key Performance Indicators

Key Performance Indicator	Measure
Deployment of 5 data set for different business verticals as examples	Generation of 5 examples of data sets to be included in the UI for data consumer consultation
Generation of 3 type of reports	Definition and implementation of 3 type of reports to automate the calculation of insights and KPIs according to the service requested
Integration of plug ins for service deployment	At least 2 plug ins will be integrated in the DaaS service flow
Generate market DaaS cases	Deploy 2 services to companies out of the consortium

4.5 Digital Marketing Data and Resources 2 (Cactus)

This use case focuses on data sharing between Cactus, a technology company specializing in web development and digital marketing, and its clients in the digital marketing sector.

4.5.1 Case description

With the evolution of digital marketing over the years, leveraging social media, mobile devices, data analytics, and personalized targeting, Cactus utilizes client data, primarily from Google and Meta Analytics, to identify the optimal digital marketing tools tailored to each client. Additionally, financial data is considered to develop a comprehensive marketing strategy that aligns with the client's overall business objectives.

4.5.2 Main technical challenges

Cactus' goal is to automate its business procedures, and while pursuing this objective, there are several important challenges that need to be addressed:

- Data Sharing: Establishing a seamless and secure process for clients to share their data with Cactus is crucial. Implementing secure data transfer protocols and providing user-friendly interfaces will facilitate efficient data exchange.
- Data Editing: Cactus needs to develop effective mechanisms to edit client data accurately and efficiently. Implementing robust editing tools and workflows will streamline the process and ensure data accuracy.
- User-Friendly Environment: Creating a user-friendly environment is essential for clients. This includes mobile responsiveness, minimizing downtime, and adhering to stringent security protocols. Prioritizing intuitive interfaces and responsive designs will enhance the overall user experience.
- Multilingual features: Given the geographical diversity of clients, it is important to consider language requirements. Contracts should be available in the client's language all the other data and information could be in English.

4.5.3 Main users/stakeholders

- Account Manager: The Account Manager at Cactus is responsible for managing multiple client accounts, engaging in regular communication with clients,

assessing their needs and requirements, and evaluating their overall situation to provide appropriate solutions and support.

- Competitors refer to other digital agencies that are actively interested in acquiring data from Cactus. These agencies operate in the same industry and compete with Cactus in providing digital marketing services to clients.
- Marketing Manager: The Marketing Manager at Cactus oversees the team of Account Managers, guiding and supervising their activities. They play a crucial role in deciding the most suitable marketing tools and strategies to be employed for each client, ensuring effective and tailored solutions are implemented.
- Customer: A customer is an individual who directly experiences and benefits from the services provided
- Sales Manager: The Sales Manager, as a Cactus employee, serves as the initial point of contact for customers, guiding them through the company's procedures and ensuring positive interaction.
- Cactus: Cactus is the company dedicated to managing the available resources.

4.5.4 Datasets

In this case description, the data used are from the following channels: Google Analytics, Meta Analytics, Google Ads, Sales Data, and the P&L statement. Specifically, the data collected from Google Analytics include Visitors, ROAS, Click-through Rate (CTR), Click per Cost (CPC), Quality Score, Bounce Rate, Channels, and Conversion Rate. From Meta Analytics, Cactus consider Budget, ROAS, Landing Page View, Cost per Lead (CPL), Reach, Impressions, and Frequency. Cactus also gathers data from Google Ads, including Quality Score, ROAS, Budget, Click-through Rate (CTR), and Click per Cost. Cactus has the ability to incorporate Sales data and Profit and Loss statements to determine the best digital marketing tools for their clients. By analyzing all these data, Cactus can identify the weaknesses and strengths of their clients and determine the tools that will help boost their sales.

A more detailed description of the datasets relevant for this pilot can be found in Annex 3.

4.5.5 User Stories and Requirements

Table 15. User stories for the Digital Marketing (Cactus) pilot.

ID	Use case name	As a <type of user>	I want to <goal/objective>	So that <benefit/result/some reason>
US_DM_CAC_1	Sell Data	Client	access data	it will be easier to make decisions and understand its business environment
US_DM_CAC_2	Sell Data	Competitor	Give/receive data	It understands what are the best KPIs in each market
US_DM_CAC_3	Sell Data	Competitor	want to have access to competitive intelligence reports	benchmark my performance against my competitors and identify areas for improvement

US_DM_CAC_4	Sell Data	Marketing Manager	have access to customer data managed by account managers	I can obtain an overall picture of the variables that determine performance and make data-driven decisions to improve the marketing strategy.
US_DM_CAC_5	Sell Data	Account manager	have access to customer data for the customers I manage	I can obtain an overall picture of the variables that determine performance and make informed decisions to improve customer satisfaction and loyalty.
US_DM_CAC_6	Sell Data	Cactus (The Company)	to establish a data marketplace	we can offer a new product to our customers and potential customers and expand our business offerings.
US_DM_CAC_7	Track Performance Ratio	Client	have access to a focused view of my advertising activities and be better informed about their progress through the use of data	I can make informed decisions and improve the effectiveness of my advertising campaigns
US_DM_CAC_8	Track Performance Ratio	Marketing Manager	have access to data that allows me to assess the status of each client managed by each account manager	I can identify whether a customer's situation is not what it should be and take action
US_DM_CAC_9	Track Performance Ratio	Account manager	to have access to data that allows me to assess the situation for each customer I manage,	I can determine whether a customer's situation is not what it should be and take action.
US_DM_CAC_10	Obtain Data from Clients	Client	have a clear process for sending my data to the company for the purpose of the audit	to avoid bureaucracy and ensure the safety of my sensitive commercial information
US_DM_CAC_11	Obtain Data from Clients	Marketing Manager	have direct access to the customer data managed by the account managers	I can oversee the audits and onboarding conducted by the account managers and ensure that they are meeting the performance goals set by the clients.
US_DM_CAC_12	Obtain Data from Clients	Account manager	have direct access to the data of the customers I manage	I can conduct audits and onboarding for these customers and ensure that they are meeting the

				performance goals set by the clients.
US_DM_CAC_13	Obtain Data from Clients	Cactus (The Company)	implement an automated data submission process for customers	improve the efficiency of our audit and onboarding procedures.
US_DM_CAC_14	Negotiate Contract	Client	an online automated process to view and propose or change the terms of my contract with Cactus	have better control and understanding of my relationship with the company
US_DM_CAC_15	Negotiate Contract	Salesman	negotiate the contract terms with potential clients	we can reach a mutually beneficial agreement
US_DM_CAC_16	Negotiate Contract	Marketing Manager	want a centralized and organized platform to store all the contracts signed by the clients managed by account managers	I can easily reference the obligations and receivables of the company with each client
US_DM_CAC_17	Negotiate Contract	Cactus (The Company)	streamline and automate the process of creating and signing contracts with each client.	create a user-friendly online platform where clients can review and sign contracts easily, while also ensuring that all necessary legal terms and conditions are included.
US_DM_CAC_18	Negotiate Contract	Cactus (The Company)	negotiate contracts that are profitable and sustainable for our business	we can continue to provide high-quality services to our clients
US_DM_CAC_19	Share Information with Client (Open Kitchen)	Client	to have access to an online platform	I can view the data that the company holds about me, and who within the company has access to it.
US_DM_CAC_20	Share Information with Client (Open Kitchen)	Cactus (The Company)	to offer its customers a way to control the actions performed within the company.	It gives customers confidence and trust in the work carried out within the company. .

Table 16. Functional requirements for the Digital Marketing (Cactus) pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_DM_CAC_F_1	Customer should be able to log into the Cactus system	US_DM_CAC_19 US_DM_CAC_20	Validation with a Business Case	Must have
REQ_DM_CAC_F_2	Customer shall have an interface to view / check all actions taken by the accounting manager	US_DM_CAC_19 US_DM_CAC_20	Validation with a Business Case	Must have
REQ_DM_CAC_F_3	Customer shall be able to view contract details and other details	US_DM_CAC_19 US_DM_CAC_20	Validation with a Business Case	Must have
REQ_DM_CAC_F_4	The marketing manager and accounting manager shall be provided enough data about customers so as to be able to make strategic decisions	US_DM_CAC_7 US_DM_CAC_8	Synthetic datasets, Validation with a Business Case	Must have
REQ_DM_CAC_F_5	Customer s shall be able to select their privacy preferences	US_DM_CAC_10 US_DM_CAC_12 US_DM_CAC_13	Validation with a Business Case	Must have
REQ_DM_CAC_F_6	Customers shall be able to configure which employees should have access to their data at any time and for how long	US_DM_CAC_10 US_DM_CAC_12 US_DM_CAC_13	Validation with a Business Case	Must have
REQ_DM_CAC_F_7	CACTUS, represented by marketing and account manager, should be able to negotiate with the client the nature and amount of data required to satisfy their requirements	US_DM_CAC_10 US_DM_CAC_12 US_DM_CAC_13	Validation with a Business Case	Must have
REQ_DM_CAC_F_8	Once registered in the (Cactus) system a customer shall be able to select which data to exchange or sell	US_DM_CAC_1 US_DM_CAC_2 US_DM_CAC_3 US_DM_CAC_5	Synthetic datasets, Validation with a Business Case	Must have

REQ_DM_CAC_F_9	The (Cactus) system shall include functionality that valuates the selected customer data (to be exchanged / sold)	US_DM_CAC_1 US_DM_CAC_2 US_DM_CAC_3 US_DM_CAC_5	Synthetic datasets, Validation with a Business Case	Must have
REQ_DM_CAC_F_10	A customer shall be granted access to data from other customers after approval of data owner through negotiation.	US_DM_CAC_1 US_DM_CAC_2 US_DM_CAC_3 US_DM_CAC_5	Synthetic datasets, Validation with a Business Case	Must have
REQ_DM_CAC_F_11	To initiate the negotiation process between a seller and customer the system shall generate a link to the Negotiation plugin (user interface).	US_DM_CAC_14 US_DM_CAC_15 US_DM_CAC_16 US_DM_CAC_17 US_DM_CAC_18	Synthetic datasets, Validation with a Business Case, Collaboration between team members	Must have
REQ_DM_CAC_F_12	The Negotiation plugin (user interface) should allow each party involved in the negotiation process to set their terms and sign a common contract.	US_DM_CAC_14 US_DM_CAC_15 US_DM_CAC_16 US_DM_CAC_17 US_DM_CAC_18	Synthetic datasets, Validation with a Business Case, Collaboration between team members	Must have
REQ_DM_CAC_F_13	The company (Cactus?) and the account manager shall have insight into the customers' requirements at any time.	US_DM_CAC_14 US_DM_CAC_15 US_DM_CAC_16 US_DM_CAC_17 US_DM_CAC_18	Synthetic datasets, Validation with a Business Case, Collaboration between team members	Must have

Table 17. Non-functional requirements for the Digital Marketing (Cactus) pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_DM_CAC_NF_1	All plugins (Privacy and Usage Control & Negotiation) implemented in the Digital Marketing (Cactus) pilot must be interoperable with other software systems and components.		Validation with a Business Case	Must have
REQ_DM_CAC_NF_2	All plugins (Privacy and Usage Control & Negotiation) implemented in the Digital Marketing (Cactus) pilot must have user interfaces that enhance user satisfaction and usability (consistency, feedback & responsiveness and system accessibility) as well as implement error protection mechanisms (guide users through dialogues, prevent mistakes and error handling / recovery through undo and redo functionality)		Validation with a Business Case	Must have
REQ_DM_CAC_NF_3	The system must be operational and accessible to users when they need it. Key elements that are important in this requirement are the following: Uptime, Fault Tolerance, Scalability, Monitoring and Alerting, Disaster Recovery, Planned Downtime. .		Validation with a Business Case	Must have

REQ_DM_CAC_NF_4	All plugins (Privacy and Usage Control & Negotiation) implemented in the Digital Marketing (Cactus) pilot must protect and preserve information from unauthorized access, disclosure, or exposure. Key elements that must be addressed are Data Encryption, Access Control, User Authentication, Secure Communication, Data Masking and Anonymization, Audit Trails and Logging.		Validation with a Business Case	Must have
REQ_DM_CAC_NF_5	All plugins (Privacy and Usage Control & Negotiation) implemented in the Digital Marketing (Cactus) pilot must support adaptability through modular and extensible software components and proper versioning and release management procedures.		Validation with a Business Case	Must have

Table 18. Pilot-specific requirements for the Digital Marketing (Cactus) pilot.

Requirement ID	Description	Source	Verification	Priority
REQ_DM_CAC_F_1	Customer should be able to log into the Cactus system	US_DM_CAC_19 US_DM_CAC_20	Validation with a Business Case	Must have
REQ_DM_CAC_F_2	Customer shall have an interface to view / check all actions taken by the accounting manager	US_DM_CAC_19 US_DM_CAC_20	Validation with a Business Case	Must have
REQ_DM_CAC_F_3	Customer shall be able to view contract details and other details	US_DM_CAC_19 US_DM_CAC_20	Validation with a Business Case	Must have
REQ_DM_CAC_F_4	The marketing manager and accounting manager shall be provided enough data about customers so as to be able to make strategic decisions	US_DM_CAC_7 US_DM_CAC_8	Synthetic datasets, Validation with a Business Case	Must have

REQ_DM_CAC_F_8	Once registered in the (Cactus) system a customer shall be able to select which data to exchange or sell	US_DM_CAC_1 US_DM_CAC_2 US_DM_CAC_3 US_DM_CAC_5	Synthetic datasets, Validation with a Business Case	Must have
REQ_DM_CAC_F_13	The company (Cactus?) and the account manager shall have insight into the customers' requirements at any time.	US_DM_CAC_14 US_DM_CAC_15 US_DM_CAC_16 US_DM_CAC_17 US_DM_CAC_18	Synthetic datasets, Validation with a Business Case, Collaboration between team members	Must have

4.5.6 Key Performance Indicators

Key Performance Indicator	Measure
Efficiency	>10% decrease in users' time to process their needed requirements compared
End-user engagement	>2 reports (deliverables D6.2 and D4.3) on stakeholder perception of how current tools and processes address challenges of data sharing and monetisation (trust, privacy, fair monetisation, transparency).
Automation	Improve automation of contracting by moving >10% of contractual clauses
Shorter process time	Reduce time to achieve data-partnership agreements by >20%
Agreement achievement	>10% agreements achieved through negotiation compared to state-of-the data usage control technology that in the face of similar usage policies would reject a partnership.
Lower compliance effort	Reduce time to reach compliance of a data-processing workflow by >30%.
National law regimes	Engagement of Member States national contract law regimes through workshop on the legal status and effects of smart contracts, with >10 relevant stakeholders (e.g., Member States, policy officers and academics) representing 4 selected Member States.
End-user engagement	>50% increase in in the number of sharing agreements perceived as transparent, measured by feedbacks from end-users.
Effectiveness	≥ 100 individual agreements using Upcast tools
Timeliness	≥ 20% reduction of time-to-market of services requiring data sharing agreements. Baseline: previously deployed services.

5 REQUIREMENTS FOR UPCAST PLUGINS

Following the methodology established in chapter 3, this chapter describes requirements for the UPCAST Plugins derived from the pilot requirements (user requirements including user stories) presented in chapter 4. In addition, these requirements are supplemented by technical requirements sourced from the state-of-art as well as other “intuitive” requirements that must be addressed in UPCAST based on the expert judgement provided by the developers of the plugins.

These requirements constitute the technical requisites that will help scope the system architecture definitions and consequently allow for the software to be developed in the forthcoming tasks of the project.

Below, we summarise the technical requirements per plugin to be developed in UPCAST.

5.1 Resource Specification

Table 19. Functional requirements for the Resource Specification plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_RS_F_1	User must be able to describe a dataset using the UPCAST vocabulary	REQ_BM_F_15, REQ_PA_F_1, REQ_BM_F_14, REQ_HF_F_4, REQ_PA_F_10	Business Case	Must have	Feasible
REQ_RS_F_2	User must be able to describe a data processing operation, either implemented as containerised software or including human processing, using the UPCAST vocabulary	REQ_PA_F_12	Business Case	Must have	Feasible
REQ_RS_F_3	User must be able to validate the resource description generated with the plugin includes all fields required by the specification of the UPCAST vocabulary.		Business Case	Must have	Feasible
REQ_RS_F_4	User must be able to upload a Domain-Specific vocabulary for describing resource, and use it to add further descriptions in the same way as the UPCAST vocabulary (RS_F_1).	REQ_BM_F_15, REQ_PA_F_1	Business Case	Must have	Feasible
REQ_RS_F_5	User can import attributes output from the UPCAST plugins “Privacy and Usage Control”, “Valuation and Pricing” and “Environmental Impact” to augment the description of a resource		Business Case	Must have	Feasible
REQ_RS_F_6	Define data operation custom parameter range	REQ_DM_JOT_F_2	Business Case	Should have	Feasible, but difficult
REQ_RS_F_7	Manage catalog of own resources		Business Case	Must have	Feasible

REQ_RS_F_8	User must be able to visualise the description of a dataset both in RDF form and in a suitable graphical form		Business Case	Must have	Feasible

Table 20. Non-functional requirements for the Resource Specification plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_RS_NF_1	The tool Resource Specification plugin must be usable by people with only a basic understanding of ontologies and vocabularies		Business Case	Must have	Feasible
REQ_RS_NF_2	UPCAST vocabulary should align as much as possible with existing vocabularies from Data Space community: IDSA Information model and DCAT		Competency Question	Must have	Feasible
REQ_RS_NF_3	Manage a joined vocabulary of at least 100 classes and 100 properties		Stress test	Must have	Feasible
REQ_RS_NF_4	Assuming knowledge of the values of the descriptive properties, and excluding the time generating output from the other UPCAST plugins, User must be able to complete the description of a resource in 20 minutes or less.		Business Case	Must have	Feasible

5.2 Resource Discovery

Table 21. Functional requirements for the Resource Discovery plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_RD_F_1	Users can search datasets with keywords on a dataset catalog	REQ_BM_F_10, REQ_PA_F_2, REQ_PA_F_14, REQ_HF_F_5	Business Case	Must have	Feasible
REQ_RD_F_2	Users can search datasets with facets on a dataset catalog	REQ_HF_F_6, REQ_BM_F_10, REQ_PA_F_2, REQ_PA_F_14	Business Case	Must have	Feasible
REQ_RD_F_3	User must be able to search based on an abstract (incomplete) description of a dataset. This should be equivalent to an OR keyword/facet search using the attributes specified in the abstract description.	REQ_PA_F_14	Business Case	Must have	Feasible
REQ_RD_F_4	User must be able to search with keywords on top of an existing Data Processing Operation Catalog.		Business Case	Must have	Feasible
REQ_RD_F_5	User must be able to search with facets on top of an existing Data Processing Operation Catalog.		Business Case	Must have	Feasible
REQ_RD_F_6	User must be able to search based on an abstract (incomplete) description of a data processing operation. This is equivalent to an OR keyword/facet search using the attributes specified in the abstract description		Business Case	Must have	Feasible
REQ_RD_F_7	Given a Data Processing Workflow, user can trigger a search for alternatives for any resource in the workflow		Business Case	Should have	Feasible

REQ_RD_F_8	User should be able to connect remotely to multiple catalogs (assuming they are available on a Web server) to browse and search from all of them.		Business Case	Should have	Feasible - hard to implement
REQ_RD_F_9	Given a Data Processing Workflow, and a set of resource catalogs, recommend resources that could replace or augment the ones in the Data Processing Workflow		Business Case	Should have	Feasible
REQ_RD_F_10	Given a dataset the plugin must be able to return a profile of the dataset.	US_DM_JOT_1	Business Case	Must have	Feasible
REQ_RD_F_11	The plugin should provide functionality to provide users with a reduced sample of the dataset.	US_DM_JOT_1	Business Case	Could have	Feasible - hard to implement
REQ_RD_F_12	Given an owned dataset, search for datasets in a catalog that can augment the input dataset in terms of join/union operations.		Business Case	Could have	Difficult

Table 22. Non-functional requirements for the Resource Discovery plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_RD_NF_1	The plugin must be usable by people with only domain specific knowledge		Stress test	Must have	Feasible
REQ_RD_NF_2	Must support at least 100.000 resources when used on a local catalog		Stress test	Could have	Feasible
REQ_RD_NF_3	Must support connection to at least 5 remote catalogs		Simulation, Business Case	Could have	Feasible

5.3 Data Processing Workflow

Table 23. Functional requirements for the Data Processing Workflow plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_DPW_F_1	The definition of all entities involved in DPWs should be supported at the required detail level, based on the semantic foundation provided.	REQ_BM_F_14	Pilots and lab test	Must have	Feasible
REQ_DPW_F_2	Usage preferences should be able to be suitably formalized, providing adequate expressiveness	REQ_BM_F_1, REQ_PA_F_5	Pilots and lab test	Must have	Feasible
REQ_DPW_F_3	Access and usage constraints should be able to be expressed in the process models in order to consistently reflect compliant execution	REQ_BM_F_2, REQ_PA_F_6	Pilots and lab test	Must have	Feasible
REQ_DPW_F_4	The definition of conditional execution of tasks and conditional data flow depending on context, purpose or intra-workflow dependencies should be supported		Pilots and lab test	Must have	Feasible
REQ_DPW_F_5	The definition of the entity that performs a processing step, but also the entity that initiates a DPW, should be supported		Pilots and lab test	Must have	Feasible
REQ_DPW_F_6	The definition of the operations performed through a data processing step should be supported	REQ_BM_F_6, REQ_BM_F_7, REQ_PA_F_11, REQ_PA_F_12	Pilots and lab test	Must have	Feasible

REQ_DPW_F_7	The definition of the data asset which is accessed or upon which an operation is performed should be supported	REQ_BM_F_6, REQ_PA_F_11	Pilots and lab test	Must have	Feasible
REQ_DPW_F_8	The definition of the data exchanged between processing steps should be supported	REQ_BM_F_6	Pilots and lab test	Must have	Feasible
REQ_DPW_F_9	DPW models should be able to accurately represent cases of cross-domain data sharing at various levels (organisations, states, regulatory domains, etc.)		Pilots and lab test	Must have	Feasible
REQ_DPW_F_10	A GUI shall be made available to model DPWs.	REQ_BM_F_5, REQ_PA_F_9	Pilots and lab test	Must have	Feasible
REQ_DPW_F_11	The DPW tool should support the definition of data workflows comprising atomic actions and decisions through a GUI	(A more detailed variant of the REQ_DPW_F_10 requirement). Hence: REQ_BM_F_5, REQ_PA_F_9	Pilots and lab test	Must have	Feasible
REQ_DPW_F_12	The DPW tool should be able to execute a workflow	REQ_BM_F_8, REQ_BM_F_9, REQ_PA_F_13	Pilots and lab test	Must have	Feasible, technically quite difficult
REQ_DPW_F_13	The DPW should be seamlessly integrated with a monitoring service to collect information on the progress of the execution of a workflow		Pilots and lab test	Must have	Feasible
REQ_DPW_F_14	The DPW should provide a control interface to allow a user to intervene (stop, pause, resume, inspect) on the execution of a workflow.		Pilots and lab test	Could have	Feasible, difficult to implement

Table 24. Non-functional requirements for the Data Processing Workflow plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_DPW_NF_1	The DPW should support the secure execution of a workflow		Pilots and lab test	Must have	Feasible, difficult to implement

5.4 Privacy and Usage Control

Table 25. Functional requirements for the Privacy and Usage Control plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_PUC_F_1	The plugin should provide the means for system operation in accordance to access and usage control policies. Moreover, policy based access control should be inline with Attribute Based Access Control (ABAC) paradigm		Business Case	Should have	Feasible - medium difficult
REQ_PUC_F_2	Access and usage control policies should be fine-grained, following hierarchical organisation of related entities; that is, policies should be defined in different granularities of the underlying concepts, particularly the resources to be protected.	REQ_DM_CAC_F_5, REQ_DM_CAC_F_6, REQ_DM_CAC_F_7,	Business Case	Must have	Feasible - medium difficult

REQ_ PUC_F_3	All applicable parameters must be taken into account, such as roles, attributes, contextual parameters, the purpose under which the underlying action should be executed, prior actions (history), etc.		Business Case	Could have	Difficult to implement
REQ_ PUC_F_4	Access and usage control policies must be machine-readable, so that they can be processed by a policy engine		Business Case	Must have	Feasible
REQ_ PUC_F_5	The plugin must incorporate a policy engine, able to make decisions as regards access and usage of data and other resources		Business Case	Must have	Feasible - medium difficult
REQ_ PUC_F_6	Policies must support the definition of complementary / forbidden actions that must/ must not take place upon and/or prior to their enforcement	REQ_DM_CAC_F_5, REQ_DM_CAC_F_6, REQ_DM_CAC_F_7,	Business Case	Must have	Feasible
REQ_ PUC_F_7	The plugin should provide for implicit definition / propagation of policies, i.e., rules defined for high-level concepts should be propagated to more specific ones without the need to add new specific rules		Business Case	Should have	Feasible

REQ_ PUC_F_8	The plugin should provide for advanced decision making. Policy decision point should provide for request transformation where possible; instead of simply allowing or denying an incoming access/usage request, it should provide for request transformation (e.g., allow access to/usage of parts of the requested data) and/or prescribe/forbid the execution of subsequent actions	REQ_PA_F_13	Business Case	Must have	Feasible - difficult
REQ_ PUC_F_9	The plugin should provide for advanced conflict resolution and rules merging, i.e., mechanisms for the elimination of deprecated policies (i.e., overridden by other policies), as well as for conflict resolution between data provider and data consumer access and usage constraints (consistent enforcement of data provider constraints, without compromising core policies)		Business Case	Must have	Feasible - medium difficult
REQ_ PUC_F_10	The plugin should provide a graphical user interface for the specification of access and usage control rules (Data Consumer side)	REQ_BM_F_1, REQ_PA_F_5, REQ_PA_F_6, REQ_HF_F_3	Business Case	Must have	Feasible
REQ_ PUC_F_11	The plugin should provide the ability to data subjects/providers to define their data usage constraints in a machine-readable form through a graphical user interface	REQ_BM_F_2	Business Case	Must have	Feasible

REQ_ PUC_F_12	By means of access and usage control rules, it should be possible to explicitly determine what is the necessary accuracy (detail level) of information that is necessary for a certain purpose and under certain conditions.		Business Case	Should have	Feasible - medium difficult
REQ_ PUC_F_13	Access and usage control should make possible that a data structure can be collected or processed in parts, based on certain criteria; that is, there must be selective handling of different parts of such structure	REQ_HF_F_4	Business Case	Must have	Feasible - medium difficult
REQ_ PUC_F_14	The plugin should enable decisions about data collection, processing, storage and communication to be made on the basis of the underlying purpose	REQ_BM_F_9	Business Case	Must have	Feasible

Table 26. Non-functional requirements for the Privacy and Usage Control plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_ PUC_NF_1	Policies must support the specification of (i) permissions, i.e., actions that are allowed to take place; (ii) prohibitions, i.e., actions that are prohibited to take place; (iii) obligations, i.e., actions that must take place	REQ_DM_CAC_F_5, REQ_DM_CAC_F_6, REQ_DM_CAC_F_7,	Business Case	Should have	Feasible - medium difficult

REQ_ PUC_NF_2	The plugin should provide for ensuring the compliance with data protection legislation; for instance, the plugin should allow the definition of policies containing any applicable GDPR-related rules as adapted to the specific organisation's needs		Business Case	Must have	Feasible - medium difficult
REQ_ PUC_NF_3	The plugin should foster context-awareness: it should provide the means for semantic definitions of applicable contextual parameters, while decision making on data collection, processing, storage and communication should be able to consider contextual aspects		Business Case	Could have	Difficult to implement
REQ_ PUC_NF_4	Access and usage control should provide the means for adjusting the detail level of data that are collected, processed, stored and communicated, in a manner as automatic as possible		Business Case	Should have	Feasible
REQ_ PUC_NF_5	Access and usage control rules should define what data are considered proportional for a certain purpose and under certain conditions		Business Case	Could have	Feasible - medium difficult

REQ_ PUC_NF_6	The plugin should provide for semantic information management: the underlying information model should (i) support a variety of concepts, including data types, purposes, roles, operations, attributes, organisations, context, etc.; (ii) provide coherent semantic definitions of the underlying concepts; (iii) support the hierarchical organisation of concepts, including different semantics, such as generalisation / particularisation, inclusion, etc.		Business Case	Should have	Difficult to implement
REQ_ PUC_NF_7	The plugin must provide mechanisms for specifying the compatibility between collection and processing purposes and provide means for defining prevention rules regarding incompatible purposes		Business Case	Should have	Feasible
REQ_ PUC_NF_8	The plugin should enable transformation of governance metadata to the underlying machine-readable access and usage control policies and vice versa		Business Case	Must have	Feasible - medium difficult

5.5 Pricing and Valuation

In this section the functional and non-functional requirements for the Pricing plugin and the Valuation plugin are presented separately.

Table 27. Functional requirements for the Pricing plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_PR_F_1	The plugin should provide similar products to a data product specified by the users found in real commercial data marketplaces. .		Business Case and/or data marketplace datasets	Must have	Feasible
REQ_PR_F_2	The plugin requires the characteristics of data products including metadata fields included in commercial data marketplaces such as their description, categories, units, update rate, geographical and time scope, etc.	State of the art	Business Case and/or data marketplace datasets	Must have	Feasible
REQ_PR_F_3	The plugin should provide flexible functionality to label data products of a data marketplace using the categories and criteria of a “source” data marketplace based on their descriptions.		Business Case and/or data marketplace datasets	Must have	Feasible - Difficult
REQ_PR_F_4	The pricing plugin will return a range of prices for a data product specified by the user based on the estimation of different price regressors fitting the price of similar commercial products in data marketplaces, already stored in the plugins database.	REQ_HF_F_1, REQ_DM_JOT_F_5	Business Case and/or data marketplace datasets	Must have	Feasible - Difficult

REQ_PR_F_5	Using Explainable AI, the pricing plugin will be able to inform the buyers about the most relevant features when building its price predictions, producing a list of relevant features and a percentage of impact in the prediction of the price.	State of the art	Business Case and/or data marketplace datasets	Must have	Feasible - Difficult
REQ_PR_F_6	End users will be able to access this functionality through a REST API		Business Case and/or data marketplace datasets	Must have	Feasible
REQ_PR_F_7	The plugin should record the history of user transactions/interactions for pricing, security, logging, insights and transparency.		Business Case and/or data marketplace datasets	Should have	Feasible
REQ_PR_F_8	The plugin must support an admin user that manages all aspects of the plugin including the database, manipulating data, updating/enriching datasets, training models, interconnections, and permission management		Business Case and/or data marketplace datasets	Should have	Feasible
REQ_PR_F_9	The database and architecture will be centralised with a central server providing all the functionality, data, models, and API responses.		Business Case and/or data marketplace datasets	Must have	Feasible
REQ_PR_F_10	The plugin should incorporate market conditions and marketplace interactions to produce a dynamic price range for datasets.	REQ_HF_F_1, REQ_DM_JOT_F_5	Validation with data marketplace	Must have	Feasible - Difficult

Table 28. Non-functional requirements for the Pricing plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_PR_NF_1	The plugin must be usable by end users with only domain specific knowledge to get price ranges and similar products in the database.		Business Case	Must have	Feasible
REQ_PR_NF_2	The architecture of the plugin and its various management scripts should be well documented and available/easily accessible to the plugin admin.		Business Case	Should have	Feasible - Difficult
REQ_PR_NF_3	Must be able to support at least 10.000 price references		Business Case	Should have	Feasible
REQ_PR_NF_4	Must support data from at least 10 data marketplaces		Business Case	Should have	Feasible
REQ_PR_NF_5	The plugin must respond in a few seconds to applications in the REST API		Business Case	Should have	Feasible
REQ_PR_NF_6	The pricing plugin must include security mechanisms to prevent abuse and misuse from users based on the activity registered in the REST API		Business Case	Should have	Feasible
REQ_PR_NF_7	The plugin needs to be compliant with the EU regulations (such as GDPR, Data Act, AI Act)		Validation with legal partners + Privacy and Usage Control plugin	Must have	Feasible

REQ_PR_NF_8	The plugin should be designed in a modular and maintainable way, allowing for flexibility to add and/or update the machine learning models, labels etc, enabling extensibility.		Validation during development and testing	Should have	Feasible

Table 29. Functional requirements for the Valuation plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_VAL_F_1	The plugin will provide functions to calculate the relative value of a set of N data sources for an integration/ML task given by a valuation function.	US_HF_1, US_HF_3, REQ_DM_CAC_F_9	Validation with use case – NIS and Digital Marketing (Cactus)	Must have	Feasible - Difficult
REQ_VAL_F_2	The plugin will be implemented using a Python library that can be integrated with new ML models and valuation functions by users.		Validation with use case - NIS	Must have	Feasible
REQ_VAL_F_3	The plugin will provide functions to carry out exact calculations of the Shapley value of data sources, and different approximation algorithms.	REQ_HF_F_2	Validation with use case - NIS	Must have	Feasible
REQ_VAL_F_4	Shapley approximation algorithms will include tunable parameters to balance precision and execution time.		Validation with use case - NIS	Must have	Feasible

Table 30. Non-functional requirements for the Valuation plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_VAL_NF_1	The plugin must be usable by ML developers with only domain specific knowledge to get price ranges and similar products in the database.		Business Case	Must have	Feasible

5.6 Environmental Impact

Table 31. Functional requirements for the Environmental Impact plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_EE_F_1	The plugin should generate an energy profile of a dataset	REQ_BM_F_20, REQ_BM_F_21, REQ_PA_F_15	Business Case and/or data marketplace datasets	Must have	Feasible
REQ_EE_F_2	The plugin should display relevant energy consumption metrics, in a graphical manner, based on processes applied to the datasets	REQ_BM_F_20, REQ_BM_F_21, REQ_PA_F_15	Validation with pilots and simulations	Must have	Feasible
REQ_EE_F_3	The plugin should use explainability techniques to explain the factors contributing to the energy consumption		Business Case and/or data marketplace datasets	Should have	Feasible

REQ_EE_F_4	The plugin requires hardware information such as server, platform (physical/cloud) and data centre characteristics		Validation with pilot and/or data marketplace infrastructure	Must have	Feasible
REQ_EE_F_5	The plugin should continuously improve its energy profiling as the dataset quantity increases, and requires dataset metadata	REQ_BM_F_20, REQ_BM_F_21, REQ_PA_F_15	Business Case and/or data marketplace datasets	Must have	Feasible - Difficult to implement
REQ_EE_F_6	The plugin should calculate the energy cost of storing and updating a dataset	REQ_BM_F_20, REQ_BM_F_21, REQ_PA_F_15	Business Case and/or data marketplace datasets	Must have	Feasible
REQ_EE_F_7	The plugin should model the energy footprint of atomic operations related to the access of resources	REQ_BM_F_20, REQ_BM_F_21, REQ_PA_F_15	Business Case and/or data marketplace datasets	Should have	Feasible
REQ_EE_F_8	The energy profile of a dataset should be used as a feature to influence its price		Validation with pilot and/or data marketplace datasets and pricing plugin	Should have	Feasible

Table 32. Non-functional requirements for the Environmental Impact plugin.

Requirement ID	Description	Related Pilot Requirements	Verification	Priority	Technical Feasibility
REQ_EE_NF_1	The plugin must be usable only by internal stakeholders, other plugin developers and relevant data marketplaces	REQ_BM_F_20, REQ_BM_F_21, REQ_PA_F_15		Must have	Feasible

REQ_EE_NF_2	The plugin should be compatible with different operating systems, environments and platforms, including both physical and cloud infrastructure		Validation with pilot and/or data marketplace infrastructure	Must have	Feasible
REQ_EE_NF_3	The plug-in's performance should not degrade significantly when monitoring large or complex dataset processes		Validation with pilot and/or data marketplace datasets + operations	Should have	Feasible
REQ_EE_NF_4	The plugin needs to be compliant with the EU regulations (such as GDPR, AI Act)		Validation with legal partners + privacy plugin	Must have	Feasible
REQ_EE_NF_5	The plug-in should be designed in a modular and maintainable way, allowing for easy updates and bug fixes		Validation with tech partners	Could have	Feasible

5.7 Integration and Exchange

Table 33. Functional requirements for the Integration and Exchange plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_IE_F_1	System must be capable of integrating data by forward chaining (chasing) and backward chaining (query rewriting).	State of the art	Synthetic datasets	Must have	Feasible - medium hard
REQ_IE_F_2	User must be able to choose whether the data to integrate comes from a local source or a remote source.		Synthetic or real datasets	Must have	Feasible
REQ_IE_F_3	User must be able to define and create schema mappings with help of a graphical interface.		Synthetic or real datasets	Must have	Feasible - medium hard
REQ_IE_F_4	User must be able to define constraints and functional dependencies for target database.		Synthetic datasets	Must have	Feasible
REQ_IE_F_5	User must be able to integrate data that may be structured under different standard formats (e.g., TSV, JSON, RDF).	REQ_BM_F_19/REQ_PA_F_15, REQ_BM_F_12, REQ_PA_F_4	Synthetic datasets	Should have	Feasible - medium hard
REQ_IE_F_6	User must be able to integrate data from various sources concurrently.	REQ_BM_NF_7	Synthetic datasets	Should have	Feasible
REQ_IE_F_7	User must be able to choose whether to integrate data through query rewriting or materialization of sources.	State of the art	Business Case	Should have	Feasible
REQ_IE_F_8	User must be able to view a preliminary view of the resulting integration.		Business Case	Could have	Feasible

REQ_IE_F_9	User must be able to choose to output their data in their desired format from a list of options.	REQ_BM_F_19/REQ_PA_F_15	Synthetic datasets	Should have	Feasible
REQ_IE_F_10	System should be able to recognise sets of dependencies that allow for more efficient data integration.		Synthetic datasets	Should have	Feasible - medium hard
REQ_IE_F_11	User must be capable of executing queries over the integrated data with a standard query language (e.g., SQL or SPARQL).		Synthetic or real datasets	Must have	Feasible - medium hard
REQ_IE_F_12	User must be able to include a number of remote sources (either databases or APIs).	REQ_BM_NF_7	Business Case	Must have	Feasible - medium hard
REQ_IE_F_13	System should operate with local sources if there is no internet connection available.		Business Case	Should have	Feasible
REQ_IE_F_14	System should compile a history of operations done by the user. In addition, it should allow the user to repeat these actions.		Business Case	Could have	Feasible - medium hard
REQ_IE_F_15	System should work independently from other UPCAST plugins.		Synthetic datasets	Should have	Feasible - medium hard

Table 34. Non-functional requirements for the Integration and Exchange plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_IE_NF_1	The plugin must be usable by people with just some domain specific knowledge.		Business Case	Must have	Feasible
REQ_IE_NF_2	The definition and creation of schema mappings must be intuitive.		Business Case	Must have	Feasible
REQ_IE_NF_3	Data integration should terminate in a reasonable amount of time.		Business Case	Must have	Feasible
REQ_IE_NF_4	System should provide guidance or suggestions when choosing between forward and backward chaining.		Synthetic datasets	Could have	Feasible
REQ_IE_NF_5	System should warn users when attempting to integrate data under constraints that are not guaranteed to terminate.		Synthetic datasets	Could have	Feasible
REQ_IE_NF_6	System should have the option to set a timeout for some of its processes (forward chaining in particular).		Synthetic datasets / Business Case	Should have	Feasible

5.8 Negotiation and Contracting

Table 35. Functional requirements for the Negotiation and Contracting plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_NE_F_1	Users must be able to make contracts semi-automatically that follow the IDSA standard (with some variance allowed).		Business Case	Must have	Feasible
REQ_NE_F_2	System should be able to detect (some) conflicts between offer and request contracts.		Business Case	Must have	Feasible
REQ_NE_F_3	Users should be able to negotiate with other parties whenever there is a conflict in their respective contracts.	REQ_BM_F_3, REQ_PA_F_7,	Business Case	Must have	Feasible
REQ_NE_F_4	Users must be able to accept, reject, or continue negotiations.		Business Case	Must have	Feasible
REQ_NE_F_5	System should evaluate privacy and usage settings from all parties.	REQ_PA_F_13, REQ_DM_CAC_F_10	Synthetic/Real policies	Must have	Feasible
REQ_NE_F_6	System should evaluate the environmental impact of the relevant datasets.		Synthetic/Real datasets	Must have	Feasible
REQ_NE_F_7	System should evaluate the pricing of all relevant datasets and inform all parties.		Synthetic/Real datasets	Must have	Feasible
REQ_NE_F_8	System should provide a visualisation of the result of the negotiation (i.e., agreement, rejection).	REQ_BM_F_4, REQ_PA_F_8	Synthetic use cases	Must have	Feasible

REQ_NE_F_9	Users must be able to generate contracts that contain and use boilerplate text whenever the contract contains clauses that can only be expressed through natural language.		Synthetic/Real templates	Should have	Feasible
REQ_NE_F_10	Once an agreement has been reached, the system should draft an agreement contract that can be reviewed by all parties.	REQ_BM_F_4, REQ_PA_F_8	Synthetic/Real use-cases	Must have	Feasible
REQ_NE_F_11	Whenever a conflict is detected between contracts, the system must highlight said conflicts so all parties can review them.		Synthetic/Real contracts	Should have	Feasible
REQ_NE_F_12	Users may define that certain parts of their contracts are non-negotiable, so the system must immediately reject contracts that conflict with any of these parts.		Synthetic/real use-cases	Could have	Feasible
REQ_NE_F_13	System must ensure that all negotiations are carried out while respecting and protecting all users' privacy and confidentiality.		Business Case	Must have	Feasible - hard to implement
REQ_NE_F_14	Users must be able to edit policies in contracts with the help of a graphical interface.	REQ_DM_CAC_F_11, REQ_DM_CAC_F_12	Business Case	Must have	Feasible

Table 36. Non-functional requirements for the Negotiation and Contracting plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_NE_NF_1	Plugin must be usable by people with domain-specific knowledge (not necessary knowledge of how the contracts are processed).		Business Case	Should have	Feasible

5.9 Safety and Security

The requirements for the Safety and Security plugin can be considered transversal as all plugins must conform to established guidelines and practices related to safe and secure data exchange. In that sense, these requirements could be regarded as system-wide requirements (Chapter 6), but we have chosen to separate these out from the other system-wide requirements since the latter is defined as a module in the UPCAST ecosystem.

Table 37. Functional requirements for Safety and Security plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_SEC_F_1	Data processing must be executed in a secure sandbox environment	REQ_BM_F_13	Validation with pilot by authentic and authorized user	Must have	Feasible
REQ_SEC_F_2	CIA requirements must be guaranteed for available datasets	REQ_BM_NF_1	CIS Controls (Center for Internet Security) provided guidelines and best practices can be used for verification for CIA	Must have	Feasible

REQ_SEC_F_3	Implementation of mechanisms to authenticate and verify the identity of users participating in the marketplace, such as through secure login systems or digital signatures.		Verification of authentication by using the identity of an individual or entity trying to access a system	Must have	Feasible
REQ_SEC_F_4	Implementation of access control policies to ensure that users have appropriate access rights and permissions based on their roles within the marketplace.	REQ_DM_CAC_NF_4	examining the implementation of user authentication, authorization mechanisms in a pilot	Must have	Feasible
REQ_SEC_F_5	Implementation of strong encryption algorithms to protect the confidentiality of data during transmission and storage.	REQ_DM_CAC_NF_4	Assess the encryption algorithms and protocols used by the organization. Verify that they are recognized and widely accepted cryptographic standards, such as Advanced Encryption Standard (AES) or Transport Layer Security (TLS) protocols.	Must have	Feasible
REQ_SEC_F_6	Implementation of mechanisms to validate the integrity of data to ensure that it has not been tampered with or modified during transit or storage.		Implementing data validation and quality control measures helps identify and correct errors or inconsistencies in data. By regularly validating data against predefined rules and conducting quality control checks, you can ensure the integrity of the data.	Must have	Feasible
REQ_SEC_F_7	Utilization of distributed ledger technologies (e.g., blockchain) to maintain an immutable record of transactions and data modifications, ensuring data integrity.		Validation with pilot by authentic and authorized user	Must have	Feasible

REQ_SEC_F_8	Employ secure communication protocols (e.g., HTTPS, TLS) for data transmission between participants, protecting against interception or unauthorized access.		Verify the strength of the encryption algorithms used in the secure communication. Ensure that the protocols implement robust encryption algorithms, such as AES (Advanced Encryption Standard), and that they are configured to use sufficiently long and secure encryption keys.	Must have	Feasible
REQ_SEC_F_9	Allow data providers to specify access controls and permissions for their data, ensuring that only authorized entities can access and utilize it.		Verify access control policies in a pilot	Must have	Feasible
REQ_SEC_F_10	Utilization of smart contracts to enforce agreements, terms and conditions			Must have	Feasible
REQ_SEC_F_11	Maintenance of comprehensive audit logs of data transactions, access attempts, and modifications for monitoring and forensic analysis.		Verify that the necessary systems, applications, are configured to generate and transmit logs to a centralized logging system in a pilot. Confirm that logs capture relevant events, including authentication attempts, access control changes, system configuration changes, and critical system activities	Must have	Feasible
REQ_SEC_F_12	Ensuring compliance with relevant data protection regulations (e.g., GDPR) and industry standards.		Validation with pilot by authentic and authorized user	Must have	Feasible

Table 38. Non-functional requirements for Safety and Security plugin.

Requirement ID	Description	Related Pilot Requirements	Verification	Priority	Technical Feasibility
REQ_SEC_NF_1	Scalability: The marketplace should be able to handle a growing number of participants, data transactions, and data volumes without compromising security.	REQ_DM_CAC_NF_5	Verify the scalability of the underlying database or storage infrastructure used in the data marketplace in a pilot. Assess the database's ability to handle increased data volumes, perform efficient queries, and scale horizontally if needed. Evaluate the data storage mechanisms, such as distributed file systems or cloud	Must have	Feasible
REQ_SEC_NF_2	The marketplace should provide efficient and responsive operations, including data retrieval, data sharing, and authentication processes, to minimize delays and ensure a smooth user experience.		Measure and analyze the response times of critical operations in the data marketplace. Identify the latency introduced by various components, such as database queries, network communications, or external service integrations. Monitor and optimize	Must have	Feasible

			response times to ensure fast and efficient user interactions.		
REQ_SEC_NF_3	Marketplace should implement fault tolerance mechanisms to handle failures in individual nodes or components.	REQ_DM_CAC_NF_3	Verify the implementation of redundancy and replication mechanisms to ensure fault tolerance. Assess the redundancy of critical system components such as databases, servers, or network infrastructure. Test failover mechanisms to ensure seamless transition to backup systems in case of failure. Validate that replicated data remains consistent across different instances or locations.	Must have	Feasible
REQ_SEC_NF_4	The marketplace should be available and accessible to users consistently, with minimal planned or unplanned downtime.	REQ_DM_CAC_NF_3	Perform load testing to evaluate the system's availability under various user loads and traffic	Must have	Feasible

			conditions. Simulate high concurrent user access, transaction volumes, and data requests to assess the system's responsiveness and ability to handle the expected workload. Measure response times and ensure that the system remains available and performs well under peak loads.		
REQ_SEC_NF_5	Marketplace should provide compatibility and interoperability with different data formats, protocols, and standards to facilitate seamless integration with various data sources and consumers.		Verify that the data marketplace supports a wide range of data formats commonly used in the industry. Test the system's ability to ingest, process, and transform data in different formats such as JSON, XML, CSV, Parquet, Avro, or specific industry-specific formats. Ensure that the system can handle various data structures, encoding	Must have	Feasible

			schemes, and data representations.		
REQ_SEC_NF_6	The system should adhere to strict privacy and data protection regulations, such as GDPR or CCPA, to protect user data and ensure compliance.			Must have	Feasible

5.10 Monitoring

Like the Safety and Security plugin described in the previous chapter, the Monitoring plugin can be regarded as a plugin offering transversal functionality supporting the UPCAST ecosystem by collecting data to monitor the behavior from systems, components and infrastructure and ensure compliance with specifications and overall health.

Table 39. Functional requirements for the Monitoring plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility
REQ_MON_F_1	The monitoring plugin must be able to collect data from different sources including access and use of datasets.	REQ_BM_NF_5	Validation with pilots and simulations	Must have	Feasible
REQ_MON_F_2	The monitoring plugin must be able to store monitoring data for a configurable duration	REQ_BM_NF_5	Validation with pilots and simulations	Must have	Feasible
REQ_MON_F_3	The monitoring plugin must be able to use a JSON-based data model for the collected monitored data		Validation with pilots and simulations	Must have	Feasible

REQ_MON_F_4	The monitoring plugin must be able to provide visualizations of the collected data		Validation with pilots and simulations	Must have	Feasible
REQ_MON_F_5	The monitoring plugin should be able to integrate with external services that may further analyze the collected data		Validation with pilots and simulations	Must have	Feasible (Difficult to implement)

Table 40. Non-functional requirements for the Monitoring plugin.

Requirement ID	Description	Related Pilot Requirements	Verification	Priority	Technical Feasibility
REQ_MON_NF_1	The monitoring service must be fault tolerant		Validation with pilots and simulations	Should have	Feasible
REQ_MON_NF_2	The monitoring service must be configurable for the type and sources of collected data.		Validation with pilots and simulations	Must have	Feasible
REQ_MON_NF_3	CIA requirements should be guaranteed for all monitoring data		Validation with pilots and simulations	Must have	Feasible

5.11Federated Machine Learning

Like the Safety and Security plugin described in the previous chapter, the Monitoring plugin can be regarded as a plugin offering transversal functionality supporting the UPCAST ecosystem by collecting data to monitor the behavior from systems, components and infrastructure and ensure compliance with specifications and overall health.

Table 41. Functional requirements for the Monitoring plugin.

Requirement ID	Description	Related Pilot Requirements / User Stories	Verification	Priority	Technical Feasibility

6 SYSTEM-WIDE REQUIREMENTS

The system-wide requirements presented in this chapter are requirements that are relevant for the UPCAST ecosystem as a whole. In addition to the system-wide requirements listed in this section, there are also requirements related to some of the plugins described in Chapter 5, notably the Safety & Security and Monitoring plugins, that can be considered system-wide in the sense that their functionality support the operations of the other UPCAST plugins.

Requirement ID	Description	Related Pilot Requirements	Verification
REQ_SYS_1	Technical interoperability with regards to support for different communication patterns, message routing and dispatching, etc. must be supported.	REQ_BM_F_13, REQ_BM_F_18, REQ_DM_CAC_NF_1	Must have
REQ_SYS_2	Semantic interoperability with regards to support for conformance to a common data model/vocabulary must be ensured.	REQ_BM_F_11, REQ_BM_F_12, REQ_DM_CAC_NF_1	Must have
REQ_SYS_3	Graphical User Interfaces (GUIs) must be provided for all plugins	REQ_BM_NF_8, REQ_PA_F_14, REQ_PA_NF_2, REQ_HF_NF_2, REQ_DM_JOT_NF_2, REQ_DM_CAC_NF_2, REQ_BM_F_1, REQ_BM_F_2, REQ_BM_F_3, REQ_BM_F_4, REQ_BM_F_5, REQ_PA_F_5, REQ_PA_F_6, REQ_PA_F_7, REQ_PA_F_8, REQ_PA_F_9	Must have
REQ_SYS_4	Support for multilingualism should be supported by the plugins	REQ_PA_NF_1	Should have
REQ_SYS_5	System components should be developed in a modular, extensible and adaptable way.	REQ_DM_CAC_NF_5	Must have

7 LEGAL FRAMEWORK AND REQUIREMENTS

This chapter presents the EU legal framework relevant to the UPCAST project and in particular to the activities and data processing operations envisaged in the pilots. It is structured as follows: The first section presents the applicable legal framework and elicits specific and actionable requirements from four legal domains: a) data protection law; b) data governance and data regulation; c) competition law; and d) automated and smart contracts. The second section delves into the five pilots envisaged in UPCAST. Depending on the legal questions raised by each pilot, the section maps out the requirements to each pilot to ensure it complies with relevant law.

7.1 Applicable legal framework and requirements

This section first introduces various categories of data for the purposes of EU law. It then briefly describes each legal framework and the main legal notions that will govern the operations within the UPCAST platform, and draws legal requirements from those notions.

Data categorisation

'Data' is a complex dimension to regulate. The term potentially refers to any piece of information existing in any form, relating to ideas, facts, people, companies, etc. The Data Governance Act defines it as 'any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording'.[39]

EU law mainly takes a sectoral/category-based approach to data regulation, with specific rules applying to different categories of data. The main distinction that was drawn – and is still quite relevant – is between data that reveals identifying information, even indirectly, about natural persons, and data that do not, i.e., the distinction between **personal** and **non-personal data**, with the former being regulated by data protection law (see below).

As to non-personal data, it is a 'negatively' formulated characterisation for data that does not (indirectly) identify natural persons. They are broadly regulated by the Free Flow of Non-Personal Data (FFD) Regulation.[40] Data about organisations and companies may be confidential and contain **trade secrets**;[41] data handled by public administrations may, on certain conditions, qualify as **open data**,[42] publicly accessible and not belonging to rightsholders. Data sets containing personal and non-personal data sets are then regulated by the Data Governance Act and the upcoming Data Act when it comes to governance and data sharing. These two legislative instruments are the first cornerstones to an EU overarching framework for data.

Data protection law

The EU legal order contains a framework for the protection of personal data. Their importance is paramount in EU law to the point where the right to the protection of personal data and the right to privacy are enshrined in the Charter of Fundamental Rights of the EU,[43] thus enjoying a quasi-constitutional status. Secondary legislation provides the main framework governing

[39] Article 2(1) of Regulation (EU) 2022/868 ('Data Governance Act').
[40] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union.
[41] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.
[42] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast).
[43] Article 8 of the EU Charter of Fundamental Rights.

data subjects' rights and obligations for entities processing personal data, and is provided in the General Data Protection Regulation (GDPR[44]). The GDPR contains harmonised rules on data protection, although certain substantive and procedural matters are still within the competences of national data protection law.

Personal data are understood as any information relating to an identified or identifiable natural person.[45] It is key to remember that information can be personal data even if it indirectly enables identifying the data subject.[46] Looking at the UPCAST platform, there are four main data protection dimensions that are worth considering: a) the rules governing the processing of personal data by the entity intending to carry it out – i.e., the data controller; b) the rules governing data protection responsibilities amongst the actors involved – mainly: data controller(s) and data processor(s); c) rights enjoyed by data subjects with regard to the processing of their personal data; and d) privacy preserving techniques. This section explores and draws requirements for each dimension.

Rules governing the processing of personal data

When it comes to personal data, the GDPR mainly governs the *processing* of data, i.e., actions performed by agents on the data. Pursuant to the GDPR, '**processing**' is a very comprehensive notion which means "any operation or set of operations which is performed on personal data or on sets of personal data […]",[47] from the collection all the way to the destruction of such data. The notion covers operations such as anonymisation, analysis, sharing, making available, trading, etc. In order to be GDPR-compliant, any processing operation needs to a) comply with a set of principles; and b) rely on suitable lawful grounds.

Article 5 GDPR lays down the **principles** guiding the processing: lawfulness; fairness; transparency; purpose limitation; data minimisation; data accuracy; storage limitation; integrity and confidentiality; and accountability. When it comes to *purpose limitation*, it is important to note that this principle requires the controller to clearly specify the purpose for which the data will be processed. This is particularly important in UPCAST to the extent that some of the pilots intend to process personal data that are already available (either publicly or in data sets held by other organisations), and that were therefore already processed for a specific purpose. Controllers are also required to only process the data for the purpose that was specified: the controller is therefore prohibited from trying to achieve a different purpose via the same processing operation(s).

The only exception to the requirement of identifying a new purpose relates to processing whose purpose is compatible with the original one. This compatibility is presumed when data are to be processed for 'archiving purposes in the public interest, scientific or historical research purposes or statistical purposes'.[48] This exception, which is regulated by Article 89 GDPR and is relevant to some of the UPCAST pilots, does however not exempt the controller from relying on a suitable **lawful ground** for processing. The notion of lawful ground refers to a legal basis that is needed to authorise the processing operation or set of operation envisaged by the data controller. Article 6 GDPR provides six lawful grounds:

- Consent given by the data subject.
- Necessity of processing for the performance of a contract.
- Necessity of processing for compliance with a legal obligation.

[44] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504
[45] Article 4(1) GDPR.
[46] Ibid.
[47] Article 4(2) GDPR.
[48] Article 5(1)(b) GDPR.

- Necessity of processing to protect vital interests.
- Necessity of processing for the performance of a public interest task.
- Necessity of processing for a legitimate interest.

Consent is the only legal basis that does not require a necessity assessment. All other legal bases require the controller to ascertain whether the intended processing is necessary to the aim at hand.

Table 42 contains a list of legal requirements relating to the general processing of personal data as well as observations relevant to UPCAST.

Table 42. Requirements relating to the processing of personal data

#	Requirement	Observations
	The data controller shall qualify the data set to ascertain whether (some of) the data it intends to process qualify as personal data.	As far as UPCAST pilot leaders are concerned, it is of course their responsibility to make such verifications before sharing data sets with third parties. When it comes to UPCAST pilot leaders being the receivers of the data sets, this responsibility lies with the data controller in that scenario, i.e., the holder of the data set prior to it being shared with an UPCAST pilot leader.
	The data controller shall verify whether some of the personal data qualify as special categories of personal data.	'Personal' data does not equate 'sensitive' data. Sensitive data (see next section) is a subset of personal data.
	Prior to the processing, the data controller shall clearly identify one or more specific and legitimate purposes for the intended data processing operation(s).	These purposes shall be clearly made explicit and not changed once the processing has begun.
	Further processing of personal data is prohibited in a manner that is incompatible with the initial purpose(s) for which they were processed.	UPCAST pilot leaders cannot rely on the compatibility principle to further process personal data (already processed) for monetisation purposes, because the extraction of monetary value from personal data processing is generally not compatible with e.g., research or other purposes.
	Prior to the processing, the data controller shall identify a suitable lawful ground amongst those provided in Article 6 and, if relevant, Article 9 GDPR.	
Consent		
	If the controller relies on the data subject's consent, it needs to make sure that the consent is (i) freely given; (ii) informed; (iii) specific; and (iv) unambiguous, pursuant to Article 7 GDPR.	The notion of *informed* consent relating to consent as a lawful ground for processing must not be confused with the notion of *informed consent* as an ethical requirement for human participants in medical research projects involving.
	In order for consent to be *freely given*, the data subject needs to be able to refuse to give his/her consent without fearing any negative consequences for such a choice.	This means that the choice shall be genuine and there shall be no detriment flowing, directly or indirectly, to the data subject as a result of his/her decision to deny it.
	In order for consent to be *informed*, data subjects shall be provided with detailed	The recommendation is to provide data subjects with as much information as

	information about the intended purpose, modalities, safeguards surrounding the processing, so that they are able to make an informed decision.	possible, in plain language, as to the intended processing.
	In order for consent to be *specific*, it needs to be limited to the processing operations that could have reasonably been foreseen by the data subject at the time of giving consent.	For scientific research purposes, it is fine not to be able to detail every possible research-based processing operation when asking for consent. It is recommended however to clearly envisage the potential research ambits the data might be used for in the future.
	In order for consent to be *unambiguous*, it shall be beyond doubt that the data subject has actively and willingly provided his/her consent, and that he/she was aware of the act by which he/she gave it.	Consent forms and privacy policies shall therefore ask consent via an autonomous act on the part of the data subject, such as an active statement and/or signature to an explicit form where all the requirements covering consent are complied with.
	Consent of children under a certain age (16, or lower depending on national law, but never lower than 13) is lawful only if it is given by the holder of the children's parental responsibility, pursuant to Article 8 GDPR.	This is potentially relevant to personal data contained in data sets with demographic data processed by MDAT. As for genetic and health data of children relevant to NIS and NHFR, see subsection below.
Legitimate interest		
	The existence of a legitimate interest of the controller needs to be assessed on a case-by-case basis.	This means that there is not a pre-determined list of legitimate interests which controllers can choose from.
	The 'legitimate interest' lawful ground needs to rely on the performance of a necessity test whereby the controller weighs the extent to which the processing is necessary to its legitimate interest against the implications of the processing for the data subjects.	This is potentially relevant to all UPCAST pilot leaders. It is recommended not to use this lawful ground whenever controllers seek to process data for monetisation purposes. This is because this aim is considered to be less important than other, non-profit aims, and hence a higher threshold may be required in the necessity test, account being taken of the consequences for the data subjects.

Ad-hoc rules and requirements apply to the processing of **special categories of personal data**, also referred to as 'sensitive' data. For UPCAST purposes, the main sub-categories of such data are data concerning health and genetic data. Pursuant to Article 9 GDPR, because such data can disclose particularly important dimensions of data subjects, their processing is governed by stronger limitations and safeguards. The GDPR even starts from the basic rule that such data cannot generally be processed *unless* an exemption applies. Amongst the exemptions, the most promising ones for UPCAST are the consent by the data subject and the processing for scientific research purposes.

Table 43 provides a list of legal requirements relating to processing sensitive data.

Table 43. Requirements relating the processing of special categories of personal data

#	Requirement	Observations
	In order for data controllers to rely on the scientific research provision to process sensitive data, they need to either a) be able	This requirement is especially relevant for the NHFR pilot. It means that, if there is no EU or national law authorising processing for

	to rely on an EU or national law to that effect; or b) obtain the explicit consent of the data subjects.	scientific research purposes, then the controller cannot simply rely on Article 9(2)(j) to make the processing lawful. The consent is still needed. However, based on the presumption of compatibility of scientific research purposes, the controller does not need to obtain a new consent if it intends to further process data for which consent was previously obtained in compliance with all the requirements mentioned in Table 42.
	If the data subject has given his/her consent to processing of his/her data for economic purposes, including for sharing in exchange for remuneration for the data subject, the data can be processed for such purposes.	This is especially relevant to the NIS pilot. Unless special legal provisions state otherwise, data subjects have of course the right to consent to processing of their sensitive data and be remunerated accordingly for that. The consent needs to comply with all the requirements mentioned in Table 42. As far as the NHFR pilot is concerned, this means that NHFR cannot process the data for monetisation purposes if it is relying on the initial consent given by data subjects to clinicians, i.e., for research purposes. Only a new consent, explicitly given for monetisation purposes, can lawfully enable processing to that effect.
	The creation of synthetic data from original personal data (e.g., regarding health or genetic data) is a data processing operation that, if akin to pseudonymisation or anonymisation, is presumed to be compatible with the initial purpose(s) for which the personal data were collected.	This is because the generation of synthetic data is an operation designed to reduce the level of data protection intrusion. As such, because its purpose is to enhance GDPR compliance, it is easy to justify this type of processing (similarly to pseudonymisation or anonymisation).
	If the synthetic data generation process was carried out with a method that still permits the re-identification of the data subjects, the synthetic data are to be considered as sensitive data, hence data governed by Article 9 GDPR.	Relevant to NHRF pilot. This is especially the case for synthetic data generated via a one-to-one transformation of the original data set containing personal data. In this case, synthetic data would be considered at best 'pseudonymous data', i.e., still personal data pursuant to the GDPR (Article 4(5)).
	If the synthetic data were generated with a method that does not permit re-identification of the data subject of the initial personal data, the synthetic data are considered as anonymous data and are not governed by data protection law.	In any case, however, because the law on synthetic data is still in its infancy, it is highly recommended that synthetic data are generated and handled with as many technical safeguards as possible taking into account the desired level of utility for further analysis/processing.

Rules governing data protection responsibilities

The GDPR assigns various responsibilities regarding compliance with data protection law. The two main roles are the data controller and the data processor.

The **data controller** is defined as the "natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing

of personal data".[49] The controller is therefore the entity that decides the "why" and "how" of data processing,[50] and as such holds the main responsibilities when it comes to ensuring that personal data processing complies with the law. When two or more entities jointly determine the purposes and means of the processing they qualify as **joint controllers**.[51] The Court of Justice of the EU in its case law has developed the concept of controllership and joint controllership.[52]

The **data processor** is defined as the "natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller".[53] The processor acts merely as an executor or else, if it determines to some extent the purpose and means outside of the instructions by the controller, it becomes a separate controller itself.

The attribution of controllership and the ability to act as a data processor is key to the UPCAST platform. Depending on the role played by the entity deploying the UPCAST plugins, it might qualify as separate controller for the data processing operations (sharing, trading, etc.) alongside individual organisations; or as joint controller for some of these operations; or as mere processor insofar as the plugins act as technical infrastructure for enabling the processing. Requirements for data protection responsibilities are further described in Table 44.

Table 44. Requirements relating to data protection responsibilities

#	Requirement	Observations
	For each data processing operation, there has to be one or more data controller(s), including joint controllers, who are responsible to comply with data protection law.	There always has to be an entity (the controller(s) which the data subject can contact and potentially hold responsible for failure to comply with data protection law.
	The data controller is the entity that decides on the purposes and means of the data processing.	These two conditions are cumulative.
	If two or more entities decide autonomously the purposes and means of a data processing (set of) operation(s), then they all qualify as separate data controllers.	This would be the case in UPCAST if two distinct organisations were to use the UPCAST plugins to share the data with two recipients. The processing operations would be separate and so would the determinations of purposes and means of the processing.
	If the purposes and means of the same data processing (set of) operation(s) are decided jointly by two or more entities, even to a different degree, then the entities qualify as joint controllers.	This would likely be the case to the extent that one organisation is willing to share its data via the UPCAST platform; and UPCAST also processes those data for the purpose of enabling, e.g., negotiation and pricing.
	Joint controllers shall determine their respective responsibilities for compliance with the obligations under the GDPR in a transparent manner.	It is recommended to draw up agreements between joint controllers to clearly set out the respective responsibilities.

[49] Article 4(7) GDPR.
[50] Article 29 Working Party, 'Opinion 1/2010 on the Concepts of "Controller" and "Processor", 00264/10/EN WP 169'.
[51] Article 26 GDPR.
[52] 5 June 2018, C-210/16, ECLI:EU:C:2018:388 ('*Wirtschaftsakademie* case'); 10 July 2018, C-25/17, ECLI:EU:C:2018:551 ('*Jehovan todistajat* case'); 29 July 2019, C-40/17, ECLI:EU:C:2019:629 ('*Fashion ID* case').
[53] Article 4(8) GDPR.

	It is the data controller's responsibility to make sure that the processor(s) envisaged provide all guarantees to ensure compliance with data protection law during the processing.	N/A
	The data processor(s) cannot autonomously engage other processors while processing the personal data on behalf of the controller(s).	N/A
	The data controller(s) and the data processor(s) shall enter into a binding agreement concerning the purpose, nature and modalities of the data processing, pursuant to Article 28(3) GDPR.	N/A

Data subject rights

The GDPR (Articles 12-22) attributes various **rights** to data subjects whose data are processed by controllers. It is key for UPCAST pilot leaders, the UPCAST platforms and the organisations intending to use it in the future, to create the conditions for data subjects to enjoy their rights already in the privacy policies.

Table 45 lists legal requirements relating to the various rights of data subjects.

Table 45. Requirements relating to data subject rights

#	Requirement	Observations
	Data subjects shall always be provided with the identity and contact details of the data controller and, where applicable, the Data Protection Officer (DPO).	This information shall be clearly mentioned in the terms & conditions and in the privacy policies of data controllers.
	Data subjects shall always be informed about the purposes and lawful ground of the processing.	Same as above.
	Data subjects shall always be informed about the categories of personal data processed.	Same as above.
	Data subjects shall be informed if the controller intends to transfer personal data to a recipient in a third country.	Same as above. Particularly relevant to NIS insofar as it envisages the transfer of health-related data to third countries for processing. The privacy policy shall be internally coherent and make sure to ask for the data subjects' consent to carry out such transfers. The data protection conditions of the countries where data may be transferred shall be laid down for data subjects to exercise an informed choice.
	Data subjects shall always be informed about the period for which their data will be stored.	Same as above.
	Data subjects shall always be informed of their right to access, rectification, erasure, restriction and objection.	Same as above.
	Data subjects shall always be informed of their right to withdraw consent, when consent is the lawful ground for processing.	Same as above.
	Data subjects shall always be informed of their right to lodge a complaint with a supervisory authority.	Same as above.

	Data subjects shall always be informed about the source of the personal data processed.	Same as above.
	Data subjects shall always be informed if their data are going to be subject to automated decision-making and, if so, about the logic of the automation and the consequences for data subjects.	Same as above.
	Data subjects shall always be informed if the controller intends to further process the personal data for a new purpose.	The new purpose needs however to be compatible with the initial one. A controller cannot simply avoid this obligation by informing the data subject.
	The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed.	Same as above.
	The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her.	Same as above.
	The data subject has the right to request the erasure of his/her data under the conditions of Article 17(1) GDPR.	Same as above. This right cannot be excluded for the original data provided by the data subject.
	The right to erasure shall not apply when the processing is, inter alia, necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, in so far as that right is likely to render impossible or seriously impair the achievement of the objectives of that processing.	This is particularly relevant to UPCAST insofar as data subjects who initially gave consent to the processing of their genomic (NHFR) data may withdraw their consent. Insofar as NHFR processes data for scientific research purposes, it may oppose the data subject's wish to withdraw consent and have the data erased, but only if it demonstrates that such erasure would seriously impair the research.
	The data subject shall have the right to obtain from the controller the restriction of processing under the conditions of Article 19(1) GDPR.	Same as above.
	Where personal data are processed for scientific or historical research purposes or statistical purposes pursuant to Article 89(1), the data subject, on grounds relating to his or her particular situation, shall have the right to object to processing of personal data concerning him or her, unless the processing is necessary for the performance of a task carried out for reasons of public interest.	Same as above.

Privacy preserving techniques

The protection of privacy and of personal data is not a black-and-white concept. The degree of protection is always a function of the nature of the processing and the techniques deployed by the controller and processor. Given the primary objective of ensuring that personal data are as protected as possible during processing, the GDPR recognises and encourages the use of various privacy preserving techniques to enhance data protection.

Not all techniques have the same practical and legal effects on the data and on data subjects. Some techniques – referred to as anonymisation techniques – have the potential to transform the data so as to make them non-personal data; other techniques, that do not strip the data

from their identifying constituents, are helpful to enhance compliance with the GDPR objectives while keeping the data as 'personal'.

Table 46 lists relevant legal requirements relating to privacy preserving techniques.

Table 46. Requirements relating to privacy preserving techniques

#	Requirement	Observations
	Anonymisation, pseudonymisation and encryption are data processing operations that require a lawful ground and a purpose but are generally considered to be compatible with the initial purpose for processing.	This is because they both help achieve the objectives of EU data protection law.
	Pseudonymisation and encryption are operations that increase the protection of personal data but do not anonymise them. Pseudonymised and encrypted data are still personal data.	These techniques are considered by the GDPR as privacy enhancing techniques and are amongst the safeguards that controllers can add to mitigate the data protection risks of a given operation. However, the GDPR continues to apply to such data.
	EU data protection law does not apply to anonymous data.	The notion of 'anonymous data' does not necessarily correspond to that of 'anonymised data'. In particular, the latter does not guarantee that the data are also anonymous in all cases.
	Anonymised data can be considered anonymous if, account being taken of the state of the art of technology, it would require more than reasonable efforts to re-identify the natural persons the original data refer to.	This means that in order to render the data anonymous, data controllers shall carefully consider the most suitable anonymisation technique available and the degree of anonymisation required to reasonably exclude re-identification.

Data governance and digital regulation

The EU is currently in the process of adopting horizontal legislation aimed at regulating the sharing, data access and use of data within the European Union (see the proposed Data Act). Data governance mechanisms have also been discussed and adopted (see the Data Governance Act regulation). The EU has also adopted platform regulations in the form of the Digital Markets Act and the Digital Services Act. A proposed AI Act also aims to regulate the provision and use of certain AI systems. But in parallel to these horizontal texts, sectoral specific legislation may also apply. The following subsection will explore these different (prospective or adopted) horizontal legal instruments within the context of the UPCAST project.

Platform regulations

The EU adopted two sets of texts relating to the regulation of platforms or digital intermediaries.

The Digital Markets Act

The digital markets act (DMA) aims to regulate the proper functioning of the internal market by **laying down harmonised rules ensuring for all businesses, contestable and fair markets in the digital sector across the EU where gatekeepers are present,** to the benefit of all business users and end users (Art. 1(1)). The regulation **applies to core platform services provided or offered by gatekeepers to business users established in the Union or end users established or located in the Union**, irrespective of the place of establishment or residence of the gatekeepers and irrespective of the law otherwise applicable to the provision of service (Art. 1(2)). This piece of legislation **regulates large technology platforms** which are

designated as 'gatekeepers'. The DMA provides rules defining and prohibiting perceived unfair business practices by such large online platforms between European businesses and consumers. The DMA will apply in parallel with other national and EU competition rules.
The Digital Markets Act will, *inter alia*, oblige certain providers of core platform services (gatekeepers) to provide enhanced portability of data generated by business and end users. Gatekeepers are provided with a set of prohibitions (Art. 5(2) to 6(13)) and obligations (Art 5.(4) to Art. 14).
In the context of UPCAST, the providers of intermediation services (platforms providing data sharing services) should not qualify as gatekeepers under the DMA.

The Digital Services Act
The Digital Services Act (DSA) **provides consumers with a new set of rules aimed at protecting them**. The regulation sets harmonised rules for a safe, predictable and trusted online environment that facilitates innovation and in which fundamental rights enshrined in the Charter, including consumer protection, are protected (Art. 1(1)). The regulation also **lays down harmonised rules regarding the provision of intermediary services** "such as it establishes a framework for the conditional exemption from liability of providers of intermediary services, rules on specific due diligence obligations tailored to certain specific categories of providers of intermediary services, and rules on the implementation and enforcement of this Regulation" (Art 1(2)).
In the context of UPCAST, the intermediation services offered by data sharing platforms may have to be assessed under the obligations stipulated in the DSA. In particular, UPCAST may qualify as a hosting service and, within this concept, as an online platform, thereby being subject to the DSA requirements relating to transparency, content moderation, recommender systems.

Current or proposed data regulations
Questions about the legal nature of data and its possible commodification have been discussed in legal literature.[1] The increasing reliance on data in many sectors of the society has exacerbated the need for a better understanding of the legal implications of data sharing and exchanges. The Commission therefore proposed two sets of texts in relation to data sharing: one on the access and use of data between several types of actors (the Data Act proposal or 'DA') and one other on the governance mechanisms of data sharing (the Data Governance Act or 'DGA'). It must be stressed that the Data Act is a proposal currently being discussed at the EU level. Its exact content can and will change depending on the amendments that are carried out by the European Parliament and the Council. However, the Data Governance Act has been adopted.

The Data Act proposal
The Data Act proposes to establish a horizontal set of rules applicable to all sectors regarding the rights to use data. It must however be noted that data rights and obligations have also been regulated at the sectoral level (see *infra: EU electricity regulation, different type approval regulations, etc.*). Nonetheless, the Data Act covers data flows from business-to-business, business-to-government, government-to-business and government-to-government. It aims to facilitate the access and the use of data by consumers and businesses, enable the use of data held by governmental institutions by other public bodies or businesses, to provide safeguards against 'unlawful data' transfers without notification by service providers and increase the interoperability standards for data re-use.

We will be looking at the Council version of the proposal:

Table 47. Council version of the Data Act proposal and observations relevant for UPCAST.

(Art. 1) The Regulation lays down harmonised rules on:	**Observations**
The making data generated by the use of a product or related service available to the user of that product or service	This Act therefore concerns **data generated by the use of a product or related service**. Any data not generated by a product or related service falls outside the scope of this Act. In practice, this seems to indicate (subject to further research and analysis of use cases and pilots) that it would not necessarily apply in all UPCAST use cases.
On making data available by data holders to data recipients	
On making data available to the Commission, the European Central Bank or Union institutions, agencies or bodies, where there is an exceptional need, for the performance of a task carried out in the public interest	
on facilitating switching between data processing services, on introducing safeguards against unlawful third-party access to non-personal data	This could apply to UPCAST as data processing services will be involved.
on providing for the development of interoperability standards for data to be accessed, transferred and used	Again, this could apply to UPCAST

The Data Act proposal applies to different parties which include Data Holders and Data Recipients (Art. 2). Public bodies are also concerned by this Act.

Some of the proposed Data Act requirements are summarised in Table 48 below[1]:

Table 48. Data Act requirements and observations on how they relate to UPCAST.

#	**Legal provisions / requirements**	**Observations**
	Chapter II: rights of users to use data of connected products and related services	
	(Art. 3) **Obligation to make data generated by the use of products or related services accessible to the user**:	**The data holders must make data generated by their use of products or related services available.** **However, it seems rather unclear whether the scope of this act could be extended to data not related to a product or related service. If such was the case, then the Act would apply to those providing data and those using the data.**
	(Art. 4) The right of users to access and use data (+metadata) **generated by the user of a product or related services**. Such access is free of charge, easily and securely, in a structured and machine-readable way. Prohibition of narrowing access rights (through agreements between data holder and user = not binding). Trade secrets can be disclosed only if the data holder and user take measures to preserve their confidentiality. Under exceptional circumstances, the data holder can refuse request for access to data covered by trade secrets.	If data users request access and use data generated by the user of a product or related services, the data holders must provide the data free of charge without delay. **This could have an impact on UPCAST where certain categories of data would be subject to no commodification.**

	Prohibition for the user to use data in order to develop a competing product + prohibition to share data with 3rd party for the same purpose. + specific rules regarding personal data generated by the use of a product or related service when user is not the data subject. The data holder can only use any non-personal data generated by the use of a product or related service on the basis of a contractual agreement with the user.	**Moreover, trade secrets can be disclosed only if all measures have been taken to preserve their confidentiality.** The data holders who wish to use non personal data generated by the use of a product or related service can do so if they have a contractual agreement with the user.
	(Art. 5) <u>Right of the user to share data with third parties:</u> Data holder must make data + metadata available to 3 party upon request of the user without undue delay, free of charge to the user, of the same quality as is available to the data holder, easily, securely, in a structured, commonly used and machine-readable format and, where applicable, continuously and in real-time. Gatekeepers (under DMA) are not eligible 3rd parties. Additional rules or prohibitions related to the use of non personal data in order to derive insights, personal data, trade secrets, are contained in this Article.	The data holders must make available the data to a third party upon request by a user or a party acting on behalf of a user. This shall be done without undue delay and free of charge to the user. In the context of UPCAST, and if the Act applies, such a user's rights may have implications on the implementation of the platforms.
	(Art. 6) <u>Obligations of third parties receiving data at the request of the user</u>: a 3rd party shall process the data only for the purposes and under the conditions agreed with the user + subject to the rights of the data subject insofar as personal data are concerned + shall delete the data when they are no longer necessary for the agreed purpose. Prohibition for the third party to: (a) coerce, deceive or manipulate the user or the data subject where the data subject is not the user (…)(b) use the data it receives for the profiling of natural persons (c) make the data it receives available to other third party parties unless this is necessary to provide the service requested by the user + other third parties take all necessary measures agreed between the data holder and the third party to preserve the confidentiality of trade secrets; (d) make the data it receives available to an undertaking designated as a gatekeeper	
	(Art. 7) <u>Scope of business-to-consumer and Business-to - Business data sharing obligations</u>: the obligations do not apply to data generated by the use of products manufactured or related services provided by enterprises that qualify as micro or small enterprises. The same applies to enterprises that qualify as medium-sized under certain conditions. Any contractual term which, to the detriment of the user, excludes the application of, derogates from or varies the effect of the user's rights shall not be binding on the user.	
	Chapter III: Horizontal obligations for data holders legally obliged to make data available in business-to-business relations	
	(Art. 8) <u>Conditions under which data holders make data available to data recipients:</u> In business-to-business relations, the data holder must make data available to a data recipient under fair, reasonable, non-discriminatory terms + transparent way + shall agree with data recipient on the terms for making the data available. Prohibition for data holder to discriminate between comparable categories of data recipients + prohibition to make data available on exclusive basis (except is request by user). Data holders + data recipients not required to provide information beyond what is necessary to verify compliance with the agreed contractual terms	If this Act were to apply to UPCAST, then the data holders wishing to make data available to data recipients would have to comply with these principles.

	Unless the law provides otherwise, an obligation to make data available to a data recipient shall not oblige the disclosure of trade secrets	
	(Art. 9) <u>Compensation for making data available:</u> such compensation between data holder and data recipient (in B to B) shall be reasonable and may include a margin. If the data recipient is a micro, small or medium enterprise any compensation agreed shall not exceed the costs set out in paragraph 1a(a). Laws may exclude compensation or provide for lower compensation. The data holder shall provide the data recipient with information for the basis of the calculation of the compensation	The notion of 'reasonable' compensation is open to interpretation. But the Act states that compensation may include a margin. However, the provision states the elements to take into account when determining the compensation (see Art 9 (1 a).
	(Art. 11) <u>Technical protection measures taken by data holder and provisions on unauthorized use or disclosure of data:</u> Data holder may apply appropriate technical protection measures, including smart contracts, to prevent unauthorised access to the data + ensure compliance with Articles 5, 6, 9 and 10 + the agreed contractual terms for making data available. Prohibition to discriminate between data recipients or to hinder the user's right to provide data to third parties based on technical protection measures. Under certain conditions/situations, the data holder may (a) request the data recipient to erase the data made available by the data holder and any copies (b) request the data recipient to end the production, offering, placing on the market or use of goods, derivative data or services produced on the basis of knowledge obtained through such data, or the importation, export or storage of infringing goods for those purposes, and destroy any infringing goods., (c) seek compensation from the data recipient.	Here the application of **smart contracts** to protect access to data and ensure the compliance with agreed contractual terms is a possible technical measure. This could have implications for UPCAST as such smart contract technologies could be used.
	(Art. 12) <u>Scope of obligations for data holders legally obliged to make data available</u>:this chapter applies to B to B relations. A data holder is obliged to make data available to a data recipient. Any contractual term in a data sharing agreement which, to the detriment of one party or to the detriment of the user, excludes the application of this Chapter, derogates from it, or varies its effect, shall not be binding on that party	
	Chapter IV: Unfair terms related to data access and use	
	(Art. 13) <u>Unfair contractual terms unilaterally imposed on a enterprise</u>: a contractual term (concerning the access to and use of data or the liability + remedies for the breach or the termination of data related obligations) which has been unilaterally imposed by an enterprise on another enterprise, shall not be binding if it is unfair. This article lists the situations where a contractual term is unfair or presumed to be unfair.This Article does not apply to contractual terms defining the main subject matter of the contract nor to the adequacy of the price, as against the data supplied in exchange. Prohibition for the parties to exclude the application of this Article, derogate from it, or vary its effects	In the context of UPCAST, an unfair contractual term will not be binding on the enterprise on which it is imposed. This article further substantiates the notion of 'unfairness'.
	Chapter V: Making data available to public sector bodies, the Commission, the European central bank or Union bodies based on exceptional need	
	(Art. 14) Obligation to make data available based on exceptional need	
	(Art. 15) Exceptional need to use data	
	(Art 18) Compliance with requests for data	
	(Art 19) <u>Obligations of public sector bodies and the Commission, the European Central Bank and Union bodies</u>: such parties shall: (a) not use the data incompatible with the purpose for which they were requested; (b) have	

	implemented technical and organisational measures preserving the confidentiality and integrity of the requested data; (c) erase the data as soon as they are no longer necessary for the stated purpose The disclosure of trade secrets to such parties are only required to the extent that it is strictly necessary to achieve the purpose of the request. In such a case, these parties must take appropriate technical + organisational measures to preserve the confidentiality of those trade secrets.	
	Chapter VI: switching between data processing services	
	(Art. 23) **Removing obstacles to effective switching between providers of data processing services**: providers of such a service must allow their customers to switch to another data processing service, covering the same service type, which is provided by a different service provider. Prohibition for them to pose obstacles which inhibit customers from: (a) terminating; (b) concluding new contractual agreements with a different provider; (c) porting its data and metadata created by the customer; (d) maintaining functional equivalence of the service in the IT-environment of the different provider or providers of data processing services covering the same service type.	If the Act applies to UPCAST, this provision could have an impact. It states that providers of a data processing service (could the UPCAST platforms and plugins have such a function?) shall take measures to ensure that customers of their service can switch to another data processing service, covering the same service type, which is provided by a different service provider. Data service providers shall remove commercial, technical, contractual, and organisational obstacles.
	(Art 23a) Scope of the technical switching obligations: The responsibilities of data processing service providers shall only apply to the services, contractual agreements or commercial practices provided by the original provider	
	(Art. 24) Contractual terms concerning the switching between providers of data processing services: the rights of the customer and the obligations of the provider of a data processing service in relation to switching between providers of such services or to an on-premise system shall be clearly set out in a written contract. The paragraph then lists the minimum elements the contract shall have. There are certain temporal rules to be respected.	In the context of UPCAST, the switching of services right may have to be enforced in practice.
	(Art 24a) **Contractual transparency obligations on international access and transfer**: providers of data processing services must provide and keep updated on their websites: (a) information regarding the jurisdiction to which physical location of all the IT infrastructure deployed for data processing of their individual services is subject; (b) a general description of the technical, legal and organisational + contractual measures adopted in order to prevent governmental access to non-personal data held in the Union where such transfer or access would create a conflict with Union law or the national law.	
	(Art 26) Technical aspects of switching: this article provides for rules regarding switching.	
	Chapter VII: Unlawful international governmental access and transfer of non-personal data	
	Chapter VIII: Interoperability	
	(Art. 28) Essential requirements regarding interoperability: Operators within data spaces shall comply with essential requirements: (a) the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty shall be sufficiently described, where applicable, in machine-readable format; (b) the data structures, data formats, vocabularies, classification schemes, taxonomies	Such requirements could have to be complied with in the context of UPCAST.

	and code lists, where available, shall be described in a publicly available and consistent manner; (c) the technical means to access the data, such as application programming interfaces, and their terms of use and quality of service shall be sufficiently described to enable automatic access and transmission of data between parties, including continuously, in bulk download or in real-time in a machine-readable format. (d) where applicable, the means to enable the interoperability of tools for automating the execution of data sharing agreements, such as smart contracts. Operators of within data spaces that meet the harmonised standards whose references have been published in the Official Journal of the European Union shall be presumed to be in conformity with the essential requirements **(Art 28a) Interoperability for the purposes of in-parallel use of data processing services** **(Art 29) Interoperability for data processing services**: open interoperability specifications and harmonised standards for the interoperability of data processing services shall: (a) be performance oriented; (b) enhance portability of digital assets between different data processing services (c) guarantee ensure, where technically feasible, functional equivalence between different data processing services Open interoperability specifications and European harmonised standards for the interoperability of data processing services shall adequately address: (a) the cloud interoperability aspects; (b) the cloud data portability aspects;(c) the cloud application aspects **(Art 30) Essential requirements regarding smart contracts for data sharing:** the vendor of an application using smart contracts or, in the absence thereof, the person whose trade, business or profession involves the deployment of smart contracts for others in the context of an agreement to make data available, shall comply with the following essential requirements: (a) robustness; (b) safe termination and interruption; (c) data archiving and continuity;(d) access control: The vendor (or person referred to in previous paragraph) of a smart contract shall perform a conformity assessment with a view to fulfilling the essential requirements under paragraph 1 and, on the fulfilment of the requirements, issue an EU declaration of conformity. The vendor of an application using smart contracts (or the person referred to in the first paragraph) shall be responsible for compliance with the requirements under paragraph 1. A smart contract that meets the harmonised standards or the relevant parts thereof drawn up and the references of which have been published in the Official Journal of the European Union shall be presumed to be in conformity with the essential requirements under paragraph 1	This could have an impact on UPCAST as the plugins may rely on smart contract technologies. In such a situation, the vendor of these applications/plugins will have to comply with the obligations set under Art. 30.

In short, it is **currently unclear how this proposed Act would apply in the context of UPCAST,** and what effects it may have on the implementation of the project. The proposal applies to the re-use of data generated by products (tangible items) or related services. But **the current proposal could evolve pending further amendments** which could **extend its scope of application,** and particularly that of the type of data that can be re-used. Further assessment will be required.

The Data Governance Act

The Data Governance Act ('DGA') complements the Data Act proposal and seeks to **facilitate the "voluntary sharing of data by individuals and businesses and harmonises conditions for the use of certain public sector data** without altering material rights on the data or established data access and usage rights".[2] The DGA also complements the Open Data Directive. As noted in a report, it "aims to foster the re-use of such data despite the existence of entitlements of third parties (such as intellectual property rights, confidentiality obligations or data protection obligations), subject to a specific legal regime".[3] The DGA regulation **lays down rules** regarding the '(a) **conditions for the re-use, within the Union, of certain categories of data held by public sector bodies**; (b) **a notification and supervisory framework for the provision of data intermediation services**; (c) a framework for voluntary registration of entities which collect and process data made available for altruistic purposes; and (d) a framework for the establishment of a European Data Innovation Board.' (Art. 1).

Data intermediaries are referred to in the DGA (Art 10).

Some of the requirements set in the DGA are summarised in Table 49 below:

Table 49. Requirements from the Data Goverance Act and observations on how they relate to UPCAST.

#	**Legal provision / Requirements**	**Observations**
	Chapter II: **Re-use of certain categories of protected data held by public sector bodies**	
	(Art. 3) <u>Categories of data:</u> **This Chapter applies to data held by public sector bodies which are protected on grounds of:** (a) commercial confidentiality, including business, professional and company secrets; (b) statistical confidentiality; (c) the protection of intellectual property rights of third parties; or (d) the protection of personal data, insofar as such data fall outside the scope of Directive (EU) 2019/1024. 2. **This Chapter does not apply to:** (a) data held by public undertakings; (b) data held by public service broadcasters and their subsidiaries (c) data held by cultural establishments and educational establishments; (d) data held by public sector bodies which are protected for reasons of public security, defence or national security; or (e) data the supply of which is an activity falling outside the scope of the public task of the public sector bodies	Under UPCAST and in the context of exchange of data held by public sector bodies, such entities could be subject to the obligations set in this chapter
	(Art. 4) **<u>Prohibition of exclusive arrangements + exceptions</u>**	Same as above
	(Art. 5) <u>Conditions for re-use:</u> The article provides a long list of conditions related to the re-use of data. For instance, competent public sector bodies shall make publicly available the conditions for allowing such re-use and the procedure to request the re-use. Such conditions for re-use shall be non-discriminatory, transparent, proportionate and objectively justified with regard to the categories of data and the purposes of re-use and the nature of the data for which re-use is allowed. *Those conditions shall not be used to restrict competition.* Public sector bodies shall ensure that the protected nature of data is preserved. They may provide for some	Same as above

	requirements (see article for list of proposed requirements). Unless national law provides otherwise, the public sector body shall make the re-use of data conditional on the adherence by the re-user to a confidentiality obligation that prohibits the disclosure of any information that jeopardises the rights and interests of third parties that the re-user may have acquired despite the safeguards put in place. Prohibition for re-users from re-identifying any data subject to whom the data relates + shall take technical and operational measures to prevent re-identification and to notify any data breach resulting in the re-identification of the data subjects. Re-use of data shall be allowed only in compliance with intellectual property rights. Where requested data is confidential, the public sector bodies shall ensure that such data is not disclosed as a result of allowing re-use (unless such re-use is allowed). Other rules are detailed in the Article.	
	(Art 6) <u>Fees</u>: public sector bodies which allow re-use of the categories of data referred to in Article 3(1) may charge fees. Any charged fees shall be transparent, non-discriminatory, proportionate and objectively justified and shall not restrict competition.	
	(Art 9) <u>Procedure for request for re use</u>: This article details the request procedure for re-use.	
	Chapter III: **Requirements applicable to data sharing services**	
	(Art. 10) <u>Data intermediation services:</u> The provision of the following data intermediation services shall comply with Article 12 and subject to a notification procedure: (a) intermediation services between data holders and potential data users, including making available the technical or other means to enable such services; those services may include bilateral or multilateral exchanges of data or the creation of platforms or databases enabling the exchange or joint use of data, as well as the establishment of other specific infrastructure for the interconnection of data holders with data users; (b) intermediation services between data subjects that seek to make their personal data available or natural persons that seek to make non-personal data available, and potential data users, including making available the technical or other means to enable such services, and in particular enabling the exercise of the data subjects' rights provided in Regulation (EU) 2016/679; (c) services of data cooperatives.	The data exchange platforms/plugins in UPCAST could qualify as data sharing services that would have to comply with the requirements set in Art. 11 and 12.
	(Art. 11) **Notification of data sharing service providers**	This first requirement obliges data sharing service providers to notify their activities
	(Art. 12) **Conditions for providing data intermediation services:** this article provides a long list of conditions. These include, for instance, the following conditions: * the provider shall not use the data other than to put them at the disposal of data users and shall provide data intermediation services through a separate legal person; *the commercial terms, including pricing, for the provision of data intermediation services to a data holder or data user shall not be dependent upon whether the data holder or data user uses other services provided by the same data intermediation services provider or by	

	a related entity, and if so to what degree the data holder or data user uses such other services; *the data collected with respect to any activity of a natural or legal person for the purpose of the provision of the data intermediation service (eg, date, time, geolocation data, duration of activity and connections to other natural or legal persons established by the person who uses the data intermediation service) shall be used only for the development of that data intermediation service (eg, for the detection of fraud or cybersecurity), and shall be made available to the data holders upon request; *the provider shall facilitate the exchange of the data in the format in which it receives it from a data subject or a data holder, shall convert the data into specific formats only to enhance interoperability within and across sectors or if requested by the data user or where mandated by Union law or to ensure harmonisation with international or European data standards + shall offer an opt-out possibility regarding those conversions to data subjects or data holders, unless the conversion is mandated by Union law; *the data intermediation services provider shall ensure that the procedure for access to its service is fair, transparent and non-discriminatory for data subjects + data holders + data users, (including prices and terms of service); *the provider shall put in place adequate technical, legal and organisational measures in order to prevent the transfer of or access to non-personal data; *the provider shall take necessary measures to ensure an appropriate level of security for the storage, processing and transmission of non-personal data + shall further ensure the highest level of security for the storage and transmission of competitively sensitive information;. The full list of conditions can be found in the Article.	

Proposed AI regulation

On the AI front, the proposed AI Act[54] establishes a **risk-based framework that aims to regulate the placing on the market, the putting into service and the use of (some) AI systems** in the European Union. It provides specific requirements for high-risk AI systems and prohibits certain types of AI practices. Its adoption may have legal implications for the UPCAST project where and when machine learning mechanisms are employed.

A system is high risk when: a) it is intended to be used as a safety component of a product or is itself a product covered by (certain) Union harmonisation legislation; AND b) the product whose safety component is an AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment prior to is putting into the market or into service. AI systems referred to in Annex III of the proposal are also considered high-risk.

In the UPCAST project, AI seems to be employed in connection with the smart contracts used to formalise data sharing agreements. It does not seem to have a safety function. Therefore, pending further analysis, it may not be qualified as high-risk, depending on the use cases. However, AI used in the context of UPCAST could possibly be qualified a non-high risk AI system which would nonetheless have to protect the fundamental rights of EU citizens.

[54] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Moreover, certain transparency requirements may have to be met (but only for those AI systems that interact with people).

Other legislation and sectoral legal instruments

Other legislation may apply in the context of data sharing. For instance, the aforementioned **Open Data Directive** ((EU) 2019/1024, 20 June 2019) establishes a set of minimum rules governing the re-use and the practical arrangements for facilitating the re-use of: (a) existing documents held by public sector bodies of the Member States; (b) existing documents held by public undertakings. This Directive applies without prejudice to provisions relating to the protection of personal data such as the GDPR. Moreover, the obligations imposed under the Directive shall only apply as long as they are compatible with the provisions of international agreements on the protection of intellectual property rights. Finally, the right for the maker of a database provided for in the Directive shall not be exercised by public sector bodies to prevent the re-use of documents or to restrict re-use beyond the limits set by this Directive, etc. The Directive states (Art. 3) that "Member States shall ensure that documents to which this Directive applies in accordance with Article 1 shall be re-usable for commercial or non-commercial purposes in accordance with Chapters III and IV". The Directive provides public sector bodies with certain obligations regarding the processing of request for re-use, the conditions for re-use, including provisions relating to the available format (Art. 5) and principles governing charging (Art 6).

The **Database directive** 96/9/EC (data base rights) may also be applicable in certain situations since UPCAST relies on data sets. In particular, it is recommended that data providers and the UPCAST platform verify whether any of the data sets made available are covered by *sui generis* database rights within the meaning of the Directive, and, if so, make sure to comply with relevant requirements.

The Free Flow of Non-Personal Data regulation aims at removing barriers to the free movement of non-personal data between different EU countries and IT systems in Europe.

Finally, **copyright and trade secret laws** will continue to apply transversally (some of the above-mentioned texts do refer to trade secrets and IP laws).
The analysis should also consider the application of **sectoral regulations or laws**. For instance, when a medical device is concerned, the Medical Device Regulation (MDR) could apply.

[1] For an analysis of the proposed Data Act, see: Ducuing Charlotte, Margoni Thomas and Schirru Lucas (eds). White Paper on the Data Act Proposal. Data Act White Paper. On the general data frameworks, see: Ducuing Charlotte, Dheu Orian, Aliki Benmayor, RNE Study on data, 2022, p. 16-46.
[2] EC Commission, Data Governance Act explanatory memorandum, 2020, p. 4-5.
[3] Ducuing, Dheu, Benyamor, RNE data study, 2022, p. 42.
[4] Ducuing, Dheu, Benyamor, RNE data study, 2022, p. 44.
[1] The detailed requirements can be found in an annex.

Automated and smart contracts

Automated contracts are characterised by processes that are wholly or in part (e.g., the drafting process, the negotiation, etc.) facilitated by automation. Smart contracts are a network of computer messages composed of conditional statements (i.e. "if condition "x" materialises itself, then effect "y" occurs") executed on e.g., a blockchain or distributed ledger technology (DLT). Contractual clauses are written in the form of code instead of human

language, and hence their consequences are triggered as soon as the condition embedded in the clause materialises itself. Although the EU institutions have recognised the importance of distributed ledger technologies and smart contracts, currently there is no specific EU-level smart contract regulation. Therefore, the relationship between automated & smart contracts and the law needs to be tackled by the law regulating contracts, which includes standard contractual provisions and general principles of contract law.

Table 50 lists legal requirements relating to the legality of automated and smart contracts. Specific requirements are laid down in relation to the legal status of smart contracts under the law of Belgium, France, Italy and Malta, as per the Grant Agreement.

Table 50. Legal requirements relating to the legality of automated and smart contracts

#	Requirement	Observations
Basic requirements		
	Parties intending to conclude a transaction by means of a smart contract have to make sure that such modality of contracting is recognised as legally valid and binding in the chosen jurisdiction.	This requirement follows from the fact that contract law, which governs the conditions for contract validity and effects, is largely Member State-based. Not all Member States apply the same requirements to contracts (including contracts with automated aspects), nor do they all regulate smart contracts in a harmonised way.
	The parties are encouraged to agree on the applicable law, especially if they are established in two different Member States / countries.	This requirement is a recommendation based on the above observation. Explicit indication of the applicable law will greatly enhance legal certainty in case of dispute.
Belgian law		
	Pursuant to Article 1108 of the Belgian Civil Code, an automated contract and a smart contract can qualify as a legal contract when the parties can exercise their independent will as to the subject matter of the agreement and the other contracting parties.	This requirement can be easily satisfied by automated and smart contracts. Any party can access the technical means for executing their will, including after having negotiated it via UPCAST plugins.
	An automated contract and a smart contract can qualify as a valid legal contract when it respects the applicable formal requirements for its validity. For those contracts that only require the exchange of consents between the parties, automated and smart contracts can qualify as valid legal contracts if they allow the parties to freely exchange their wills.	The law prescribes that certain types of contracts are valid or opposable only if accompanied by certain formalities that smart contracts cannot always provide. However, data sharing and data use contracts aren't subject to such formalities.
	Pursuant to Article 5.27 of the Belgian Civil Code, an automated contract and a smart contract can qualify as a valid legal contract if they have a certain subject matter and a lawful cause.	This requirement can be easily satisfied by both (partially) automated contracts and by smart contracts.
	Pursuant to Article 5.69 of the Belgian Civil Code, contracts have the force of law between the parties who concluded them. An automated contract and a smart contract can therefore qualify as a valid legal contract if it provides the parties with the possibility to have its terms enforced before a court.	Automated contracts can in principle satisfy this requirement to the extent that their terms can foresee the resort to arbitration or judicial authorities. Their (partially) automated nature would not hinder such possibilities. This may prove difficult in smart contracts given their immutable and generally non-reversible character.

	Pursuant to Article 5.57, the sanction for failure to comply with the contract validity requirements is the nullity of the contract.	This requirement should be fairly easy to comply with for automated contracts. It may however prove difficult in smart contracts. Nullity implies that the effects of the invalid act (the contract) are assumed as though they never took place. However, the irreversible character of smart contracts sits at odds with this condition.
French law		
	Pursuant to Article 1129 of the French Civil Code, for a contract to be valid, the parties need to a) share their genuine consent; b) have the legal capacity to enter into a contract; and c) base the contract on a certain and lawful subject matter.	Same considerations as above regarding exchange of consents; capacity; and subject matter.
	Articles 1174, 1366 and 1367 of the French Civil Code allow contracts that require the written form to have this requirement satisfied if they are concluded in an electronic form.	This can be the legal basis for considering blockchain-based contracts as in principle compliant with the written form requirement in French law.
	Pursuant to Article 1178 of the French Civil Code, a contract that doesn't comply with its validity requirements shall be considered null and void.	Same considerations as above regarding the challenges posed by nullity of contracts for smart contracts.
	Pursuant to Article 1193 of the French Civil Code, the parties to a contract need to have the possibility to amend the contract upon their wills.	This requirement should be fairly easy to comply with for automated contracts. It may however prove difficult with smart contracts. When a block in the blockchain is validated by a node, thereby triggering the execution of the contractual terms, that block can no longer be amended.
	Pursuant to Article 1195 of the French Civil Code, the parties have the possibility to renegotiate the contract when its terms have become excessively onerous due to an unforeseeable event.	This requirement should be fairly easy to comply with for automated contracts. It may however prove difficult with smart contracts for the same reason as above, i.e., that the contract terms cannot be amended. In certain conditions, this requirement may still be complied with by having the parties negotiate a new contract that trumps the effects of the former.
Italian law		
	Unless otherwise specified, pursuant to the Italian Civil Code, the parties have the freedom to determine the form they wish their contract to take.	This implies that smart contracts can at least be used as the vehicle through which a contract can be concluded and subsequently enforced.
	Pursuant to Article 8-ter(2) of Law No. 12/2019, upon the digital identification of the contracting parties, smart contracts are considered to be legally valid contracts as their registration in the blockchain satisfies the requirement of written form.	This Italian law establishes that the registration of a smart contract on the blockchain satisfies the 'written form' requirement that applies to certain types of contracts. Provided that the smart contract satisfies other applicable contract law requirements, it can be considered as a legally valid contract if the parties digitally identify themselves.
	Pursuant to Article 1331 of the Italian Civil Code, the parties can agree that the contract is	In this case, the smart contract, because of its immutable character, may represent one

	based on an irrevocable proposition by one of the parties.	form through which that party can make its proposition irrevocable.
	Pursuant to Article 1418 of the Italian Civil Code, those contracts that do not comply with their validity requirements are null and void.	Same considerations as above regarding nullity and smart contracts.
Maltese law		
	For a smart contract to be legally recognised and binding, the parties need to have legal capacity to enter into a contract.	Same considerations as above regarding exchange of consents; capacity; and subject matter.
	For a smart contract to be legally recognised and binding, all the parties intending to conclude the contract need to provide and demonstrate their free and undistorted consent to the agreement.	Same considerations as above regarding exchange of consents; capacity; and subject matter.
	For a smart contract to be legally recognised and binding, the contract shall have a subject (tangible or intangible) that is lawful.	Same considerations as above regarding exchange of consents; capacity; and subject matter.
	For a smart contract to be legally recognised and binding, the contract shall have a lawful consideration.	Same considerations as above regarding exchange of consents; capacity; and subject matter.
General aspects of smart contracts		
	The formulation of smart contract shall adhere as much as possible to conditional statements in order to maximise effectiveness and reduce legal uncertainty.	Smart contracts are likely to be ill-suited to replicating the linguistic formulations of typical natural language contracts based on principles and concepts open to human interpretation. The more the statements are conditional (i.e., if a given condition occurs, a given effect is produced), the more the code will be able to capture the essence of the contractual provisions in the smart contract.
	Given the legal fragmentation of contract law and smart contract regulation, it is recommended that the parties conclude a preliminary human language contract.	This is because the case law in the Member States has still not provided clarity as to the legally binding status of smart contracts.
	The parties need to establish the applicable law in case of dispute, including the possibility to resort to a court of arbitration.	Smart contracts need to be encapsulated within general contract law also as far as dispute resolution is concerned. Because blockchain nodes are rarely located in only one country, and especially in light of the typical transactions envisaged in UPCAST, it is strongly recommended that the parties establish ex ante and agree on the competent courts and the range of options regarding arbitration and/or other extra-judicial dispute resolution bodies.

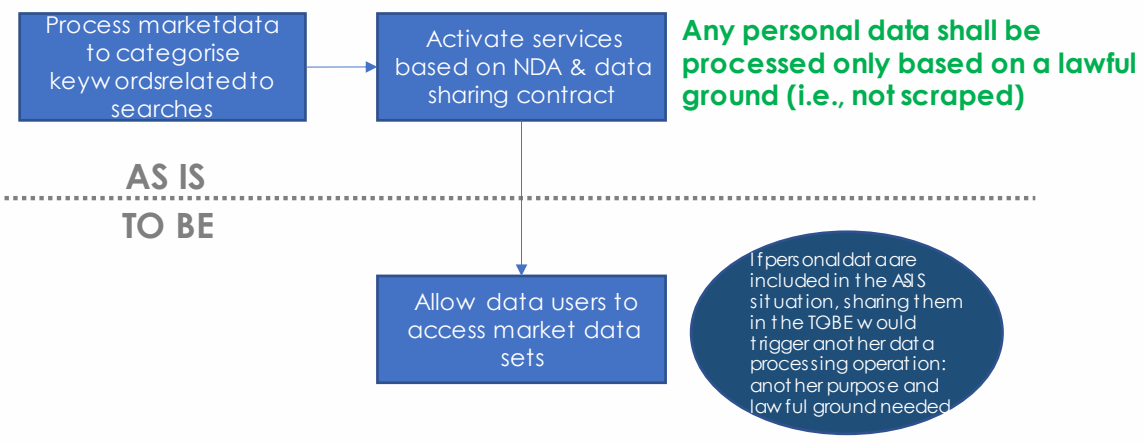
7.2 Legal requirements and UPCAST pilots

This section zooms in on the five UPCAST pilots and provides some considerations and requirements tailored to the specific pilot workflows. The analysis is supported by figures that compare the 'as-is' scenario to the 'to-be' scenario.

Pilot: Digital Marketing Data and Resources (JOT & CACTUS)

This section describes legal requirements for the two digital marketing pilots of JOT and CACTUS. Figure 12 highlights the main legally relevant workflows relating to the two pilots.

Figure 12. Legally relevant workflows relating to the Digital Marketing Data and Resources pilot.



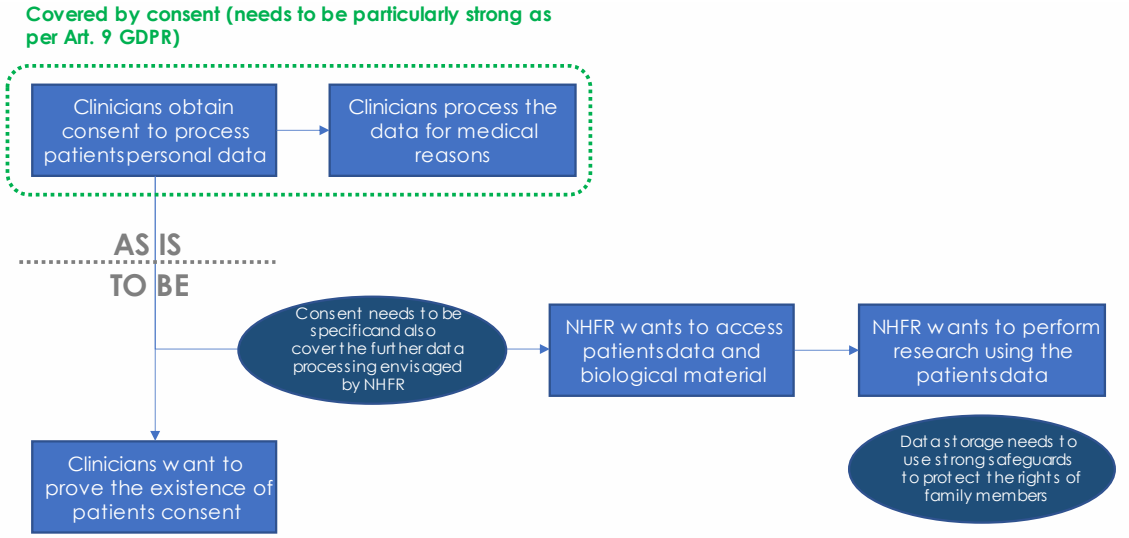
The main considerations relating to these pilots are the following:

- To the extent that the data sets used by JOT and CACTUS in the AS-IS situation contain **personal data** (e.g., customer accounts passwords, email addresses, IP addresses, etc.), they **shall be processed based on a lawful ground** (Article 6 GDPR). Such data cannot be scraped from the Internet despite being publicly available. The legitimate interest lawful ground cannot be used if the purpose is linked to monetisation/commercialisation.
- The legal requirements mentioned in the previous point shall be met also when continuing the same activity in the TO-BE scenario.
- JOT and CACTUS should **continue to sign non-disclosure agreements** with customers for the provision of services based on company data or any data that may be confidential.
- JOT and CACTUS shall ensure that the data contained in the market research data sets **do not include trade secrets** and, if they do, comply with the Trade Secret Directive.
- To the extent that in the TO-BE situation JOT and CACTUS intend to provide data users access to data sets containing personal data, they shall either a) **anonymise those data** prior to the sharing; or b) **only allow access if the processing operation can be based on a lawful ground** pursuant to Article 6 GDPR.
- In the latter case, JOT and CACTUS shall **ask data users to sign a contract setting out data protection policies and requirements for the usage of the personal data** included in the data sets, making sure that the data users agree to **not use those data for other purposes than those specified in the contract** and that are linked to the lawful ground for processing.

Pilot: Biomedical and Genomic Data Sharing (NHRF)

Figure 13 highlights the main legally relevant workflows relating to this pilot.

Figure 13. Legally relevant workflows relating to the Biomedical and Genomic Data Sharing pilot.



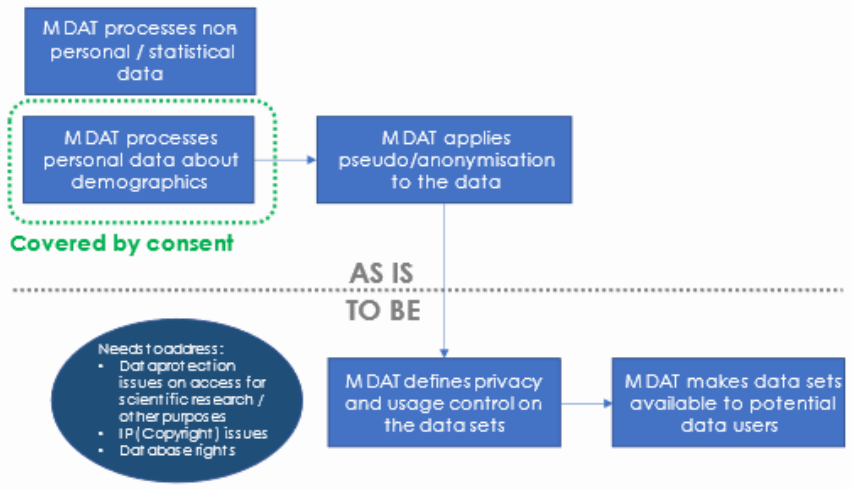
The main considerations relating to this pilot are the following:

- NHRF must engage with clinicians and ensure that, in order to lawfully process health-related data for their purposes, **the consent forms used by clinicians ask an informed consent for further processing** (by NHRF). The more specific NHRF can be when detailing the purposes and scenarios of the further processing, the better; however, the GDPR scientific research regime allows controllers to provide information only as far as reasonably foreseeable (not every single research purpose can be foreseen when requesting consent).
- **Storage** of biologic material and of health-related data **needs to be guaranteed according to the strongest available safeguards** in order to protect the data subjects and the family members potentially identifiable via genomic data.
- Unless covered by a thorough consent from data subjects or properly anonymised, **the genomic data at hand cannot be traded for monetisation purposes** on another legal basis. The scientific research exemption would cease to apply in such a scenario.
- It is recommended that NHRF **only considers the synthetic data generated in laboratories as candidates for data trading and monetisation**. However, NHRF must **ensure that the synthetic data generation process does not allow re-identifying the data subjects**. For instance, the resulting synthetic data must not allow a one-to-one matching with the source personal data.

Pilot: Sharing Public Administration for Climate (MDAT)

Figure 14 highlights the main legally relevant workflows relating to the Public Administration pilot.

Figure 14. Legally relevant workflows relating to the Public Administration pilot.



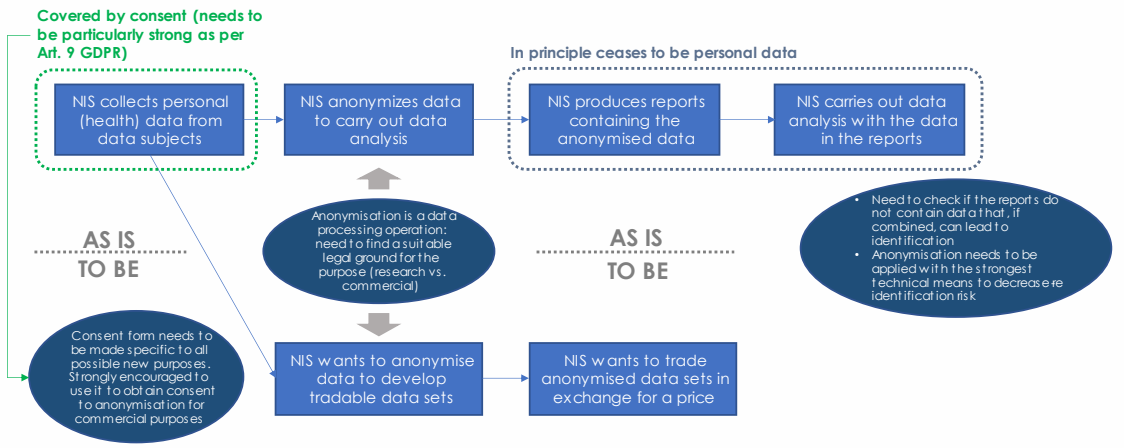
The main considerations relating to this pilot are the following:

- In order to be able to lawfully share data sets containing demographic data, MDAT needs to **continue applying anonymisation** to the personal demographic data.
- To the extent that MDAT does not anonymise these personal data, MDAT needs to rely on a **lawful ground for further processing** those data (i.e., sharing the data with users).
- MDAT needs to make sure that the data aggregated in their data sets do not contain **data covered by intellectual property rights or trade secrets**.

Pilot: Health and Fitness Data Trading (NIS)

Figure 15 highlights the main legally relevant workflows relating to the Health and Fitness Data Trading pilot.

Figure 15. Legally relevant workflows relating to the Health and Fitness Data Trading pilot.



The main considerations relating to this pilot are the following:

- NIS needs to **continue to obtain consent** from data subjects for the processing of their sensitive data, making sure that the consent **complies with the requirements of Article 6 and Article 9 GDPR**.
- As for the 'to-be' scenario, NIS intends to process these sensitive data for profit. As a result, NIS must thoroughly inform data subjects of this intent. **The consent forms need to be exhaustive as regards a) NIS' intent to make profit** from the processing and trading of the data; and b) **NIS' intent to remunerate data subjects** for their contribution to NIS' business model.
- On top of this, as NIS intends to anonymise data before trading them, it is strongly encouraged that the **consent forms also ask data subjects to provide their consent to the use of anonymisation techniques** on their sensitive data.
- NIS needs to **make sure that the envisaged anonymisation techniques** conform to the state of the art and that, taking into account reasonable re-identification efforts, **do not in principle allow an external attacker to re-identify the data subjects**.
- When producing reports based on anonymised data, NIS needs to **make sure that the information contained in the reports does not**, in isolation or in combination, **allow re-identifying the data subjects**.

8 CONCLUSIONS

This report presents the initial requirements definition process conducted in WP1 of the UPCAST project. The objectives of this process have been to derive and properly define requirements to steer the next stages of technical development of UPCAST plugins, requirements related to deployment and demonstration of the plugins as well as legal requirements to ensure compliance with EU laws and regulations related to data management.

The requirements are derived from a three-step approach involving (1) a common understanding of the AS-IS situation (current state and needs of pilots & research topics related to plugins), (2) a TO-BE vision (pilot use cases and user stories and technical discussions on how these could be supported using the plugins), and (3) a requirements elicitation and consolidation process (including verification, filtering and weighing of all initial requirements).

Figure 16 shows an overview of the coverage of pilot requirements per UPCAST plugin.

	Biomedical and Genomic Data Sharing	Public Administration	Health and Fitness	Digital Marketing (JOT)	Digital Marketing (Cactus)
Resource Specification					
Resource Discovery					
Data Processing Workflow					
Privacy and Usage Control					
Valuation and Pricing					
Environmental Impact					
Negotiation and Contracting					
Integration and Exchange					
Safe and Secure Exchange					
Monitoring					

Figure 16. Matrix showing coverage of pilot requirements for each UPCAST plugin.

Some plugins requirements were easier to define than others. This relates to the fact that the functionality of some plugins (e.g., Resource Specification) is more concrete and easier to isolate than for others that are more transversal and offer supporting functionality (in the background) for multiple plugins and usage scenarios (e.g., Monitoring and Environmental Impact).

The initial set of requirements presented in this report will now be subject to further scoping and detailing in WP1 where the main features of the MVP and initial conceptual and technical architecture components (T1.1 and T1.5), final pilot design and functionalities (T1.2), as well as initial input to the vocabulary and data model (T1.3) will be prepared.

9 REFERENCES AND ACRONYMS

9.1 References

A. Daouadji, K. . -K. Nguyen, M. Lemay, & M. Cheriet. (2010). Ontology-Based Resource Description and Discovery Framework for Low Carbon Grid Networks. *2010 First IEEE International Conference on Smart Grid Communications*, 477–482. https://doi.org/10.1109/SMARTGRID.2010.5622090

Abedjan, Z., Golab, L., & Naumann, F. (2017). Data Profiling: A Tutorial. *Proceedings of the 2017 ACM International Conference on Management of Data*, 1747–1751. https://doi.org/10.1145/3035918.3054772

Alhazmi, A., Blount, T., & Konstantinidis, G. (2022). ForBackBench: A Benchmark for Chasing vs. Query-Rewriting. *Proc. VLDB Endow.*, *15*(8), 1519–1532. https://doi.org/10.14778/3529337.3529338

Alhazmi, A., & Konstantinidis, G. (2022). *OBDA vs Forward Chaining: The ForBackBench Framework*.

Andres, S., Iordanou, C., Laoutaris, N., & others. (2023). Understanding the Price of Data in Commercial Data Marketplaces. *IEEE International Conference on Data Engineering*.

Azcoitia, S. A., Iordanou, C., & Laoutaris, N. (2022). Measuring the Price of Data in Commercial Data Marketplaces. *Proceedings of the 1st International Workshop on Data Economy*, 1–7. https://doi.org/10.1145/3565011.3569053

Azcoitia, S. A., & Laoutaris, N. (2022). A Survey of Data Marketplaces and Their Business Models. *SIGMOD Rec.*, *51*(3), 18–29. https://doi.org/10.1145/3572751.3572755

Azcoitia, S. A., Paraschiv, M., & Laoutaris, N. (2022). Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/3557915.3561470

B. Pande, K. Padamwar, S. Bhattacharya, S. Roshan, & M. Bhamare. (2022). A Review of Image Annotation Tools for Object Detection. *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 976–982. https://doi.org/10.1109/ICAAIC53929.2022.9792665

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., … Zheng, J. (2016). The Ontology for Biomedical Investigations. *PLOS ONE*, *11*(4), e0154556. https://doi.org/10.1371/journal.pone.0154556

Benedikt, M., Konstantinidis, G., Mecca, G., Motik, B., Papotti, P., Santoro, D., & Tsamoura, E. (2017). Benchmarking the chase. *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 37–52.

Bertossi, L., & Li, L. (2013). Achieving Data Privacy through Secrecy Views and Null-Based Virtual Updates. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 987–1000. https://doi.org/10.1109/TKDE.2012.86

Bonatti, P. A., & Sauro, L. (2013). A Confidentiality Model for Ontologies. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, & K. Janowicz (Eds.), *The Semantic Web – ISWC 2013* (pp. 17–32). Springer Berlin Heidelberg.

Bonatti, P., Kirrane, S., Petrova, I. M., Sauro, L., & Schlehahn, E. (2022). *The SPECIAL Usage Policy Language, version 1.1*. SPECIAL H2020 EU project. https://ai.wu.ac.at/policies/policylanguage/

Carbon, S., Champieux, R., McMurry, J. A., Winfree, L., Wyatt, L. R., & Haendel, M. A. (2019). An analysis and metric of reusable data licensing practices for biomedical resources. *PLoS ONE*, *14*(3). Scopus. https://doi.org/10.1371/journal.pone.0213090

Carvalho, R., & Lioudakis, G. (2020). *Final specification and prototyping of the process re-engineering framework* (BPR4GDPR Deliverable D4.2).

Čebirić, Š., Goasdoué, F., & Manolescu, I. (2015). Query-Oriented Summarization of RDF Graphs. In S. Maneth (Ed.), *Data Science* (pp. 87–91). Springer International Publishing.

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., & Groth, P. (2020). Dataset search: A survey. *The VLDB Journal*, *29*(1), 251–272. https://doi.org/10.1007/s00778-019-00564-x

Chirkova, R., & Yu, T. (2017). Exact Detection of Information Leakage: Decidability and Complexity. In A. Hameurlain, J. Küng, R. Wagner, S. Madria, & T. Hara (Eds.), *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXII: Special Issue on Big Data Analytics and Knowledge Discovery* (pp. 1–23). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-55608-5_1

Consens, M. P., Fionda, V., Khatchadourian, S., & Pirrò, G. (2015). S+EPPs: Construct and Explore Bisimulation Summaries, plus Optimize Navigational Queries; All on Existing SPARQL Systems. *Proc. VLDB Endow.*, *8*(12), 2028–2031. https://doi.org/10.14778/2824032.2824128

Corcho, O. (2006). Ontology Based Document Annotation: Trends and Open Research Problems. *Int. J. Metadata Semant. Ontologies*, *1*(1), 47–57. https://doi.org/10.1504/IJMSO.2006.008769

Cremaschi, M., Avogadro, R., & Chieregato, D. (2022). s-elBat: A Semantic Interpretation Approach for Messy taBle-s. *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, SemTab 2021, Co-Located with the 21st International Semantic Web Conference, ISWC 2022, Virtual Conference, October 23-27, 2022.*, 59–71. https://ceur-ws.org/Vol-3320/paper7.pdf

de Cesare, S., & Geerts, G. L. (2012). Toward a Perdurantist Ontology of Contracts. In M. Bajec & J. Eder (Eds.), *Advanced Information Systems Engineering Workshops* (pp. 85–96). Springer Berlin Heidelberg.

Diao, Y., Guzewicz, P., Manolescu, I., & Mazuran, M. (2019). Spade: A Modular Framework for Analytical Exploration of RDF Graphs. *Proc. VLDB Endow.*, *12*(12), 1926–1929. https://doi.org/10.14778/3352063.3352101

Doe, S. (2018). Practical Privacy: Report from the GDPR World. *Legal Information Management*, *18*(2), 76–79. Cambridge Core. https://doi.org/10.1017/S1472669618000178

Dudáš, M., Svátek, V., & Mynarz, J. (2015). Dataset Summary Visualization with LODSight. In F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker, & A.

Zimmermann (Eds.), *The Semantic Web: ESWC 2015 Satellite Events* (pp. 36–40). Springer International Publishing.

Dwork, C. (2008). Differential Privacy: A Survey of Results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and Applications of Models of Computation* (pp. 1–19). Springer Berlin Heidelberg.

Fensel, D., Facca, F. M., Simperl, E., & Toma, I. (2011). *Semantic web services* (Vol. 357). Springer.

Filipczuk, D., Gerding, E. H., & Konstantinidis, G. (2023). *Consent Management in Data Workflows: A Graph Problem*.

Gangl, M. (2019). *THE IMPACT OF THE GDPR ON THIRD-PARTY CONTRACTS IN THE CLOUD SERVICE INDUSTRY* [Tilburg University]. https://arno.uvt.nl/show.cgi?fid=149355

Ghorbani, A., & Zou, J. (2019). Data shapley: Equitable valuation of data for machine learning. *International Conference on Machine Learning*, 2242–2251.

Grau, B. C., & Kostylev, E. V. (2016). Logical Foundations of Privacy-Preserving Publishing of Linked Data. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 943–949.

Hepp, M. (2008). GoodRelations: An Ontology for Describing Products and Services Offers on the Web. In A. Gangemi & J. Euzenat (Eds.), *Knowledge Engineering: Practice and Patterns* (pp. 329–346). Springer Berlin Heidelberg.

Hils, M., Woods, D. W., & Böhme, R. (2020). Measuring the Emergence of Consent Management on the Web. *Proceedings of the ACM Internet Measurement Conference*, 317–332. https://doi.org/10.1145/3419394.3423647

Hinze, A., Heese, R., Luczak-Rösch, M., & Paschke, A. (2012). Semantic Enrichment by Non-experts: Usability of Manual Annotation Tools. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, & E. Blomqvist (Eds.), *The Semantic Web – ISWC 2012* (pp. 165–181). Springer Berlin Heidelberg.

Iannella, R., & Villata, S. (2018). *ODRL Information Model 2.2*. W3C. https://www.w3.org/TR/odrl-model/

IDSA. (2021). *Usage Control in the International Data Spaces* (Position Paper Version 3.0). IDSA. https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-Usage-Control-in-the-IDS-V3..pdf

Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., Li, B., Zhang, C., Spanos, C., & Song, D. (2019). Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms. *Proc. VLDB Endow.*, *12*(11), 1610–1623. https://doi.org/10.14778/3342263.3342637

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., & Spanos, C. J. (2019). Towards Efficient Data Valuation Based on the Shapley Value. In K. Chaudhuri & M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (Vol. 89, pp. 1167–1176). PMLR. https://proceedings.mlr.press/v89/jia19a.html

Kabilan, V., & Johannesson, P. (2003). Semantic Representation of Contract Knowledge Using Multi Tier Ontology. *Proceedings of the First International Conference on Semantic Web and Databases*, 378–397.

Khalili, A., & Auer, S. (2013). User interfaces for semantic authoring of textual content: A systematic literature review. *Journal of Web Semantics*, *22*, 1–18. https://doi.org/10.1016/j.websem.2013.08.004

Kim, J., Kim, J., Lee, D., & Chung, K.-Y. (2014). Ontology driven interactive healthcare with wearable sensors. *Multimedia Tools and Applications*, *71*(2), 827–841. https://doi.org/10.1007/s11042-012-1195-9

Konrath, M., Gottron, T., Staab, S., & Scherp, A. (2012). SchemEX — Efficient construction of a data catalogue by stream-based indexing of linked data. *The Semantic Web Challenge 2011*, *16*, 52–58. https://doi.org/10.1016/j.websem.2012.06.002

Konstantinidis, G., & Ambite, J. L. (2011). Scalable query rewriting: A graph-based approach. *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, 97–108.

Konstantinidis, G., Holt, J., & Chapman, A. (2021). Enabling Personal Consent in Databases. *Proc. VLDB Endow.*, *15*(2), 375–387. https://doi.org/10.14778/3489496.3489516

L. Guo, Q. Liu, K. Shi, Y. Gao, J. Luo, & J. Chen. (2021). A Blockchain-Driven Electronic Contract Management System for Commodity Procurement in Electronic Power Industry. *IEEE Access*, *9*, 9473–9480. https://doi.org/10.1109/ACCESS.2021.3049562

L. Youseff, M. Butrico, & D. Da Silva. (2008). Toward a Unified Ontology of Cloud Computing. *2008 Grid Computing Environments Workshop*, 1–10. https://doi.org/10.1109/GCE.2008.4738443

Lamparter, S., & Schnizler, B. (2006). Trading Services in Ontology-Driven Markets. *Proceedings of the 2006 ACM Symposium on Applied Computing*, 1679–1683. https://doi.org/10.1145/1141277.1141674

Lawson, J., Cabili, M. N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S. R., Boyles, R. R., Brookes, A. J., Brush, M., Burdett, T., Clissold, H., Donnelly, S., Dyke, S. O. M., Freeberg, M. A., Haendel, M. A., Hata, C., Holub, P., Jeanson, F., … Courtot, M. (2021). The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genomics*, *1*(2), 100028. https://doi.org/10.1016/j.xgen.2021.100028

Lioudakis, G., Papagiannakopoulou, E., Koukovini, M., Dellas, N., Kalaboukas, K., Bracciale, L., Raso, E., Bianchi, G., Loreti, P., Barracano, P., Alexakis, S., Medeiros de Carvalho, R., & Hassani, M. (2021). GDPR Compliance Made Easier: The BPR4GDPR Project. *ARIS2 - Advanced Research on Information Systems Security*, *1*(1), 5–23. https://doi.org/10.56394/aris2.v1i1.1

Lioudakis, G. V., Koukovini, M. N., Papagiannakopoulou, E. I., Dellas, N., Kalaboukas, K., de Carvalho, R. M., Hassani, M., Bracciale, L., Bianchi, G., Juan-Verdejo, A., Alexakis, S., Gaudino, F., Cascone, D., & Barracano, P. (2020). Facilitating GDPR Compliance: The H2020 BPR4GDPR Approach. In I. O. Pappas, P. Mikalef, Y. K. Dwivedi, L. Jaccheri, J. Krogstie, & M. Mäntymäki (Eds.), *Digital Transformation for a Sustainable Society in the 21st Century* (pp. 72–78). Springer International Publishing.

Liu, J., Chabot, Y., Troncy, R., Huynh, V.-P., Labbé, T., & Monnin, P. (2023). From tabular data to knowledge graphs: A survey of semantic table interpretation

tasks and methods. *Journal of Web Semantics*, *76*, 100761. https://doi.org/10.1016/j.websem.2022.100761

Loetpipatwanich, S., & Vichitthamaros, P. (2020). Sakdas: A Python Package for Data Profiling and Data Quality Auditing. *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)*, 1–4.

Louati, A., Aufaure, M.-A., Lechevallier, Y., & Chatenay-Malabry, F. (2011). Graph Aggregation: Application to Social Networks. *HDSDA*, 157–177.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-Diversity: Privacy beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data*, *1*(1), 3-es. https://doi.org/10.1145/1217299.1217302

Mader, C., Pullmann, J., Petersen, N., Lohmann, S., & Lange-Bever, C. (2022). *International Data Spaces Information Model*. Fraunhofer IAIS/EIS, Fraunhofer FIT. https://international-data-spaces-association.github.io/InformationModel/docs/index.html

Martin, D., Paolucci, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D., Parsia, B., Payne, T., Sabou, M., Solanki, M., Srinivasan, N., & Sycara, K. (2005). Bringing Semantics to Web Services: The OWL-S Approach. In J. Cardoso & A. Sheth (Eds.), *Semantic Web Services and Web Process Composition* (pp. 26–42). Springer Berlin Heidelberg.

Mihindukulasooriya, N., Poveda-Villalón, M., García-Castro, R., & Gómez-Pérez, A. (2015). Loupe—An Online Tool for Inspecting Datasets in the Linked Data Cloud. *International Workshop on the Semantic Web*.

Möller, M. L., Berton, N., Klettke, M., Scherzinger, S., & Störl, U. (2019). jHound: Large-Scale Profiling of Open JSON Data. *Datenbanksysteme Für Business, Technologie Und Web*.

Mottin, D., Lissandrini, M., Velegrakis, Y., & Palpanas, T. (2016). Exemplar queries: A new way of searching. *The VLDB Journal*, *25*(6), 741–765. https://doi.org/10.1007/s00778-016-0429-2

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data Lake Management: Challenges and Opportunities. *Proc. VLDB Endow.*, *12*(12), 1986–1989. https://doi.org/10.14778/3352063.3352116

O. Perrin & C. Godart. (2004). An approach to implement contracts as trusted intermediaries. *Proceedings. First IEEE International Workshop on Electronic Contracting, 2004.*, 71–78. https://doi.org/10.1109/WEC.2004.1319511

OECD. (2019). *The Path to Becoming a Data-Driven Public Sector*. https://doi.org/10.1787/059814a7-en

Papagiannakopoulou, E. (2020). *Final specification and prototyping of the policy framework* (BPR4GDPR Deliverable D3.3).

Park, J., & Brenza, A. (2015). Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art. *Information Technology and Libraries*, *34*(3), 22–42. https://doi.org/10.6017/ital.v34i3.5889

Petrova, G. G., Tuzovsky, A. F., & Aksenova, N. V. (2017). Application of the Financial Industry Business Ontology (FIBO) for development of a financial organization ontology. *Journal of Physics: Conference Series*, *803*.

Q. Song, Y. Wu, & X. L. Dong. (2016). Mining Summaries for Knowledge Graph Search. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 1215–1220. https://doi.org/10.1109/ICDM.2016.0162

R. Greenwell, X. Liu, & K. Chalmers. (2016). Pricing Ontology for Task-Oriented Cloud Sourcing. *2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, 63–70. https://doi.org/10.1109/FiCloud.2016.17

Rehm, H. L., Page, A. J. H., Smith, L., Adams, J. B., Alterovitz, G., Babb, L. J., Barkley, M. P., Baudis, M., Beauvais, M. J. S., Beck, T., Beckmann, J. S., Beltran, S., Bernick, D., Bernier, A., Bonfield, J. K., Boughtwood, T. F., Bourque, G., Bowers, S. R., Brookes, A. J., … Birney, E. (2021). GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics*, *1*(2), 100029. https://doi.org/10.1016/j.xgen.2021.100029

Riondato, M., García-Soriano, D., & Bonchi, F. (2017). Graph summarization with quality guarantees. *Data Mining and Knowledge Discovery*, *31*(2), 314–349. https://doi.org/10.1007/s10618-016-0468-8

Rizvi, S., Mendelzon, A., Sudarshan, S., & Roy, P. (2004). Extending Query Rewriting Techniques for Fine-Grained Access Control. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, 551–562. https://doi.org/10.1145/1007568.1007631

Rosenthal, A. S., & Sciore, E. (2000). View security as the basis for data warehouse security. *Design and Management of Data Warehouses*.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). The Shapley Value in Machine Learning. In L. D. Raedt (Ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial*

Intelligence, IJCAI-22 (pp. 5572–5579). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/778

S. Campinas, T. E. Perry, D. Ceccarelli, R. Delbru, & G. Tummarello. (2012). Introducing RDF Graph Summary with Application to Assisted SPARQL Formulation. *2012 23rd International Workshop on Database and Expert Systems Applications*, 261–266. https://doi.org/10.1109/DEXA.2012.38

Shapley, L. S. (1952). *A Value for N-Person Games*. RAND Corporation. https://doi.org/10.7249/P0295

Simić, S., Marković, M., & Gostojić, S. (2021). Smart Contract and Blockchain Based Contract Management System. *7th Conference on the Engineering of Computer Based Systems*. https://doi.org/10.1145/3459960.3459975

Simon, R., Barker, E., Isaksen, L., & De Soto CaÑamares, P. (2017). Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2. *Journal of Map & Geography Libraries*, *13*(1), 111–132. https://doi.org/10.1080/15420353.2017.1307303

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., Lewis, S., & The OBI Consortium. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–1255. https://doi.org/10.1038/nbt1346

Sweeney, L. (2002). K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Tauqeer, A., Kurteva, A., Chhetri, T. R., Ahmeti, A., & Fensel, A. (2022). Automated GDPR Contract Compliance Verification Using Knowledge Graphs. *Information*, *13*(10). https://doi.org/10.3390/info13100447

Z. Liu & A. Zhang. (2020). Sampling for Big Data Profiling: A Survey. *IEEE Access*, *8*, 72713–72726. https://doi.org/10.1109/ACCESS.2020.2988120

9.2 Acronyms

Acronyms List	
DoA	Description of Action
DM	Data Marketplace
DPW	Data Processing Workflow
HPC	High Performance Computing
ML	Machine Learning
PC	Project Coordinator
PDP	Policy Decision Point
PMB	Project Management Board
PMP	Policy Management Point
PPR	Project Periodic Report
QM	Quality Management
RM	Risk Management
TM	Technical Manager
WPL	Work Packages Leaders
US	User story
REQ	Requirement
XAI	Explainable AI

Table 51. Acronyms

ANNEX 1: TEMPLATES FOR USER STORIES AND REQUIREMENTS

User stories are formulated in this form:

As a <type of user> I want to <goal/objective> so that <benefit/result/some reason>

Functional and non-functional requirements are formulated using this template:

Requirement ID	*Every requirement should have a unique identifier*
Requirement Short Title	*A short title of the requirement*
Requirement Description	*A description of the requirement*
Source	*State of the Art/AS-IS Interview/TO-BE Interview/User Stories/Architecture(DoA)*
Type	*Does not apply/Functional/Non-functional*
Phase/subphase/components of the UPCAST Plugin(s)	*Does not apply/Pricing/Federated Learning/etc.*
Means of verification	*Synthetic datasets/Validation with a Business Case/Collaboration between team members etc.*
KPIs affected (updated from AS-IS interview)	*Fill in if this requirement directly affects a KPI*
Depends on	*Fill in if this requirement depends by the achievement of other requirements*
Priority	*Could have/Should have/Must have*
Technical feasibility	*Feasible (Easy to implement, difficult to implement)/ not feasible.* *Techncial partners provide input here*

ANNEX 2: AS-IS INTERVIEWS

Date	<Date>
Case Partner	<Use case partner>
Interviewer	<Names of interviewers>
Participants	<List of participants>
Modality	<Online | Physical>

Purpose

- To better understand the business cases and how it will use/benefit from the tools developed in the project.
- To prepare follow-up TO-BE interviews focusing on user requirements (written as user stories).

Questions

Introductory questions
What is the business case about?
Who are the stakeholders?

Describe the current processes or existing products that you aim to extend (if it is not a new product/service) using the tools
Process Perspective
What are the business objectives related to your processes?
Describe the challenges, issues and risks that you handle to enact your processes
Which are the main steps of the process? For each step, what are the scalability requirements/constraints (both vertical and horizontal)? Vertical scaling means to add more resources (e.g., CPU, memory) to existing machines, horizontal scaling means adding more machines to the pool of resources.
What development process do you use? Do you have issues with moving from development to the operation environment?
What is technical background of people involved into the main steps of process construction and monitoring? If there are people with different backgrounds, how communication between these groups is organised?
Is the process still as it was described in the UPCAST Description of Work? Do you foresee any changes?
Are there any “critical” steps? (e.g., steps in which the UPCAST Plugins could help to obtain the KPI of interest)
How should/could the processes change, with the support obtained from UPCAST?

Which is the innovation that UPCAST will provide to the management of your processes?
Which are the envisioned risks to leverage the UPCAST plugins to improve your processes? I.e., what considerations must be made when you integrate the UPCAST plugins in your existing environment?
Data Perspective
How is the data required for your processes generated or obtained? (i.e., what are your data sources?)
What types of internally generated data do you use? Are there personal data (e.g., of customers, employees, etc.) and IPR-protected data?
If data is generated externally, how do you manage to obtain it? Is it free? Is some of this data personal, and if so, sensitive (e.g., genetic data, biometric data, data related to health, politics, religion, etc.)? Are there any privacy constraints to obtain/analyze data?
Are there any relevant standards or vocabularies for encoding/representing your data?
Do you have any issues with data provenance? Data provenance is the documentation of where a piece of data comes from and the processes and methodology by which it was produced. If so, which issues?
If the data is in the form of one, or more, DB/s, which are the relevant queries, in natural language, that should be asked to it?
Is data, even only in some steps, real-time?

What monitoring and visualization of the data processing do you use? For example, do you use monitoring tools that enable you to detect errors in the input data or visualisation tools to generate data reports of the results of a data analysis task.
Do you use case/conditional or loops in your data processing? For example, if data do not satisfy certain threholds or have a particular value, they are filtered out of the dataset used in the data processing.
Do you have full traceability of the data you use in your data processing? If not, in which steps of the process do you “lose track of data”?
Are you aware of data pipelines that lay behind the enactment of your actual processes? This relates to both internal and external data pipelines.
In which way knowing the data pipeline of your processes could help you to improve their enactment?
Which parts of your current data processing solutions (if any) do you consider re-usable in the context of UPCAST? Reusable means that an AS-IS data processing solution can be reused in the TO-BE setting of the business case.
Technological Perspective
Which technologies (hardware/software/machineries etc.) do you currently employ to manage your process?
Do you leverage cloud features to enact your process?
Do you envision the introduction of new technologies to use the UPCAST plugins in the context of your processes?

Results and KPIs Achievement
How are you going to measure the results gained by UPCAST? Are the KPIs defined for this business case in the UPCAST proposal still relevant?
Describe the concrete tasks that need to be performed to achieve the KPIs defined for this business case.
Integration with other business cases
Do you think that your results could impact on the other Business Cases of the UPCAST project, or other companies in general?
Do you think that the other partners of the project (e.g., the technology providers) could be of help for yours?
Do you offer something (e.g., software, hardware, data) to the other partners?
Do you need something from the other partners?
General Aspects
How well is the business case aligned with your company strategy?
Can you / do you want to realize the business case if the UPCAST Project did not exist? What kind of support/backing do you get from your company?
What is the value proposition? Who is going to pay for it (if none, how do you change the life of others that may use your service)?

What are your concrete plans for dissemination and impact creation – business aspects and potential clients, communication with clients, etc?
What can we start to do now in the context of UPCAST?

ANNEX 3: DATASETS

Biomedical and Genomic Data Sharing

Description of dataset	Provider	Dataset Size	Format	Data Origin	Time Coverage	Language	Metadata standard / vocabularies	Personal data	Access Restrictions
Transcriptomic data: The number of reads (sequencing data) that were aligned to specific gene regions, for each sample in a RNA-seq experiment. They are used to quantify the expression levels of genes or transcripts in a sample.	NHRF	1 -200 MB	csv/tsv	1. In-house (NHRF) Next Generation Sequencing experiments, 2. Downloaded from databases	N/A	English	Alternative candidates: Minimum Information About a Sequencing Experiment (MIAME), The Global Alliance for Genomics and Health (GA4GH), International Nucleotide Sequence Database Collaboration (INSDC), ISA Framework standards.	1. N (existing in-house data) 2.Y (to be downloaded from databases) (Anonymised)	Credentials
Genomic data: Genomic variant information of somatic cancer biosample, containining the position of the mutation, the reference and alternate alleles, and any annotations or attributes associated with each variant.	NHRF	10-200 MB	csv/tsv	1. In-house (NHRF) RNA-seq experiments, 2. Downloaded from databases	N/A	English	Metadata will be created following Standards of the Global Alliance for Genomics and Health (GA4GH) and other recommendations (MANE transcript, Sequence Ontology , Data Use Ontology etc)	Y (Anonymised)	Credentials

Clinical data: Demographic and clinical characteristics/cancer biosample	NHRF	1-10 MB	Excel	1. Clinical collaborators 2.Databases	N/A	English	Phenopackets: The Phenopacket specification is an open machine-readable schema that supports the global exchange of disease and phenotype information. National Cancer institute's Thesaurus (NCIt) is used for cancer biosamples, and is the de facto standard for cancer knowledge representation and regulatory submission.	Y (Anonymised)	Credentials

Public Administration

Description of dataset	Provider	Dataset Size	Format	Data Origin	Time Coverage	Language	Metadata standard / vocabularies	Personal data	Access Restrictions
General demographics: Population and gender distribution, Household composition, Population and occupation distribution, Economically active population and workplace location	Hellenic Statistical Authority (ELSTAT)	N/A	.xls	Decade Census 2011 and 2021	2011, and 2021 remain to be analysed and updated	greek	N/A	N	N
Urban statistics and living conditions: Continuous residential urban fabric, Discontinuous residential urban fabric, Industrial, commercial, public, military and private units, Transport infrastructure, Other artificial areas, Green urban areas and sports and leisure facilities, Agricultural areas, Share of land, Natural areas	Eurostat	N/A	.xls	Annual (according to the availability of the national data)	1990-2021	english	N/A	N	N
Population and Social Conditions: Dwellings construction period, Dwellings heating availability, Dwellings insulation availability, Dwellings surface (m2), Domestic density m2/person (incl. private and rented), Households heating source, Households water heating source, Households energy for cooking, Dwellings lacking basic amenities, Thessaloniki greater area 1989-2021 (compare with other Greek cities and Greece in total), Empty conventional dwellings, Thessaloniki greater area 1989-2021 (compare with other Greek cities and Greece in total)	Hellenic Statistical Authority (ELSTAT) & Eurostat	N/A	.xls	Decade Census 2011 and 2021 & Annual (according to the availability of the national data)	2011, and 2021 remain to be analysed and updated (ELSTAT) & 1989-2021 (EUROSTAT)	greek & english	N/A	N	N
Transport statistics - Households - Private Vehicles: Households (incl. number of members) and car availability, Households and number of cars, Households and parking availability	Hellenic Statistical Authority (ELSTAT)	N/A	.xls	Decade Census 2011 and 2021	2011, and 2021 remain to be analysed and updated	greek	N/A	N	N

Transport statistics: Length of bicycle network (dedicated cycle paths and lanes) - km, Cost of a combined monthly ticket (all modes of public transport) for 5-10 km in the central zone - EUR, Cost of a taxi ride of 5 km to the centre at daytime - EUR, Number of private cars registered, Number of deaths in road accidents, People killed in road accidents per 10000 pop.	Eurostat	N/A	.xls	Annual (according to the availability of the national data)	1990-2021	english	N/A	N	N
Urban mobility statistics: Current total number of detections per iTravel device, iTravel devices characteristics, Predefined paths between the iTravel devices, Current travel times for selected paths in Greece produced using itravel detections, Current speeds on openstreetmap network links, The current traffic conditions on openstreetmap network links, Floating Car Data in Thessaloniki, Greece (Historical datasets), Floating car data in Thessaloniki, Greece in -almost- real time	CERTH-HIT OpenData Hub Greece	N/A	.csv	Data generated by iTravel devices	2016/2017-2022/2023	english	N/A	N	additional license (http://opendata.imet.gr/about)

Road freight mobility statistics: Total transported goods in thousand tones, Transported goods for products of agriculture, hunting, and forestry; fish and other fishing products, Metal ores and other mining and quarrying products; peat; uranium and thorium, Food products, beverages and tobacco, Textiles and textile products; leather and leather products, Wood and products of wood and cork (except furniture); articles of straw and plaiting materials; pulp, paper, and paper products; printed matter and recorded media, Coke and refined petroleum products (Thessaloniki NUTSIII 2015-2021), Chemicals, chemical products, and man-made fibers; rubber and plastic products; nuclear fuel , Other nonmetallic mineral products, Basic metals; fabricated metal products, except machinery and equipment, Machinery and equipment n.e.c.; office machinery and computers; electrical machinery and apparatus n.e.c.; radio, television and communication equipment and apparatus; medical, precision and optical instruments; watches and clocks, Transport equipment, Furniture; other manufactured goods n.e.c. Secondary raw materials; municipal wastes and other wastes, Mail, parcels, Equipment and material utilized in the transport of goods, Goods moved in the course of household and office removals; baggage and articles accompanying travellers; motor vehicles being moved for repair; other non-market goods n.e.c, Grouped goods: a mixture of types of goods which are transported togetherm, Unidentifiable goods: goods which that cannot be	Eurostat	N/A	.xls	Annual (according to the availability of the national data)	2008-2021	english	N/A	N	N

identified and therefore cannot be assigned to previous groups, Other goods n.e.c.									

Environmental Data: Total number of hours of sunshine per day (Thessaloniki; 2001, 2004, 2010-2015), Average temperature of warmest month - degrees, Average temperature of coldest month – degrees (Thessaloniki; 2001, 2004, 2010-2015, Rainfall - litre/m² (Thessaloniki; 2001, 2004, 2010-2015), Number of days ozone O3 concentrations exceed 120 μg/m³ (Thessaloniki; 2001, 2010-2014, Number of days particulate matter PM1, concentrations exceed 50 μg/m³ (Thessaloniki; 2001, 2010-2014), Annual average concentration of NO2 (μg/m³) (Thessaloniki, 2010-2014), Annual average concentration of PM10 (μg/m³) (Thessaloniki, 2010-2014), Total use of water - m³ (Thessaloniki; 2004), Price of a m³ of domestic water – Euro (Thessaloniki, 2001, 2004-2009)	Eurostat	N/A	.xls	Annual (according to the availability of the national data)	1989-2021	english	N/A	N	N
Air Pollution Emissions: Agia Sophia station (AGS), AUTH station (APT), Panorama station (PAO), Kalamaria station (KAL), Kordelio station (KOD), Sindos station (SIN), Neochorouda station (NEO), Aristotelous (ARI)	Ministry of the Environment and Energy	N/A	.dat	Daily	2001-2005, 2006-2010, 2011-2015, 2016-2020	greek	N/A	N	N

Health and Fitness

Description of dataset	Provider	Dataset Size	Format	Data Origin	Time Coverage	Language	Metadata standard / vocabularies	Personal data	Access Restrictions

Heart rate monitoring data: Heart rate data collected by monitoring users while training (fitness, medical)	NIS	1-50MB per batch/user	json/csv	Collected from sensors of users from 3 different APPS	one training per batch (from 30min to 4hr)	English	N/A	Y (Anonymised)	Credentials
Acceleration monitoring data: Monitoring Acceleration data from accelerometer for trainee's better performance	NIS	5-150MB per batch/user	json/csv	Collected from sensors of users from 3 different APPS	one training per batch (from 30min to 4hr)	English	N/A	Y (Anonymised)	Credentials
Formatted-Combined Data: HR and Acceleration data formatted for our own analysis with added magnitude parameter. With added context data.	NIS	5-400MB per batch / user / training	json/csv	Formatted raw data from DMP ID 1 and 2	one exercise or one training per batch (from 30min to 4hr)	English	N/A	Y (Anonymised)	Credentials

Digital Marketing (JOT)

Description of dataset	Provider	Dataset Size	Format	Data Origin	Time Coverage	Language	Metadata standard / vocabularies	Personal data	Access Restrictions
Visitors: In Google Analytics, a user is a visitor who has initiated a session on your website: the moment a person lands on any page of your site, they are identified as either a new or returning user.	Customer's Systems	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
ROAS: ROAS, or return on ad spend is the metric used to track the impact of your paid advertising on your bottom line. Similar to return on investment, return on ad spend looks specifically at how much revenue you generate compared to how much you spend on paid channels	Customer's Systems	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials

Click through Rate (CTR): Clickthrough rate (CTR) can be used to gauge how well your keywords and ads, and free listings, are performing. CTR is the number of clicks that your ad receives divided by the number of times your ad is shown: clicks ÷ impressions = CTR	Customer's Systems	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Click per Cost (CPC) Cost-per-click (CPC) bidding means that you pay for each click on your ads.	Customer's Systems	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Quality Score: Quality Score is a diagnostic tool meant to give you a sense of how well your ad quality compares to other advertisers.	Customer's Systems	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Visitors: A user is a visitor who has initiated a session on your website: the moment a person lands on any page of your site, they are identified as either a new or returning user.	Customer's Systems	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Bounce Rate: Bounce rate is the inverse of engagement rate. For example, let's say someone visits your website, reads some of your content for less than 10 seconds, and then leaves. While they were on your website, they didn't trigger any events or visit any other pages.	Customer's Systems	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
JOT Marketing Campaigns	Customer's Systems	N/A	JSON	Customer's Systems	N/A	English	N/A	N	Credentials

Digital Marketing (Cactus)

Description of dataset	Provider	Dataset Size	Format	Data Origin	Time Coverage	Language	Metadata standard / vocabularies	Personal data	Access Restrictions

Visitors: In Google Analytics, a user is a visitor who has initiated a session on your website: the moment a person lands on any page of your site, they are identified as either a new or returning user.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
ROAS: ROAS, or return on ad spend is the metric used to track the impact of your paid advertising on your bottom line. Similar to return on investment, return on ad spend looks specifically at how much revenue you generate compared to how much you spend on paid channels	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Click through Rate (CTR): Clickthrough rate (CTR) can be used to gauge how well your keywords and ads, and free listings, are performing. CTR is the number of clicks that your ad receives divided by the number of times your ad is shown: clicks ÷ impressions = CTR	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Click per Cost (CPC) Cost-per-click (CPC) bidding means that you pay for each click on your ads.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Quality Score: Quality Score is a diagnostic tool meant to give you a sense of how well your ad quality compares to other advertisers.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Visitors: A user is a visitor who has initiated a session on your website: the moment a person lands on any page of your site, they are identified as either a new or returning user.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials

Bounce Rate: Bounce rate is the inverse of engagement rate. For example, let's say someone visits your website, reads some of your content for less than 10 seconds, and then leaves. While they were on your website, they didn't trigger any events or visit any other pages.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Channels: Channels are groupings of different sources (the origin of traffic, e.g. a search engine such as 'google' or a domain name) and mediums (the general category of sources, e.g. 'organic' for all organic search or 'referral' for all web referrals).	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Conversion Rate: Conversion rates are calculated by simply taking the number of conversions and dividing that by the number of total ad interactions that can be tracked to a conversion during the same time period.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Budget: A budget is the amount of money you want to spend on showing people your ads. It is also a cost control tool. It helps control your overall spend for a campaign or ad set, the same way a bid strategy helps control your cost per result.	Cactus' Customer	N/A	JSON	META API	N/A	English	N/A	N	Credentials
ROAS: ROAS is simply the total revenue generated from your Meta ads (your return) divided by your total ad spend.	Cactus' Customer	N/A	JSON	META API	N/A	English	N/A	N	Credentials
Landing Page View: A 'landing page view' is when a person lands on your ad's destination URL (landing page) after clicking a link in your ad	Cactus' Customer	N/A	JSON	META API	N/A	English	N/A	N	Credentials
Cost per Lead (CPL): CPL is calculated on Meta by dividing your spend by leads generated in campaigns using the lead generation campaign objective.	Cactus' Customer	N/A	JSON	META API	N/A	English	N/A	N	Credentials

Reach: Reach is the number of people who saw any content from your Page or about your Page.	Cactus' Customer	N/A	JSON	META API	N/A	English	N/A	N	Credentials
Impressions: Impressions on Meta tell you how many times your content was displayed on a screen	Cactus' Customer	N/A	JSON	META API	N/A	English	N/A	N	Credentials
Frequency: Meta defines frequency as the average number of times each person saw your ad.	Cactus' Customer	N/A	JSON	META API	N/A	English	N/A	N	Credentials
Quality Score: Quality Score is a diagnostic tool meant to give you a sense of how well your ad quality compares to other advertisers	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Roas: ROAS stands for return on ad spend—a marketing metric that measures the amount of revenue your business earns for each dollar it spends on advertising.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Budget: Through Google Ads budget management, you bid on keywords that people search with on Google – related to your business for a chance to show ads in Google search results.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Click through Rate (CTR): CTR is the number of clicks that your ad receives divided by the number of times your ad is shown: clicks ÷ impressions = CTR	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Click per Cost (CPC): Cost-per-click (CPC) bidding means that you pay for each click on your ads.	Cactus' Customer	N/A	JSON	GOOGLE API	N/A	English	N/A	N	Credentials
Sales Data: Sales data is any information that is machine-readable and of benefit to anyone who tries to track,understand and predict the sales of a company	Cactus' Customer	N/A	CSV	Customer's Systems	N/A	English	N/A	N	Credentials

Profit & Loss: Profit and loss (P&L) statement refers to a financial statement that summarizes the revenues, costs, and expenses incurred during a specified period, usually a quarter or fiscal year.	Cactus' Customer	N/A	CSV	Customer's Systems	N/A	English	N/A	N	Credentials