



ICARUS:

“Aviation-driven Data Value Chain for Diversified Global and Local Operations”

D1.2 – The ICARUS Methodology and MVP

Workpackage:	WP1 – ICARUS Data Value Chain Elaboration		
Authors:	UCY, Suite5, UBITECH		
Status:	Final	Classification:	Public
Date:	03/08/2018	Version:	1.00













Disclaimer:

The ICARUS project is co-funded by the Horizon 2020 Programme of the European Union. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Communities. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. © Copyright in this document remains vested with the ICARUS Partners.

ICARUS Project Profile

Grant Agreement No.:	780792
Acronym:	ICARUS
Title:	Aviation-driven Data Value Chain for Diversified Global and Local Operations
URL:	http://www.icarus2020.aero
Start Date:	01/01/2018
Duration:	36 months

Partners

	UBITECH (UBITECH)	Greece
	ENGINEERING - INGEGNERIA INFORMATICA SPA (ENG)	Italy
	PACE Aerospace Engineering and Information Technology GmbH (PACE)	Germany
	SUITE5 DATA INTELLIGENCE SOLUTIONS LIMITED (SUITE5)	Cyprus
	UNIVERSITY OF CYPRUS (UCY)	Cyprus
	CINECA CONSORZIO INTERUNIVERSITARIO (CINECA)	Italy
	OAG Aviation Worldwide LTD (OAG)	United Kingdom
	SingularLOGIC S.A. (SILO)	Greece
	ISTITUTO PER L'INTERSCAMBIO ISI SCIENTIFICO (ISI)	Italy
	CELLOCK LTD (CELLOCK)	Cyprus
	ATHENS INTERNATIONAL AIRPORT S.A (AIA)	Greece
	TXT e-solutions SpA (TXT) – 3 rd party of PACE	Italy

Document History

Version	Date	Author (Partner)	Remarks
0.10	18/04/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Initial draft of the Table of Contents (ToC)
0.11	23/04/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Update draft of the Table of Contents (ToC)
0.20	06/05/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Contribution to Section 2
0.21	07/05/2018	George Pallis (UCY)	Feedback for Section 2
0.22	15/05/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Updated Section 2 and contribution to Section 3
0.23	20/05/2018	Fenareti Lampathaki (Suite5), Dimitrios Miltiadou (UBITECH), George Pallis, Marios Dikaiakos (UCY)	Feedback for Section 2 and Section 3
0.30	10/06/2018	Fenareti Lampathaki (Suite5), Konstantinos Perakis (UBITECH)	Initial MVP features definition in Section 4 to initiate the demonstrators' assessment
0.40	12/06/2018	Loukas Pouis, Dimosthenis Stefanidis, George Pallis (UCY)	Updated Section 2 and Section 3
0.41	14/06/2018	Loukas Pouis, Dimosthenis Stefanidis, George Pallis (UCY)	Contribution for Section 1 and 5
0.42	15/06/2018	Marios Dikaiakos (UCY)	Feedback for Sections 2 and 3
0.43	15/06/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Updated Sections 2 and 3
0.44	22/06/2018	Dimitrios Alexandrou (UBITECH)	Deliverable Content Review
0.50	06/07/2018	Dimitrios Alexandrou (UBITECH)	Deliverable Content Updates (Sections 2 and 3)
0.60	11/07/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Updated Sections 1, 2, 3 and 5
0.70	25/07/2018	Fenareti Lampathaki (Suite5)	Updated Section 4 with inputs from demonstrators, OAG, ENG
0.80	30/07/2018	Fenareti Lampathaki (Suite5), Dimosthenis Stefanidis, George Pallis (UCY)	Final version containing updates from internal review
1.00	03/08/2018	Dimitrios Alexandrou (UBITECH)	Final version to be submitted to the EC

Executive Summary

ICARUS deliverable D1.2 “The ICARUS Methodology and MVP” constitutes a report of the performed work and the produced results of Task 1.4 “ICARUS Methodology and MVP Definition”. In particular, the purpose and scope of this deliverable is to formulate the ICARUS methodology and the Minimum Viable Product (MVP) in order: a) to reveal how the different concepts interrelate and b) to display the high-level usage scenarios of the ICARUS concept.

In this deliverable, the initial ICARUS methodology, which consists of different phases with each phase having its own steps and aspects, is elaborated. These phases interact and parts of them can be abstracted based on each stakeholder’s objective. Furthermore, the key challenges and the relevant methods are briefly presented for each step. More precisely, the phases and steps that constitute the methodology are the following:

- **Phase I - Data Collection:**
 - Step I.1: Data Retrieval
 - Step I.2: Data Anonymization
 - Step I.3: Data Quality Check
 - Step I.4: Data Curation
 - Step I.5: Data Check-In
- **Phase II - Data Enrichment:**
 - Step II.1: Semantic Enrichment and Annotation
 - Step II.2: Data Linking
- **Phase III - Asset Storage**
- **Phase IV - Asset Exploration and Extraction:**
 - Step IV.1: Asset Indexing and Searching
 - Step IV.2: Asset Export
- **Phase V - Data Analytics:**
 - Step V.1: Data Analysis
 - Step V.2: Data Visualization
- **Phase VI - Added Value Services:**
 - Asset Sharing
 - Recommendations
 - Notifications
- **Phase VII - Service Collection:**
 - Step VII.1: Service Check-In
 - Step VII.2: Testing and Assessment
 - Step VII.3: Service Asset Review

Furthermore, D1.2 presents the defined high-level usage scenarios of ICARUS, which are based on the ICARUS methodology, the input from the ICARUS pilots and the insights provided in D1.1 “Domain Landscape Review and Data Value Chain Definition”. Such scenarios were actually co-created in the face-to-face brainstorming sessions that were organized in the ICARUS plenary meeting in Nicosia on May 2018. More precisely, six scenarios (general workflow diagrams) as representative of all core differentiated ICARUS stakeholders (either providers or consumers) and three examples (more technical sub-diagrams) for each scenario have been defined.

In addition, the present deliverable outlines the process and the initial outcomes of the Minimum Viable Product (MVP) definition. The ICARUS platform features have been extracted from the defined methodology and the high-level scenarios and have been assessed (qualitatively and quantitatively) from the ICARUS demonstrators and the core data provider in order to get feedback with regard to their added value for their own operations and for the aviation industry, in general. The preliminary MVP consolidation contains 66 features, as depicted in the following table:

Phase I	<ul style="list-style-type: none"> • PLATF_F_01 Retrieval of data directly from an aviation stakeholder's back-end system • PLATF_F_02 Uploading of data assets as files extracted by the aviation stakeholder's back-end system • PLATF_F_06 (Semi-)Automatic quality check of the data and assessment of quality level • PLATF_F_09 (Semi-)Automatic on-the-fly anonymization in ICARUS • PLATF_F_11 (Semi-)Automatic extraction of and navigation within the Data Model of a Data Asset • PLATF_F_14 Easily applicable data manipulation / transformation methods • PLATF_F_15 Easily applicable data cleaning methods
Phase II	<ul style="list-style-type: none"> • PLATF_F_13 (Semi-) Automatic Transformation / Mapping of data assets and extracted concepts to the ICARUS common schema • PLATF_F_18 Searchability and identification of related additional data assets • PLATF_F_20 Indication of the linkable denominators, upon which linking of the data assets can be performed • PLATF_F_21 (Semi-)Automatic data asset linking
Phase IV	<ul style="list-style-type: none"> • PLATF_F_22 Definition of simple and advanced "information" queries • PLATF_F_25 Filtering of data assets based on different criteria • PLATF_F_26 Access and inspection of data assets "extracts" depending on their license • PLATF_F_27 Transformation of a data asset to other supported data formats and export
Phase V	<ul style="list-style-type: none"> • PLATF_F_31 Automatic check whether the data asset is appropriate for a specific algorithm • PLATF_F_32 Automatic check for data licences compatibility to run under a specific algorithm • PLATF_F_35 Execution of an analytics task / algorithm according to specific preferences and settings for computation resources • PLATF_F_38 Visualization of the analytics results to gain insights on the data and / or comparison how the same results are visualized in different diagrams • PLATF_F_39 Definition of customized dashboards by selecting which visualizations should appear • PLATF_F_40 Definition of an end-to-end workflow / recipe • PLATF_F_41 Export of analytics results in machine-readable format • PLATF_F_43 Saving own projects / analysis for future reference • PLATF_F_44 Execution of scheduled analytics • PLATF_F_45 Navigation to analytics on data asset usage
Phase VI	<ul style="list-style-type: none"> • PLATF_F_48 Delivery of notifications regarding new data assets checked in, and/or existing data assets updated, related to own data assets or to analysis and visualisations performed • PLATF_F_49 Delivery of notifications regarding updates and modifications in the terms of use (e.g. licences) of data assets exploited through the platform • PLATF_F_51 Proposition of additional data assets for the enrichment of existing data assets and / or for analysis and visualisation • PLATF_F_55 Automatic license compatibility check for data assets that build on other assets • PLATF_F_56 Step-by-step guidance on how to define the appropriate license of a data asset • PLATF_F_57 Negotiation of a data sharing agreement • PLATF_F_60 Approval of a change in the terms of a data sharing agreement • PLATF_F_61 Acceptance of terms of use of a public data asset and availability to download

The results of this deliverable, including the defined ICARUS methodology, high-level usage scenarios and MVP will be leveraged as input to the use cases, the architecture and specification tasks in WP2 and WP3. In particular, this deliverable (D1.2) will feed the ICARUS deliverables D2.1 "Data Management and Value Enrichment Methods", D2.2 "Intuitive Analytics Algorithms and Data Policy Framework", D3.1 "ICARUS

Architecture, APIs Specifications and Technical and User Requirements” and D7.1 “Initial Project Exploitation Plan–v1”. Furthermore, the activities of this deliverable will continue their execution, and the updates of the work and the final results will be presented in the ICARUS deliverable D1.3 “Updated ICARUS Methodology and MVP” on M15 of the project.

Table of Contents

1	INTRODUCTION.....	9
1.1	DOCUMENT PURPOSE AND SCOPE	9
1.2	DOCUMENT APPROACH	9
1.3	DOCUMENT RELATIONSHIP WITH OTHER ICARUS WORK PACKAGES.....	10
1.4	DOCUMENT STRUCTURE	10
2	ICARUS METHODOLOGY	12
2.1	PHASE I: DATA COLLECTION	14
2.1.1	Step I.1: Data Retrieval	15
2.1.2	Step I.2: Data Anonymization	16
2.1.3	Step I.3: Data Quality Check.....	17
2.1.4	Step I.4: Data Curation.....	18
2.1.5	Step I.5: Data Check-In.....	19
2.2	PHASE II: DATA ENRICHMENT	20
2.2.1	Step II.1: Semantic Enrichment and Annotation	20
2.2.2	Step II.2: Data Linking	21
2.3	PHASE III: ASSET STORAGE.....	21
2.4	PHASE IV: ASSET EXPLORATION AND EXTRACTION	22
2.4.1	Step IV.1: Asset Indexing and Searching	22
2.4.2	Step IV.2: Asset Export	23
2.5	PHASE V: DATA ANALYTICS	23
2.5.1	Step V.1: Data Analysis	24
2.5.2	Step V.2: Data Visualization.....	25
2.6	PHASE VI: ADDED VALUE SERVICES	26
2.6.1	Asset Sharing.....	26
2.6.2	Recommendations.....	27
2.6.3	Notifications.....	28
2.7	PHASE VII: SERVICE COLLECTION	28
2.7.1	Step VII.1: Service Check-In	29
2.7.2	Step VII.2: Testing and Assessment.....	29
2.7.3	Step VII.3: Service Asset Review	30
3	ICARUS HIGH-LEVEL SCENARIOS.....	31
3.1	SCENARIO 1: DATA UPLOAD AND DOWNLOAD.....	31
3.2	SCENARIO 2: DATA UPLOAD AND ANALYSIS.....	35
3.3	SCENARIO 3: DATA UPLOAD AND SHARING	40
3.4	SCENARIO 4: SERVICE UPLOAD AND SHARING.....	44
3.5	SCENARIO 5: DATA UPDATE	47
3.6	SCENARIO 6: DATA EXPLORATION, ANALYSIS AND SHARING	50
4	ICARUS MVP DEFINITION.....	56
4.1	FEATURES EXTRACTION, INTEGRATION AND HOMOGENIZATION	57
4.2	INITIAL FEATURES VALUE ASSESSMENT	73
4.3	PRELIMINARY MVP CONSOLIDATION	75
5	CONCLUSION	78
6	REFERENCES.....	79

List of Figures

FIGURE 1-1: DOCUMENT APPROACH	9
FIGURE 1-2: D1.2 RELATIONSHIP WITH OTHER DELIVERABLES AND WORK PACKAGES	10
FIGURE 2-1: ICARUS METHODOLOGY	14
FIGURE 2-2: ICARUS DATA COLLECTION PHASE	15
FIGURE 2-3: ICARUS DATA ENRICHMENT PHASE.....	20
FIGURE 2-4: ICARUS ASSET EXPLORATION & EXTRACTION PHASE	22
FIGURE 2-5: ICARUS DATA ANALYTICS PHASE	24
FIGURE 2-6: ICARUS SERVICE COLLECTION PHASE	29
FIGURE 4-1: ICARUS MVP APPROACH.....	57
FIGURE 4-2: ICARUS MVP FEATURES ASSESSMENT BY THE 4 DEMONSTRATORS FOR OWN BUSINESS VALUE.....	74
FIGURE 4-2: ICARUS MVP FEATURES ASSESSMENT BY 3 DEMONSTRATORS AND OAG FOR AVIATION INDUSTRY BUSINESS VALUE.....	74
FIGURE 4-3: ICARUS MVP FEATURES ASSESSMENT: CORRELATION BETWEEN THE DEMONSTRATORS' OWN BUSINESS VALUE AND THE AVIATION INDUSTRY BUSINESS VALUE.....	75

List of Tables

TABLE 4-1: PRELIMINARY ICARUS MVP ON M6.....	75
--	----

1 Introduction

1.1 Document Purpose and Scope

The ICARUS Deliverable D1.2 aims at formulating the ICARUS methodology and the Minimum Viable Product (MVP). More precisely, its purpose is to reveal how concepts interrelate and to display high-level usage scenarios of the ICARUS concept. The work in D1.2 is part of WP1 and more specifically, the context of the task T1.4 “ICARUS Methodology and MVP Definition”.

One of the major outcomes of T1.4 is the ICARUS Methodology. A detailed analysis of the as-is and to-be processes is performed, illustrated as workflow diagrams for different stakeholder groups (and prospective platform users), revealing the logical flow of information and operations inside ICARUS platform. In addition, a set of diagrams are presented in order to have a better understanding of the methodology components that will be supported by the platform and will be showcased during the demonstrators’ realization. The specific diagrams are more technical and disclose the data flow across the envisaged system, regarding the inputs and the data structures needed, as well as the expected outputs for every possible process and interaction among the stakeholders. This task is based on the definition of high-level usage scenarios that support the methodology. The methodology defined leads towards the formulation of the ICARUS MVP, which covers the most important needs of the users and prioritizes the features to be transferred into implementation and deployment.

1.2 Document Approach

The deliverable follows a clear and easily comprehensive approach in order to derive the outcomes of T1.4. Figure 1-1 depicts a high-level and abstract overview of the approach followed.

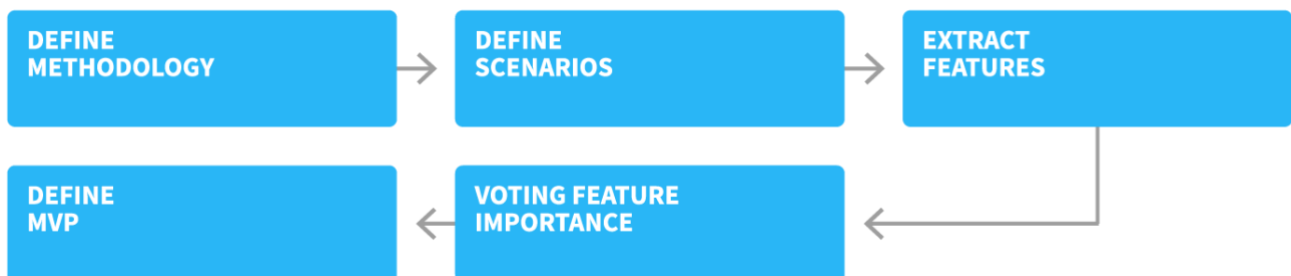


Figure 1-1: Document Approach

At first, the key findings of D1.1 have offered valuable input to the definition of the ICARUS Methodology (as presented in Section 2). The methodology is divided in various phases, with each phase having its own specific steps and aspects.

The consortium defined several high-level usage scenarios of ICARUS based on the ICARUS methodology, ICARUS demonstrators and key findings from D1.1 “Domain Landscape Review and Data Value Chain Definition”. In particular, six high-level scenarios were defined in detail, as representative scenarios of all core differentiated ICARUS stakeholders. (as detailed in Section 3).

The derived methodology along with the possible high-level scenarios and the demonstrators’ requirements have facilitated the extraction of features that could be possibly included in the ICARUS MVP. Considering these features, a voting process among the members of the consortium was organized in order to specify feature importance and lead to the selection of the ones that constitute the MVP (presented in Section 4).

1.3 Document Relationship with other ICARUS Work Packages

This deliverable (D1.2 - “The ICARUS Methodology and MVP”) is the outcome of the Task 1.4 “ICARUS Methodology and MVP Definition” which remains active until the 15th month of the project. Figure 1-2 depicts the relationship of Deliverable D1.2 with other Deliverables and Work Packages (WPs) in ICARUS.

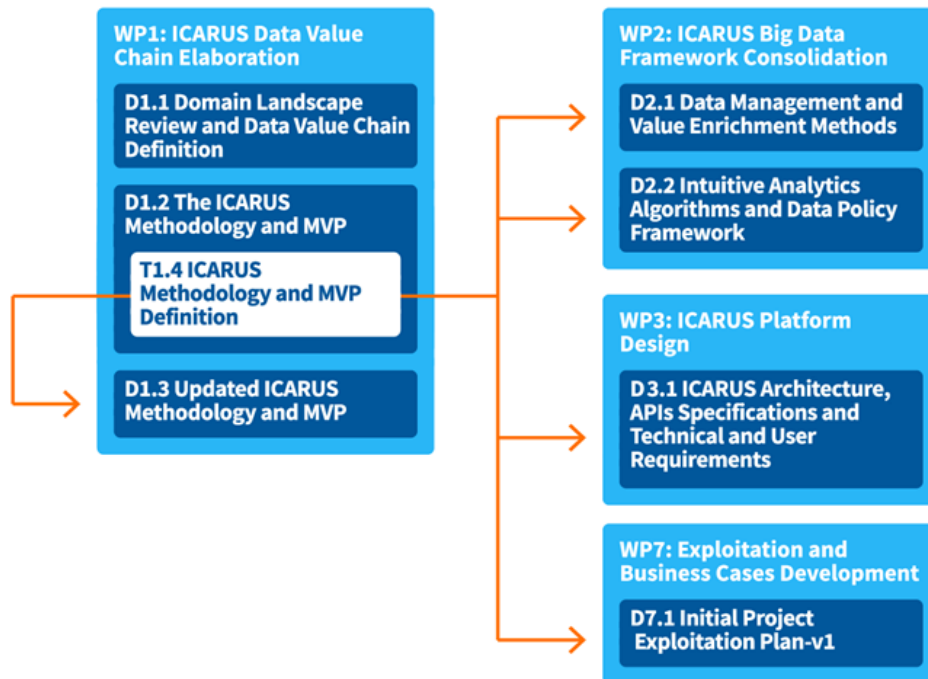


Figure 1-2: D1.2 Relationship with other Deliverables and Work Packages

Deliverable D1.3 (“Updated ICARUS Methodology and MVP”) will directly use D1.2 results to provide an updated version of the ICARUS methodology and the final version of the ICARUS MVP. Furthermore, with the defined methodology, high-level usage scenarios and MVP, this deliverable (D1.2) will provide input to the use cases, the architecture and specification tasks in WP2 and WP3. In particular, it will feed the ICARUS deliverables D2.1 (“Data Management and Value Enrichment Methods”), D2.2 (“Intuitive Analytics Algorithms and Data Policy Framework”), D3.1 (“ICARUS Architecture, APIs Specifications and Technical and User Requirements”) and D7.1 (“Initial Project Exploitation Plan–v1”).

1.4 Document Structure

The following sections of the specific deliverable are structured as follows:

- Section 2 presents the definition of the ICARUS Methodology and is divided in seven phases (Phase I - Data Collection, Phase II - Data Enrichment, Phase III - Asset Storage, Phase IV - Asset Exploration and Extraction, Phase V - Data Analytics, Phase VI - Added Value Services, Phase VII - Service Collection), with each phase having its own specific steps;
- Section 3 presents the defined high-level usage scenarios of ICARUS for different stakeholder groups, that present the logical flow of information and operations inside ICARUS platform. More precisely, it describes six scenarios (general workflow diagrams) as representative of all core distinct ICARUS user types and three examples (more technical sub-diagrams) per each scenario;

- Section 4 presents the platform features that are extracted from the methodology and are related the high-level scenarios, the initial internal assessment of the features, and the preliminary consolidation of the ICARUS MVP;
- Section 5 concludes this deliverable.

2 ICARUS Methodology

This section describes the initial definition of the ICARUS methodology which consists of a set of interactive phases. The aim of ICARUS Methodology is to provide a well-constructed and meaningful process that will guide the development of ICARUS Platform and Ecosystem and will ensure that consensus upon ICARUS offerings is reached among ICARUS partners. In addition, it illustrates how the different stakeholders (either providers or consumers) interact with ICARUS, as well as when the ICARUS administration team that has undertaken the role of a moderator, is involved in the process.

The aforementioned methodology, considers the various challenges that are presented in the Big Data Value Association (BDVA) [1] results, namely:

- **Data protection and privacy:** data protection and management must be aligned throughout the data lifecycle. Control, auditability and lifecycle management are essential for governance, cross-sector applications and compliance to E.U.'s General Data Protection Regulation (GDPR [2]). Furthermore, methods for anonymization have to be deployed for preventing the processing of Personally Identifiable Information (PII) when necessary in order to comply with laws and regulations such as GDPR.
- **Data quality:** methods for assessing and optimizing data quality have to be created, together with curation frameworks and workflows to be utilized.
- **Semantic annotation and interoperability:** datasets need to be enhanced with semantic annotation in digital formats, without imposing extra effort on data providers.
- **Data analytics:** recent progress in Machine Learning and Deep Learning needs to be considered, in order to improve the efficiency and reliability of data analytics processes for advanced business applications.
- **Interactive visual analytics:** the results produced in the analysis of datasets are not always clear in advance, and single optimal solutions are unlikely to exist. Interactive visual interfaces have great potential for facilitating the empirical inference and verification of results, by modifying the scale and the means of any aggregation.

The discrete phases of the methodology interact with the **ICARUS Data Value Chain**. In particular, the Data Value Chain of ICARUS consists of three core tiers:

- **Data Tier 1: Primary Aviation Data** comprises of aircraft sensor data, scheduled route plans, airport traffic, fuel emissions, passenger data that pile up in heaps of data in every flight. Typical data providers include airports, airlines and original equipment manufacturers (OEMs).
- **Data Tier 2: Extra-Aviation Data** features data collected by airport services providers and aviation-related service providers (e.g. drones, helicopters, etc.). Such data concern passengers' profiles (purchases in duty free shops, parking history, social media activities, etc.) which are complemented by Linked Open Data (indicatively weather, environment) and other historical data.
- **Data Tier 3: Aviation-derived Data** contains data and knowledge from businesses and organizations in other sectors such as Health, Tourism, Security industries and Public organizations (e.g. local municipalities), which can be combined with aviation data from tiers 1 and 2 to produce new derived data and create new knowledge that would be impossible to infer otherwise.

Each phase of the methodology that is presented in this section, focuses on a specific task. A brief description of the methodology phases is described below:

- **Phase I - Data Collection:** refers to the insertion of data in ICARUS, ensuring the privacy of sensitive information, the assessment and improvement of data quality, as well as protecting and safeguarding the data intellectual property rights (IPRs) through appropriate licensing.
- **Phase II - Data Enrichment:** involves the enrichment of data with machine-readable metadata and additional information about its entities and relationships, based on aviation-specific ontologies, vocabularies and semantic models. It also refers to the linking of data assets with other relevant sources that exist in ICARUS.
- **Phase III - Asset Storage:** describes the appropriate aspects and challenges that need to be considered in order for ICARUS to host the data and service assets in a secure and efficient manner.
- **Phase IV - Asset Exploration and Extraction:** includes the indexing of assets that will be hosted in ICARUS, as well as the asset searching and exporting mechanisms, so as to explore data and service assets.
- **Phase V - Data Analytics:** refers to the analysis of data assets upon which state-of-the-art data analytics algorithms can be applied, as well as visualization approaches through respective charts and plots.
- **Phase VI - Added Value Services:** consists of an asset sharing mechanism that is responsible for sharing assets between asset providers and consumers, as well as recommendation and notification services that aim to enhance asset discoverability and assist the stakeholders.
- **Phase VII - Service Collection:** describes the process in which the stakeholders (e.g. IT aviation companies) can register their own implemented algorithm that will be translated into a service asset in ICARUS.

The phases are not to be considered in a sequence despite their numbering, as these phases interact with each other and based on each stakeholder's objectives a different subset of them may be mobilized. Figure 2-1 illustrates an overview of the ICARUS methodology that brings together the holistic ICARUS perspective whose many aspects will be further elaborated in the frame of WP2.

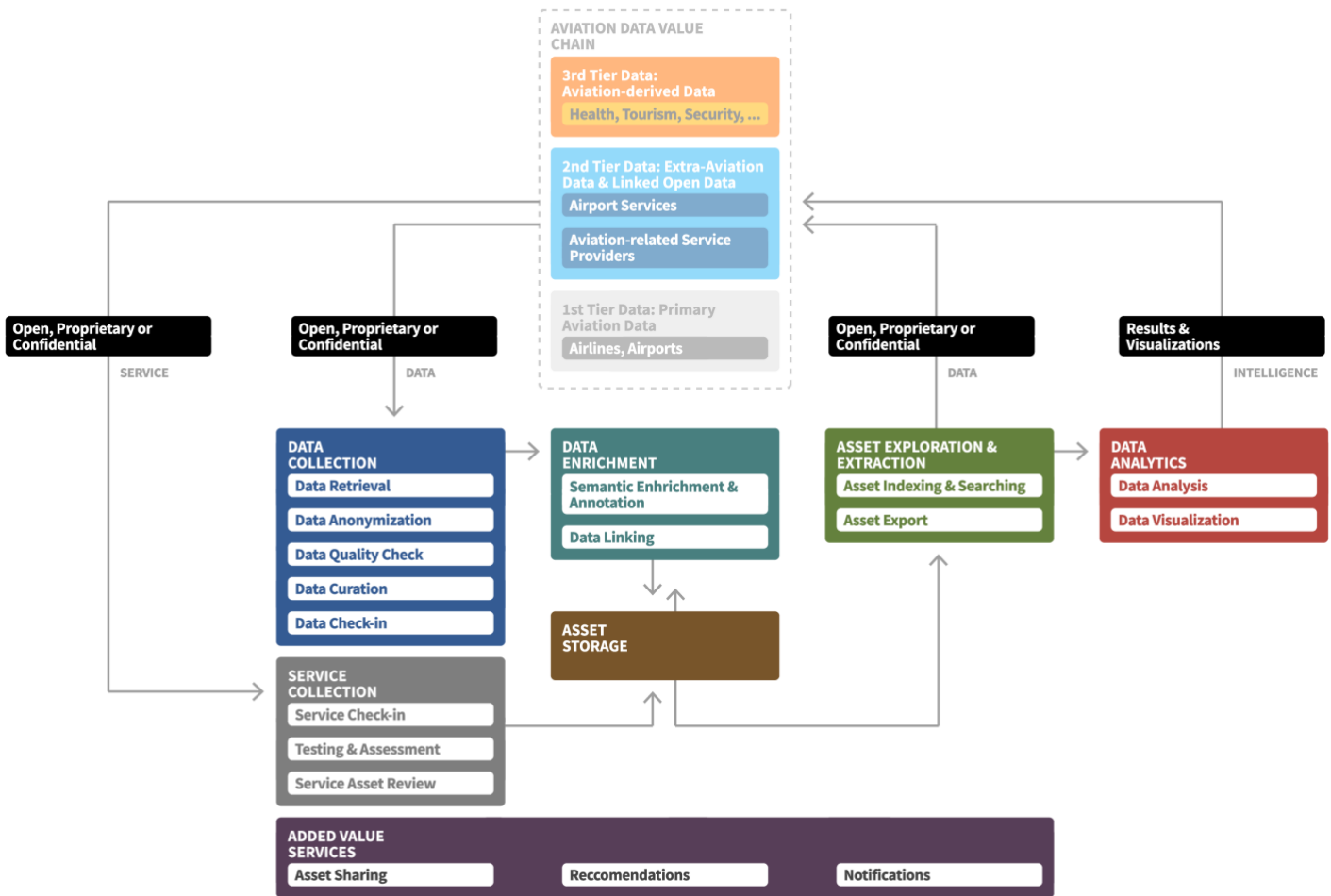


Figure 2-1: ICARUS Methodology

2.1 Phase I: Data Collection

Data Collection is the phase of the ICARUS Methodology that is responsible for accessing and aggregating datasets from various sources across different aviation-related domains. Such data can be either open data to the public (e.g. weather data) or private data owned by businesses and organizations (e.g. passengers' data owned by airports or airlines). Specifically, Data Collection consists of mechanisms that facilitate data accessibility and retrieval from various sources in a manner that enables privacy protection of the individuals described in the data through data anonymization. Furthermore, the specific phase is responsible to perform data quality checks of the data to ensure the integrity and veracity of the data, as well as data curation mechanisms to filter and clean the data from inconsistencies and errors. Finally, the data need to be registered (checked-in) in a way that data policy definition is obeyed and compliant with the data provider's IPR.

During this phase, data providers are directly involved, since it provides the procedure that facilitates the insertion of data to ICARUS platform. It should be stated that in this phase, the data are not permanently stored in the ICARUS storage (as it is addressed in the Phase III of Asset Storage, Section 2.3), but on the contrary, they are temporarily stored throughout these steps. In this phase, the visibility of data is restricted to the data providers only.

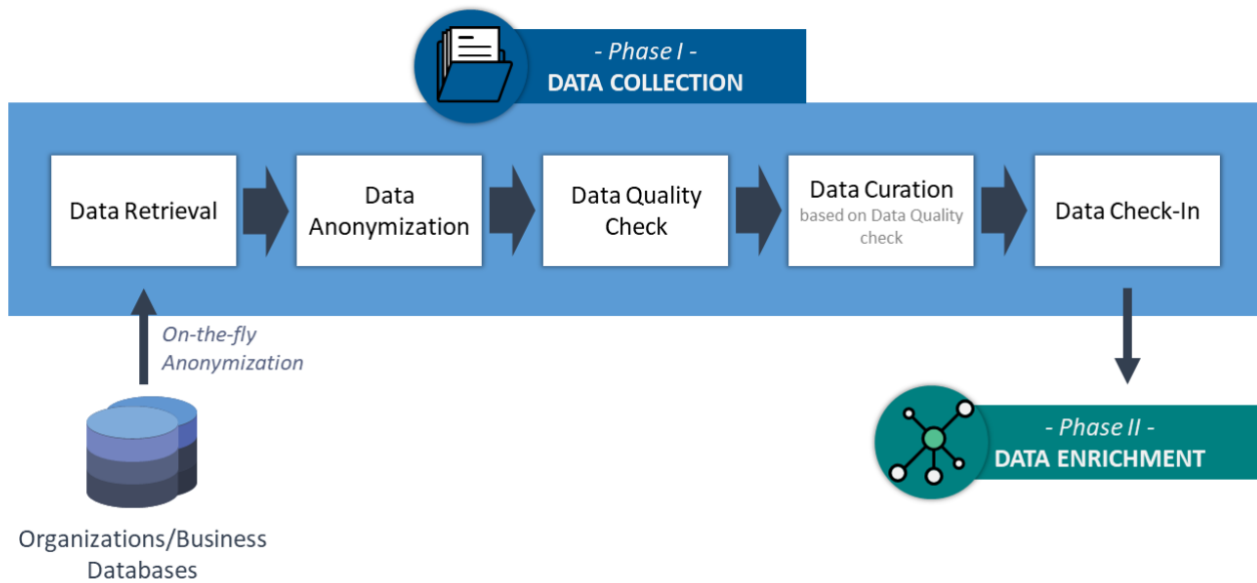


Figure 2-2: ICARUS Data Collection Phase

2.1.1 Step I.1: Data Retrieval

Data Retrieval corresponds to the initial step of the Data Collection phase that provides the required technical bridges for accessing and aggregating public or private data. Businesses and organizations in the aviation sector, store their data to private Database Management Systems (DBMSs). For the purpose of retrieving the specific data, various web services can be used such as REST APIs and SOAP, as well as mechanisms a) for uploading files in various supported formats (e.g. CSV, JSON, XML, etc.) manually or b) for retrieving open data from provided links.

Data retrieval in ICARUS can be ensured through the implementation and deployment of RESTful APIs that have gained significant traction over the last years. In order for an API to be directly usable in ICARUS, it requires to be well-designed following the de-facto design principles and guidelines that have been generally adopted, e.g. in the W3C Web API Design Cookbook [3], in the Apigee Web API Design Guidelines [4], etc. Although the latest trends in APIs (indicatively featuring hypermedia, GraphQL, GRPC) should be considered for data retrieval, any API that complies with the principles of readability, simplicity and thorough documentation, should be potentially appropriate in ICARUS.

An important consideration is the different policies for releasing an API. ICARUS needs to support partner APIs which can be used only by specific business partners, as well as open APIs which are available to the public. Additional concerns that need to be properly addressed for APIs prior to their inclusion in ICARUS include:

- **API usability:** APIs need to be easily comprehensive and efficiently utilized, providing sensible error messages when required.
- **API versioning:** when new features are added or modifications are made to an API, there must be sufficient provided period of time for transition to the latest version and also, old versions need to be supported (at least for a period of time).
- **API availability:** APIs need to ensure uninterrupted availability, so as the services provided by the APIs to be properly delivered.

2.1.2 Step I.2: Data Anonymization

Data Anonymization is a critical step in the phase of Data Collection as it has become a vital challenge, taking into account the compulsory compliance with emerging data regulations. Examples of such regulations are the GDPR in the European Union [5], Cybersecurity law in China [6], regulation for protecting the privacy of customers of broadband and other telecommunications services in the USA [7].

In ICARUS, it is very important to ensure privacy when sharing business critical data, as well as personal private data of individuals (e.g. passengers' information). Towards this direction, the primary objective of Data Anonymization is to prevent the inference of business-sensitive information and potential associations deriving from the data so as to allow the stakeholders to take advantage of the data without endangering the business confidentiality or compromising the privacy of individuals. Through the specific data anonymization step, a malicious party will not be able to infer anything private, while a honest party will be able to analyze the information inside the data. However, it should be noted that if the data is confidential (i.e. data will be only utilized for private purposes by the data provider), without sharing or allowing anyone else to view the data, while being this the case, the step of anonymization is neither relevant nor applicable.

In order to address the privacy issues in ICARUS, the attributes that may compromise the privacy will be identified. The most critical personally identifiable information (PII) are the attributes that directly define an individual (unique attributes e.g. ID number, passport number, etc.) [8]. However, data regulations like GDPR have indicated that the definition of PII is evolving, as technology has been capturing much more information than these unique attributes. Therefore, the quasi-identifiers of an individual, i.e. attributes that are not unique per individual but can be combined with other information to create a unique identifier (e.g. gender, age, address, etc.), as well as other sensitive information must be also considered.

Taking under consideration the fact that ICARUS project intends strongly to provide analytics while maintaining the privacy of sensitive information, there are plenty of anonymization approaches that may be considered. Thus, the values of various attributes can be replaced by more generic ones (generalization e.g. the age of a passenger who is 25 years old can be replaced with age range such as 20-30 years old) and unique attributes can also be replaced by using encryption techniques (pseudo-anonymization). Additionally, there are various techniques that may add randomization and noise to the data such as permutation (i.e. shuffling the values of attributes), perturbation (i.e. adding noise to the values of attributes), differential privacy (i.e. adding noise to query response), etc. Moreover, there are various anonymization properties in order to ensure the level of anonymity of the resulting anonymized dataset such as k-anonymity, l-diversity and t-closeness [9].

There is a diversity between the aviation-related stakeholders that ICARUS aims to reach, as these stakeholders vary among large enterprises which have experience with data anonymization and managing big data in general (e.g. airports, airlines), but there are also unexperienced stakeholders or SMEs (e.g. small travel agencies) without in-place mechanisms for this purpose. Therefore, ICARUS will consider the following approaches to achieve the anonymization of data:

- Aviation-related businesses and organizations can possibly have in-house mechanisms which comply with their internal policies for sharing sensitive information, in order to anonymize their data before uploading it to ICARUS. In this case, the responsibility of anonymizing the data is on the data providers' side.
- The data providers without in-house anonymization mechanisms may need to install locally an offline tool for lightweight anonymization provided by ICARUS, in order to anonymize their dataset at a batch

fashion, prior to uploading. Otherwise, the data can be uploaded in ICARUS in order to be anonymized using online mechanisms, which may require more computational resources. ICARUS may provide both options and the data providers will need to choose based on their internal policies. Furthermore, the data providers will have the ability to create manually a “check-list” of anonymization tasks.

- In the case of streaming data, ICARUS may provide semi-automatic mechanisms for anonymizing the data on-the-fly, i.e. a sequence of processes which act as a pipeline during the uploading of the data, so as for the data to be anonymized at the same time that the uploading process is being materialized.

Even though the main challenge during the specific step is the data protection and privacy, another important challenge is to maintain the usefulness of the datasets (i.e. the usefulness and usability of the data) without affecting privacy. Anonymization is worthless if the data usefulness is not considered and all meaningful information is lost. For example, an empty dataset would have perfect privacy, but no utility, while on the other hand, the original dataset would have full utility, but no privacy. This is also known as the “privacy vs. utility tradeoff”.

Furthermore, an important fact to be considered before proceeding to the anonymization process is the nature of the data per se [10]. The most common types that need to be considered in ICARUS are:

- **Relational data** (aka. tabular data) represents information about entities (e.g., passengers), their characteristics (quasi-identifiers e.g., age, address) and other sensitive information.
- **Transactional data** associates people with the sets of items purchased in a transaction. Unlike typical relational data, transactional data involves multiple entities and can have variable lengths and higher dimensionality. For this reason, transactional data is more difficult to be anonymized, as it has unique properties; for instance, transactions involving multiple items, with each item having its own attributes.
- **Graph-based data** that represents sensitive associations between entities (e.g., people in social networks). This is even more difficult as many pieces of information can be utilized to identify individuals in a graph like the labels of vertices and edges, neighborhood subgraphs and their combinations.

2.1.3 Step I.3: Data Quality Check

Data Quality Check is the process that aims to discover inconsistencies and other anomalies in the data through validation rules and various statistical checks, in order to ensure the integrity and reliability of the dataset. Nowadays, data quality checking [11] is a continuous process, as businesses and organizations operating inside the aviation ecosystem are collecting massive amount of data that derives from multiple sources. Having data perfectly complete and accurate is not only time consuming but also prohibitively expensive as decisions have to be made effectively and efficiently [12]. Instead, the data needs only to meet the quality standards that have been set for it. In ICARUS, the process of Data Quality Check will be semi-automated. In case if the data provider intends to utilize the dataset for private use only, the specific step becomes optional.

In order to check the data quality, various dimensions must be considered such as:

- **Accuracy:** the degree to which data reflects the “real world” object correctly (e.g. passenger’s age is incorrect);
- **Completeness:** the proportion of stored data against the potential of “100% complete” (e.g. critical data such as passenger names must have no missing values);
- **Consistency:** the absence of difference when comparing two or more representations of an object against a definition (e.g. date format is MM/DD/YYYY in the USA and DD/MM/YYYY in Europe);

- **Timeliness:** the degree to which data represents reality from required point in time (i.e. data can be valid for a specific time period and must be up-to-date when published);
- **Uniqueness:** nothing will be recorded more than once, based upon how that object is identified (i.e. duplicate entities having different attribute values);
- **Validity:** data is valid if it conforms to the syntax (format, type, range) of its definition

The starting point of this semi-automatic process is the data profiling which is required by the data provider, in order to initially assess the data quality, including the particular standards that the data must conform to. Towards this direction, validation rules can be defined to verify that the dataset meets the standards that the data provider has specified. For example, automatic verification rules can ensure that the data has no missing values or duplicates. Other validation rules may include range constraints (i.e. the values must range between predefined limits), value representation constraints (e.g. a UK phone number contains exactly 10 digits, excluding the prefix code numbers) and data type constraints (e.g. an attribute is set to contain only integer numbers, but it appears to contain real numbers as well). In addition, automatic cross-check rules can be used to parse the data and the metadata to identify inconsistencies based on the data provider's description. For instance, the temporal coverage of the dataset in the metadata may be between 2015 and 2018, but after a quick data parsing, it could be indicated that the data contains timestamps within 2014. Finally, it is possible to have a "crowd" quality assessment, in which data consumers that have already used/purchased the data can provide feedback about the data quality.

2.1.4 Step I.4: Data Curation

The Data Curation constitutes the active and ongoing management over the data lifecycle and aims to ensure that the necessary data quality levels are met for maximum usefulness and usability. In particular, it incorporates: (a) data cleaning, correcting or removing the dirty or coarse data in order to reduce the content of noise or errors and unmask important features from the data; (b) data filtering to filter/remove irrelevant or unnecessary data.

Generally, this step is highly connected with the previous step of data quality check (Step I.3, Section 2.1.3). Based on the detected errors and inconsistencies of the data quality check, the data provider may select different data cleaning methods to be applied upon the dataset. Furthermore, guidelines or suggestions for improving the data quality may be provided by ICARUS to the data providers. As in the previous steps, the specific one of data curation may be optional if the data provider does not desire to share the data with others or if the data quality check mechanism of ICARUS (Step I.3, Section 2.1.3) does not identify any inconsistencies or errors inside the dataset.

After the data quality check, the data transformation workflow must be defined by the data provider. Depending on the volume, the degree of heterogeneity and the quality of the data, many data transformation and cleaning methods [13] may be applied. The most popular methods for data cleaning are data transformation (e.g. replace, reduce, etc.), missing-data imputation, data normalization, data deduplication, outlier elimination and syntax errors correction [14]. However, even after the different curation methods are applied, the data quality must be re-evaluated. Multiple iterations of all steps may be required, as some errors only become apparent after applying the different curation methods.

In the aviation sector, there are many challenges when dealing with data curation. Data variety that may arise based on the data creation within different contexts and with different requirements, may lead to a critical issue

of data curation. This is the specification of a cleaning workflow that may be applied to dirty data so as to eliminate anomalies. Furthermore, considering that the aviation data providers have high volume of data, performing a data curation process can be computationally intensive, especially if 100% complete elimination of the anomalies and errors is desired [15]. In addition, another possible issue that needs to be considered is the data usability, as the various curation methods that may delete or transform part of the data, may cause information loss. Therefore, data curation is not a straight-forward process and it is often underestimated by data providers.

2.1.5 Step I.5: Data Check-In

The Data Check-In step will facilitate the registration of the data assets in ICARUS. The registration will be accompanied by the definition of the assets' metadata regarding the data profiling and the data asset rights to ensure the safeguarding of the asset's IPR (intellectual property rights) through the appropriate licensing. In particular, the data licenses will effectively expand or restrict the actions that a data consumer is allowed to perform upon each dataset and grant permissions based on the idea that certain terms have to be met.

In D1.1 "Domain Landscape Review and Data Value Chain Definition", an initial data profiling template has been presented containing metadata to appropriate levels of detail, in order to describe the data assets. The template contains details about the data volume, variety, veracity, type, format, as well as the speed at which data is generated or updated. Additionally, it contains details about the availability of historical data along with their frequency, the data temporal and spatial coverage, the language of the data, the relevant standards to which the data comply, the existing dependencies to other sources, the data asset availability (ownership and mechanisms to access the data) and the data asset rights (privacy aspects, license, pricing, need of anonymization). This kind of information would be appropriate to be provided during the specific step by the data provider.

The information presented in the data profiling template is critical, in order to properly handle the data assets. A major consideration is the privacy aspects that set the data visibility, as the data can be **confidential** (not to be shared) and used only by the data provider, **proprietary** that can be shared with the appropriate licensing and/or after purchasing, as well as **open** for the public. In particular, depending on the defined data visibility (public, private or confidential), different levels of detail may be required for the data profiling. Furthermore, confidential data assets may have the ability for "fast lane" check-in.

Another important consideration is the updating policy that the data provider is requested to define, specifying the frequency of the updates and the rights of each data consumer that has purchased a previous version of a specific dataset. Furthermore, the data provider may need to define the timeframe of data, defining the time period for which the data asset is valid and if the dataset needs to be deleted after a period of time. Finally, a data provider may have the option of creating a sample of the dataset, in order to make it available for potential buyers so as to allow them to experiment with the data prior to purchase, in case the dataset fulfils their needs.

After the previous steps of ensuring the data quality and the anonymization of the data, the specific process prepares the data so as to proceed to the next phase of Data Enrichment (Phase II, Section 2.2). However, depending on whether the data will be shared or not, the above mentioned steps may become obligatory and the data provider is responsible to complete the specific steps before proceeding to the next phase.

2.2 Phase II: Data Enrichment

The specific phase is responsible for the semantic annotation and linking of the data with other relevant data assets. More specifically, during the specific phase, the platform will provide the required semantic annotation mechanisms for enriching the data by utilizing concepts from aviation-related vocabularies, ontologies and semantic models. Furthermore, ICARUS will provide mechanisms for linking data with other similar datasets, either open or private data (depending on the data license), in order to enrich the information and make the data assets more useful.

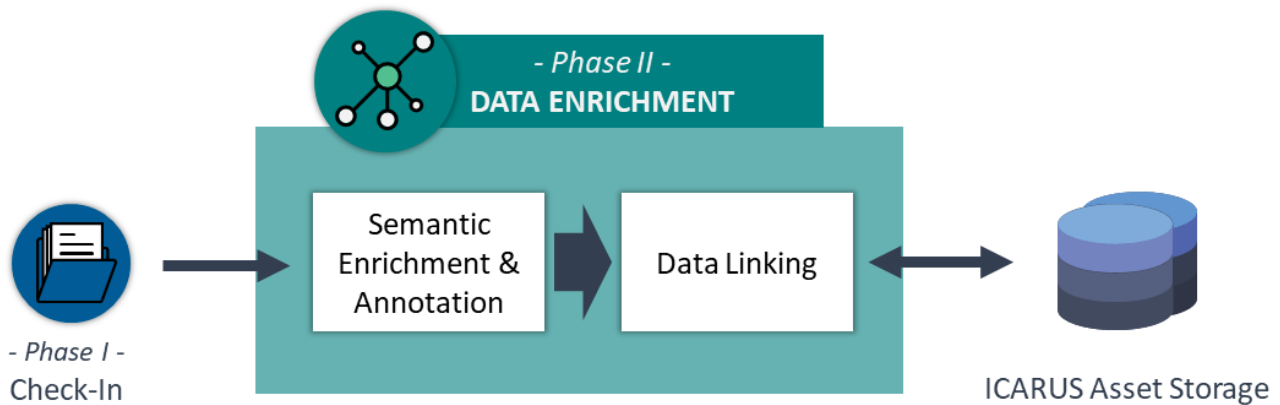


Figure 2-3: ICARUS Data Enrichment Phase

2.2.1 Step II.1: Semantic Enrichment and Annotation

During the Semantic Enrichment and Annotation step the users of ICARUS platform are offered with the required modalities for attaching additional information on various concepts (e.g., people, locations, organizations, etc.) to the data. This step follows directly after the data check-in (Step I.5, Section 2.1.5) to facilitate the enrichment of the data assets with machine-processable information by linking background information from aviation-related vocabularies, ontologies and semantic models (like the IATA's Airline Industry Data Model [16] and the NASA's Advanced Air Traffic Management Ontology [17]). This additional semantic information provides an aviation-oriented extra flavor to the data assets and the relationships that link them, making the concepts unambiguously defined and related to the aviation sector, as well as alleviates their maintenance, interpretation, and reusability.

As soon as the data asset is checked-in, a semantic data model [18] may be semi-automatically extracted, and mappings can be performed to popular and standardized vocabularies. ICARUS may assist by providing suggestions to the mapping, especially with the support of the ICARUS aviation semantic model that will be defined by the end of WP1 activities. In particular, concepts can be semi-automatically extracted from the data (e.g. ATH stands for Athens International Airport) and then mapped to the vocabularies and ontologies of ICARUS collection. Furthermore, the extracted concepts can be classified as people, locations, organizations and so on, and then, they can be unambiguously identified and linked with broader context according to domain-specific knowledge. As for the relationships between the extracted concepts, they can be identified and further interlinked with related known and newly recognized entities, enriched with machine-readable data.

However, the nature of the aviation data imposes many challenges in the tasks of semantic enrichment. These aviation data sources provide enormous amount of data; hence, the annotation must focus on selecting only the relevant data to be semantically enriched without considerably increasing the volume of the original dataset.

2.2.2 Step II.2: Data Linking

The Data Linking step encloses the utilization of methods that interlink datasets with other data sources in order to create new datasets. The specific process aims at identifying a) different datasets that refer to the same real-world object in a specific domain or b) pinpoint a relationship between them. The linking may be performed in datasets provided by different data providers or by the same data provider within ICARUS. The data can be linked directly at 1st level (e.g. dataset A is linked to dataset B) and indirectly at 2nd level (e.g. if dataset B is linked to dataset C, then a potential linking between datasets A and B will involve dataset C). The resulting linked dataset contains data from each of the source datasets and the mapping can be on either the “row” (i.e. a single structured data item in a table) or the “column” (i.e. a set of data values, one value for each data item) levels. Furthermore, before linking different data assets, a stakeholder may receive an indication of the linkable denominators (e.g. variables) upon which a semi-automatic data asset integration (data linking) can be performed.

The most critical challenge regarding data linking in ICARUS is privacy and licensing of the data. The licensing of confidential and proprietary data need to be carefully considered before linking with other data assets, so as to prevent any violation of the data providers’ terms and conditions. Furthermore, the variety of the data assets needs to be considered as the data assets may be described and shared with different identifiers from different data providers, hence, the identification of relevant data assets becomes more difficult.

The process of data linking involves the assessment of the dataset compatibility with other datasets that exist within ICARUS. In particular, the assessing of linking compatibility incorporates the degree of similarity of different data assets as an evaluation measure. This is under the assumption that the higher similarity between two data descriptions, the higher the probability that the two descriptions actually refer to the same object. The construction of the identity link that determines the data linking of the entities is referred as instance matching [19] and consists of three main techniques:

- **Value matching:** These techniques focus on the determination of equivalence between property values of instances.
- **Individual matching:** These techniques focus on deciding whether two different entities represent the same object or not. These techniques compare two different entities and utilize the results of the value matching technique applied on the property values of the specific entities.
- **Dataset matching:** These techniques focus on all the individuals of the datasets under investigation and try to construct an optimal alignment between the whole set of individuals. These techniques take as input the results of individual matching and introduce several improvements by applying optimization algorithms, similarity propagation and more.

2.3 Phase III: Asset Storage

Within ICARUS, data and service assets will be hosted, after the data and service collection phases (phases I and VII in sections 2.1 and 2.7 respectively). The assets will be stored along with their metadata in a structured fashion, as well as log files that record the stakeholders’ historical usage of assets (e.g. assets viewed / used /

purchased, analytics performed in data assets, etc.). This information may be useful for the asset providers in order to gain information about the usage of their assets.

With the intension of storing the various assets, many challenges need to be addressed. The aviation-related data is characterized by high volume and thus, scalability is an essential property for ICARUS, so as to handle large amount of data. In addition, the data can be produced in high velocity and can be provided by many providers, hence, high performance and efficiency is crucial to handle the high-traffic requirements. Furthermore, a key issue is the privacy of the assets, as not all stakeholders will have the same access privileges (e.g. stakeholders that have not purchased or agreed terms for proprietary assets). Therefore, it is vital to restrict unauthorized access, either for stakeholders without the appropriate permissions or malicious parties that may try to access the assets. Besides that, an important consideration is how to manage data that are not to be permanently stored (e.g. private data) or data retrieved from 3rd parties. This will be based on the data providers' policies for sharing business critical data and ICARUS needs to support it. Moreover, replication of data assets can be considered for fault tolerance, if the data providers allow for it. Finally, another important consideration is the data assets updates and versioning.

2.4 Phase IV: Asset Exploration and Extraction

This phase is strongly connected with the previous phase of Asset Storage (Phase III, Section 2.3) as it includes the indexing of assets that will be hosted in ICARUS platform, as well as the asset searching and exporting mechanisms, through ICARUS storage mechanism. The specific phase is responsible for the efficient exploration of data and service assets as well as for the facilitation of the extraction of these assets from the ICARUS asset storage, if the stakeholder has the appropriate permissions.

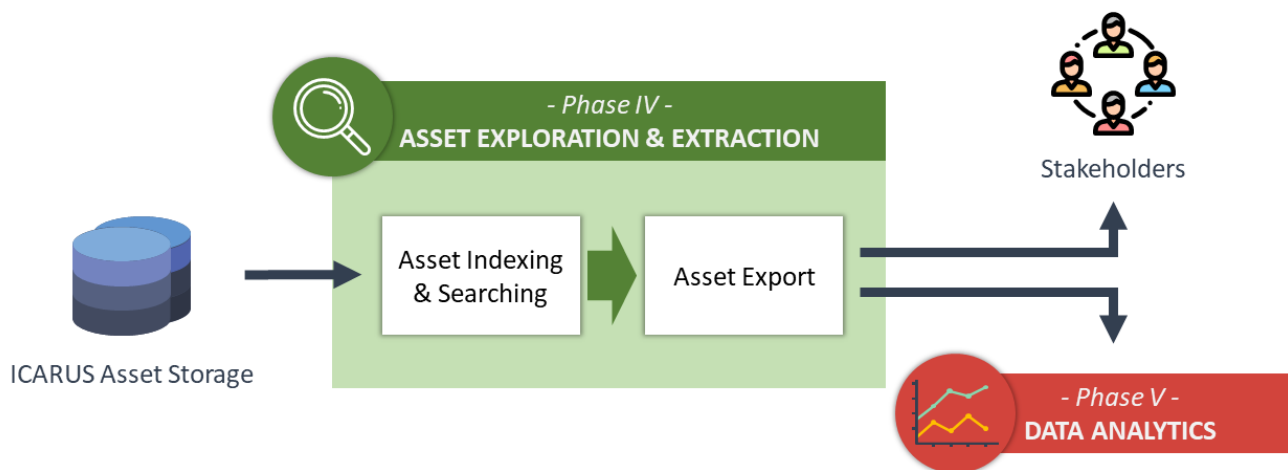


Figure 2-4: ICARUS Asset Exploration & Extraction Phase

2.4.1 Step IV.1: Asset Indexing and Searching

This step aims to facilitate the retrieval of assets from the ICARUS asset storage. The first part of the specific step incorporates the data indexing which is responsible for the creation of an index of the data assets and their metadata. Furthermore, based on the type of the data, different types of indexes should be used for maximum efficiency. Some of the most common types of indexes will be utilized: Bitmap index, Dense index, Sparse index and Reverse index.

The second part of the step is responsible for the searching in the ICARUS asset storage and the retrieval of assets based on the search queries provided by the stakeholders. Each stakeholder is able to search for open or proprietary assets, as well as his/her own assets. Before retrieving the asset, the stakeholder is able to view its license and detailed description, before proceeding accordingly (e.g. export, analyze, purchase the asset). In addition, the stakeholders can include pins and favorites among the ICARUS data and service assets.

The searching service needs to support many different types of queries, so as to enable stakeholders to easily explore the ICARUS repository and enhance data discoverability. Each stakeholder can search for a specific asset, providing its exact name or its asset provider. On the other hand, the stakeholder can be able to explore all the relevant options that exist in the repository, providing appropriate keywords based on the stakeholder's interest. Additionally, a stakeholder may desire to perform a query based on specific criteria that are not relevant to the content. For instance, a stakeholder may search the repository looking specifically for open or proprietary assets only, or even assets that have a specific license, as well as assets that their price lies between specific ranges. Moreover, this service may support structure-based queries that are related to the asset profiling and metadata, allowing a stakeholder to search for specific features of the assets (e.g. specific locations).

2.4.2 Step IV.2: Asset Export

During Step IV.2 the stakeholders of ICARUS ecosystem are able to export assets and download them, through the available APIs, in order to use them outside ICARUS. The design of such APIs is very important in this step, as this will provide the stakeholders with efficient mechanisms to export the assets. Furthermore, the stakeholders can transform a data asset to another data format (e.g. from CSV to XML) before downloading it.

The stakeholders are able to export assets only if they own the assets or if the assets are open to the public. Additionally, the stakeholders are able to export the assets that they have purchased, but only if they have the right to download them, without violating the license scheme offered by the asset provider. The latter will be decided by the sharing service that is described in Phase VI in Section 2.6.1 (Asset Sharing), which is responsible to supervise the asset exchange activities and generate micro-mutual contracts between the asset providers and consumers.

2.5 Phase V: Data Analytics

The phase of Data Analytics is utilized for the analysis of data, using statistical processing, data mining, machine learning and knowledge extraction techniques, so as to identify useful sets of relationships and trends. It is responsible for leveraging data to extract meaningful insights in order to enable evidence-based decision making. This phase consists of techniques to transform and analyze the data in order to acquire knowledge or solve various tasks (e.g. classification, regression, etc.). Additionally, it incorporates a set of visualization capabilities to better understand the data and the patterns that are derived from the datasets. These steps of the process are utilized iteratively, as each iteration may produce new insights that require further analysis.

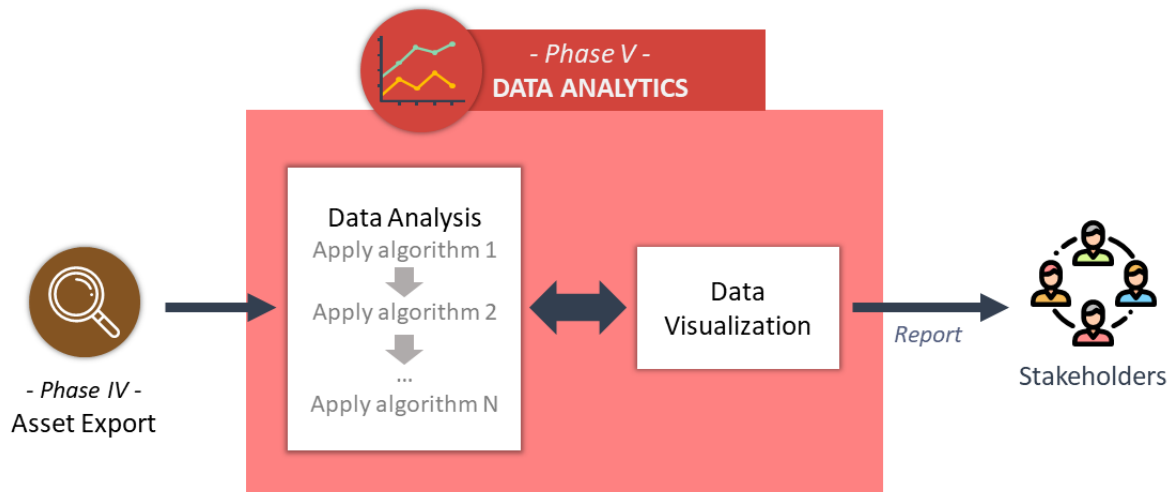


Figure 2-5: ICARUS Data Analytics Phase

2.5.1 Step V.1: Data Analysis

In the step of Data Analysis in ICARUS, a data asset or even a data collection of linked data assets is analyzed in order to extract meaningful information and knowledge. A stakeholder may select from a variety of predefined algorithms (e.g. logistic regression, gradient tree boosting, etc.), as well as statistical methods (e.g. variance of the data, correlation and covariance between variables, etc.) to perform various tasks (e.g. classification, regression, clustering, statistical analysis etc.) according to specific preferences (e.g. accuracy, efficiency) and settings of computational resources, in order to extract insights from the datasets. These algorithms can be either well-known algorithms or variations that have been specifically designed for the aviation sector in ICARUS. Regarding the computational resources, a stakeholder can select to deploy and customize a secure experimentation space in ICARUS in order to analyze his data assets. When the analysis is complete, the stakeholder can erase the secure experimentation space including all the traces of confidential data.

Once a stakeholder selects a specific algorithm, an automatic check may be performed in order to verify that the selected algorithm is appropriate for the specific data asset and is compatible with the data licenses. In addition, a stakeholder may use his/her own custom implemented algorithms, after the approval by a member of the ICARUS administration team or can use custom algorithms provided by other users (more details about service assets in Phase VII, Section 2.7). Furthermore, a stakeholder can define a schedule (either incremental or full scheduled) in order for a data analytics task to be executed on a periodic basis. Furthermore, the analysis of data can create new datasets that can be stored and shared in ICARUS or exported from ICARUS platform (via an API, or as CSV, JSON etc.), as long as it does not violate any license terms of the original data that were utilized in the beginning of the analysis. Finally, a stakeholder can have the ability to compare the results of various algorithms executions.

A typical process of data analysis may include multiple algorithms and techniques which are applied on the data. The data consumers can build a chain of algorithms in which different algorithms are applied sequentially and the input in each algorithm is the output of the previous execution. Its purpose, is to perform a series of transformations of the data in order to produce the expected outcome. For instance, for a classification problem that uses a dataset containing both numerical values and textual content, there are important transformations that need to be realized prior to applying a classification algorithm (e.g. random forests classifier). The numerical data need to be normalized in order to scale between the same ranges (e.g. between 0 and 1), since if the data

attributes scale between different ranges, the specific fact may affect the classification accuracy negatively. As for the textual data, it needs to be transformed into numerical vectors, as classification algorithms require the input to be mapped into numerical variables.

A critical challenge in this step is the fact that the aviation data value chain is characterized by high dimensionality and large volume of data. These two features cause issues such as high computational cost and algorithmic instability, creating the requirement for more computational resources. Furthermore, due to the fact that data comes from various sources, it is possible to contain noise (despite the Step I.4 of Data Curation, Section 2.1.4) and create issues of heterogeneity, experimental variations and statistical bias.

2.5.2 Step V.2: Data Visualization

Data Visualization involves the creation and study of the visual representation of data. In general, people are able to process visual information much faster than textual content, as the results of machine learning methods are not easily interpreted or comprehended, especially for business users without experience in data science. For this reason, Data Visualization [20] is an essential step of the methodology, aiming to enable the representation of complex information in a way that is easier to interpret, using visually engaging images of graphical plots and charts. In addition, a stakeholder may define and store custom dashboards by selecting which visualization should be presented. Furthermore, a stakeholder can create aviation data value chain reports that involve intuitive reporting of the results produced by the analysis (Step V.1: Data Analysis) and interactive visualization capabilities of this step. These reports can be utilized either for private use of the data consumers or for sharing knowledge with other stakeholders, as it is addressed in the Asset Sharing service of the Added Value Services phase (Phase VI, Section 2.6.1). Furthermore, the reports may be exported as downloadable files (e.g. PDF) (Phase IV, Section 2.4.2). Finally, a data provider can navigate to analytics about the usage of her data assets (e.g. “how people use my data” etc.).

The typical process of the visual analytics in ICARUS will consist of an iterative procedure of selecting specific data, forming a hypothesis, selecting a visualization technique supported by ICARUS and interacting with the visualizations. There are many techniques of visualizing data and each one depends on the stakeholders’ goal. For example, the relationships and mutual impact between specific elements can be visualized using scatter plots. Additionally, line graphics may be utilized so as to illustrate the time evolution of a specific phenomenon or to compare two or more features with respect to a specific variable.

Data Visualization is not a simple process as there are many challenges to be overcome. Considering that the aviation-related data is characterized by high complexity and high dimensionality, it becomes even more difficult to visualize data.

For this purpose, there are many dimensionality reduction techniques [21] which may be utilized to transform the high-dimensional data to a space of fewer dimensions. In particular, these techniques can be divided into two categories:

- **Feature selection:** These techniques focus on selecting a subset of relevant features.
- **Feature extraction:** These techniques focus on building a new set of features from the original feature set that still contain most of the useful information.

However, the dimensionality reduction techniques may not always be effective due to information loss. Furthermore, the heterogeneity and diversity of this data create visual noise, making the objects in the data too relative to each other and thus, it is difficult to visually separate them.

2.6 Phase VI: Added Value Services

This phase encloses additional services that ICARUS aims at providing. More precisely, it involves a secure and trustworthy sharing service that is responsible for sharing both data and service assets (either for free or under payment scheme), ensuring the asset provider's IPR. Furthermore, it consists of a set of recommendation services to assist the stakeholders and enhance the asset discoverability. Additionally, this phase includes notification services to notify stakeholders about asset requests and updates. These are not sequential steps, but they are additional services that are provided throughout all methodology phases.

2.6.1 Asset Sharing

Asset Sharing is a service that aims to link asset providers and asset consumers. In particular, the Asset Sharing service may utilize a blockchain-based sharing framework (e.g. [22]) to generate micro contracts, in order to facilitate and ensure the secure and trustworthy exchange of assets with respect to the sharing modalities under which the specific assets were provided. These modalities refer to confidential assets that are not for sharing with others, proprietary assets that are private and every interested stakeholder must purchase them, as well as public assets that are open to everyone. Furthermore, if a data/service consumer wishes to acquire a public asset, then it must accept the terms of use of the asset. Additionally, if an asset provider modifies a data sharing agreement, then the Asset Sharing service is responsible for checking and approving the modification.

Regarding the assets, they can be either data or service, as well as reports that were produced after performing analytics and visualization tasks on available data (more details in Step V.3 of Data Visualization, Section 2.5.2). Furthermore, the Asset Sharing service may provide the ability to the stakeholders of automatic renewal or cancellation of a data sharing agreement. In addition, it may provide an assessment of the reputation of the data assets by confirmed buyers/users.

As described in Step I.5 of Data Check-In and Step VII.1 of Service Check-In (sections 2.1.5 and 2.7.1 respectively), the asset providers should register their ownership policies as well as the required annotations. The previously mentioned steps ensure that the assets are fully described with respect to their IPRs, as well as privacy and quality through data anonymization and data quality validation methods (Step I.2 and Step I.3 respectively). These preconditions are essential for the proper operation of the asset sharing service. Additionally, the data linking is another important consideration, as multiple data assets having different licenses may be combined. Therefore, license compatibility needs to be checked so as to reassure that there is no violation of the data providers' terms for the data assets that will be linked together.

Regarding the secure exchange of assets, a mechanism based on blockchain technology [23] may be utilized to attach real value to proprietary assets that their providers would like to share through ICARUS. The blockchain technology might be beneficial to ICARUS since it may enhance privacy, as no third party or intermediary is involved to validate transactions between asset providers and consumers. Moreover, since information about transactions is cryptographically secured and stored across a network of computers instead of a single server, blockchain provides integrity and prevents fraud and unauthorized activities, making the transactions highly secured. Additionally, transactions can be completed faster and more efficiently, as the specific process may be automated.

For these reasons, a blockchain mechanism can be responsible for the supervision of all asset exchange activities by making decisions based on predefined ontologies and rule sets, for the generation of on-the-fly micro contracts between the collaborating stakeholders. This will ensure that there is no violation of the privacy and the asset provider's IPRs and that asset quality and delivery is guaranteed.

The Asset Sharing service is also responsible for providing the bridge for the negotiation among asset providers and consumers as well as for the "sealing" of a data sharing agreement/contract. If the interested stakeholders reach an agreement, then the blockchain mechanism is responsible to validate the transaction. On the other hand, assets that are characterized as confidential and non-shareable, will be kept only in property of their asset providers, as the specific service will guarantee that no one else has permission to access them.

2.6.2 Recommendations

Recommendation services will be provided in ICARUS, aiming at assisting the stakeholders in various ways. In particular, this service will provide recommendations a) for data assets that are based on each stakeholder's interests, as well as b) for the type of data analytics procedure to follow, based on the stakeholder's expected outcome and c) for asset licensing.

The recommendation process in ICARUS can be inspired by a typical process of a recommendation system, which consists of an iterative procedure that includes the following steps:

- **Collect information:** relevant information about the assets and the stakeholders is collected, in order to enable justified and accurate recommendations. The collection relies on different types of input, such as explicit feedback (explicit input by stakeholders), implicit feedback (indirect input through monitoring the stakeholders' behaviour) and hybrid feedback (a combination of both) [24][25].
- **Make recommendations:** using the output of the previous step, it builds a recommendation system that is able to provide recommendations/predictions to assist the stakeholders. The recommendation filtering techniques are divided in three categories according to the type of data and the computational algorithm methods: content-based filtering, collaborative filtering and hybrid filtering [26][27].

This iterative procedure aims to improve the quality and accuracy of the recommendations in each iteration. In order for this to be achieved, there are some common challenges that need to be addressed. The most common problems are the availability of descriptive data, content overspecialization, cold start (i.e. there are no information for new users or datasets), sparsity (i.e. not enough information) and scalability.

In ICARUS, an important recommendation service includes recommendation for relevant data assets. These recommendations are based on the datasets currently explored and datasets that could be linked to them, either directly or through an intermediary dataset, in order to enhance data discoverability and ensure that each stakeholder is able to reach datasets that will be more useful and suitable for its needs. This can be achieved by either content-based filtering which can relate the metadata that describe the data assets, or by collaborative filtering which can relate the stakeholders' behaviour and suggest data assets that other similar stakeholders have used or purchased.

Another type of recommendation that ICARUS aims at offering is to recommend data analytics procedure to be followed by the stakeholders. This type of recommendation includes the suggestion of machine learning algorithms, statistical methods and visualization modalities provided by ICARUS, based on the stakeholder's expected outcome and the selected data asset, in order to assist stakeholders to extract meaningful information

more quickly and efficiently, depending on their objective. This type of recommendation may be achieved by collecting information about other stakeholders who had similar goal or even used a similar dataset for their analysis. In addition, significant input to the recommendation mechanism may be provided in the form of common practices by experienced data scientists that belong to ICARUS consortium.

Furthermore, ICARUS may assist the asset providers by suggesting appropriate licenses or guiding them in order to define the appropriate license for their assets. In particular, the licenses can be divided in the following types [28][29]:

- **Attribution (BY):** Licensees must provide the author or licensor with the credits in a manner specified by them;
- **Share-alike (SA):** Licensees may distribute derivative works only under a license identical to the original (not more restrictive);
- **Non-commercial (NC):** Licensees may use the material only for non-commercial purposes;
- **No Derivatives (ND):** Licensees may not alter, transform, or build upon the specific piece of work.

The recommendations may be provided at the asset check-in steps (Step I.5 and Step VII.1 in sections 2.1.5 and 2.7.1 respectively) depending on the asset profiling that is provided by the asset provider. In addition, there can be recommendations for suitable licenses types (e.g. BY, SA, NC, ND) when sharing derivative data and reports, which are possibly the result of linked datasets with different licensing schemes. In the specific case though, the different licenses may have many restrictions, so it may not be a matter of simply suggesting suitable licenses, but mostly being forced to follow a specific license. In either case, the asset providers are able to decide which is the most suitable sharing license, based on suggestions about appropriate licenses for the specific asset.

2.6.3 Notifications

ICARUS aims at providing notification services to the aviation-related stakeholders in order to trigger events and enhance asset discoverability and proactive delivery. More precisely, this service may allow any stakeholder in the aviation data value chain to post requests for specific assets and the asset providers may be notified through the analytics of the asset utilization. On the other hand, asset consumers may receive notifications for assets that have been just uploaded and might be relevant to them, but this needs to be carefully monitored in order not to overwhelm the asset consumers with irrelevant notifications. In addition, the asset consumers that have purchased specific assets, will be notified in the future for potential updates and modifications in the terms of use (e.g. licenses) of their assets. Depending on the updating policy defined by the asset provider, the stakeholders that purchased the asset will be notified about the update and decide if they would like to keep the old version, receive the updated version or utilize both versions (depending also on the agreement with the provider). Finally, stakeholders may receive notifications regarding successful or unsuccessful execution of scheduled analytics.

2.7 Phase VII: Service Collection

The phase of Service Collection is responsible for managing the service assets in ICARUS. In this phase, stakeholders may upload their custom implemented algorithms, once they have defined in detail the terms and conditions of sharing and have provided all the metadata that relate to the service assets. Afterwards, the service assets need to be tested and assessed prior to their inclusion in ICARUS, e.g. based on some predefined Key Performance Indicators (KPIs). The service assets need to be eventually approved or rejected by a member of ICARUS administration team, that will undertake the role of moderator, in a semi-automated manner. This whole

phase is an iterative procedure that involves both the service providers and the administration team of ICARUS, as there are many clarifications and modifications that may be required in order for a service asset to be approved and available by the platform.

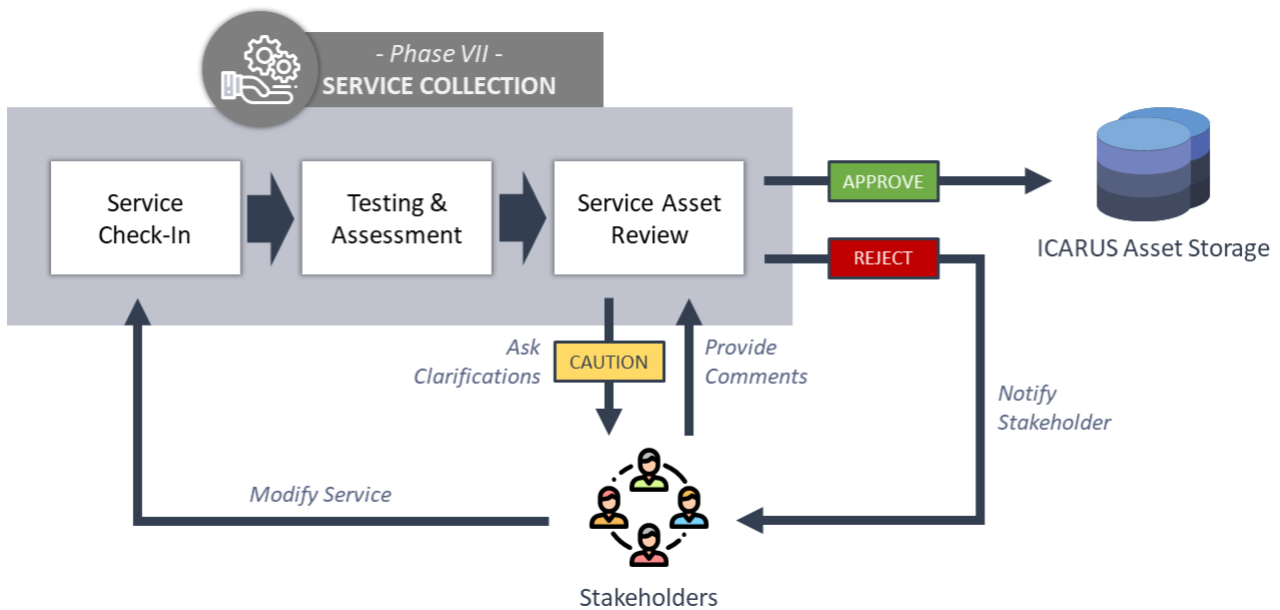


Figure 2-6: ICARUS Service Collection Phase

2.7.1 Step VII.1: Service Check-In

This step is responsible for facilitating the registration of service assets in ICARUS. As in the case of the data assets (addressed in Step I.5 of Service Check-In, Section 2.1.5), stakeholders may upload their own service (e.g. code for a new custom algorithm, end-to-end workflows containing data, algorithms and visualizations), which will be translated into service assets, either for private usage only or for sharing it as either open or proprietary.

Before the uploading of the service assets though, service providers need to provide the terms and service access policies which eventually will define and build the service license. This includes the service visibility, which may be confidential for private use only or may be shared to other stakeholders as open for everyone or as proprietary. In the latter case, the service provider needs to set a price in order to allow for potential stakeholders to use the service asset. Furthermore, the service provider must provide the metadata related to the service, including general info, training and testing datasets to evaluate the service asset performance, as well as KPIs that are filled in by the service provider to show the usefulness of the specific service asset.

2.7.2 Step VII.2: Testing and Assessment

Once the service asset is checked-in and uploaded, it should be tested and assessed in a manual or semi-automated manner by ICARUS moderator. For the evaluation, multiple tests should be defined and executed. Predefined KPIs provided by ICARUS (e.g. upper bound limitation for the ICARUS resources that service asset utilizes), as well as KPIs provided by the service provider (e.g. reduces the time needed by 10%, comparing to the algorithms provided by ICARUS in a specific task), are very important in the specific step, as they will reveal the quality of the service provided. The specific KPIs should consider the various aspects of a service, regarding the execution time (e.g. if the algorithm finishes its execution before a maximum time limit), the resources required (e.g. if the algorithm utilizes more resources than those which are available for a user), the accuracy

achieved in specific tasks to be executed in the test datasets (e.g. classification accuracy on a given test dataset must be at least 90%), as well as other important aspects such as access control policies, terms of use and so on. The tests in this step will help to conclude if the service should be allowed to be hosted in ICARUS ecosystem.

2.7.3 Step VII.3: Service Asset Review

During the specific step, ICARUS moderator needs to be involved in the process and decide if the service provided will be approved and offered by ICARUS or if it will be rejected. This decision will be based on the service's assessment and the different tests that are executed to evaluate the quality and performance of the service in the previous step (Step VII.2). In this way, the moderator will be able to decide based on specific predefined criteria that will ensure high Quality of Service as well as data integrity and data privacy as this service will make use of ICARUS resources in order to be executed. If a service asset is approved by ICARUS moderator, then it will be immediately included in ICARUS platform.

If a service asset is to be rejected, the service provider should be informed about the decision and the reasons that led to the disapproval. Thus, the service provider will be able to proceed with the appropriate modifications in order for the service to be eventually approved. Also, there is the case where ICARUS may request some clarifications from the provider prior to the decision. Therefore, this step may lead to many iterations of this phase, in order for the final decision to be made.

3 ICARUS High-level Scenarios

The specific section presents six high-level usage scenarios of ICARUS. All of the high-level usage scenarios are based on the ICARUS methodology, the input from the ICARUS pilots and the insights provided in D1.1 “Domain Landscape Review and Data Value Chain Definition”.

High-level scenarios were initially drafted as workflow diagrams (i.e. clear sequences of steps) that a stakeholder would follow in order to achieve his/her business objectives. The graphical notation of the workflow follows the color scheme of the various phases of ICARUS Methodology so as to provide a comprehensive matching modality throughout the various indicative flows. After collecting feedback on the initial drafts, the outlines of six scenarios were chosen as representative of all core differentiated ICARUS users that correspond to real life ICARUS utilization cases relevant to the stakeholders’ needs.

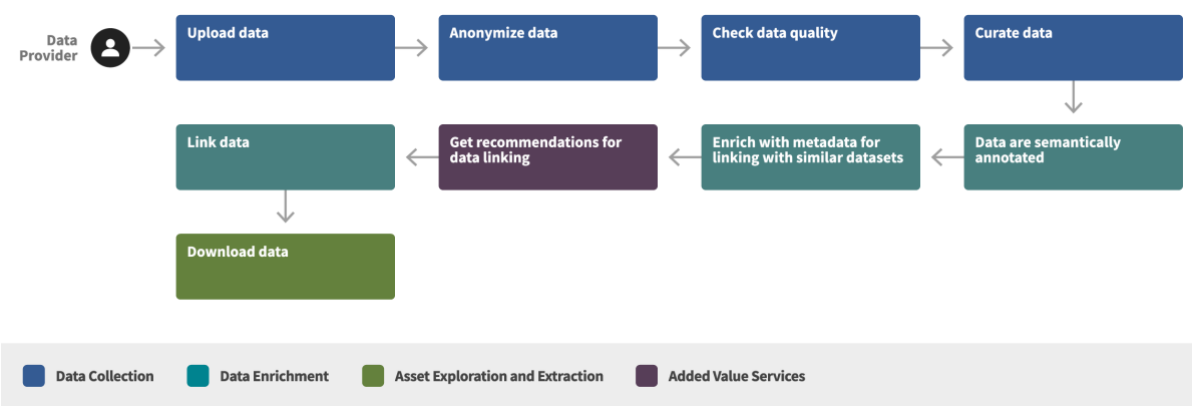
For each scenario, a general workflow diagram has been designed and the benefits and challenges of users are highlighted. Afterwards, three different examples have been instantiated for each scenario, referring to different potential stakeholders that may utilize ICARUS in different ways, depending on their needs and objectives. Each of these examples is based on the general workflow of the specific scenario, but each example modifies the general workflow slightly in order to show the flexibility of ICARUS to meet the exact stakeholders’ demands in each case.

ICARUS users have been grouped into four high-level categories that are not mutually exclusive but are utilized to better distinguish and describe the different workflows. These high-level categories are the following:

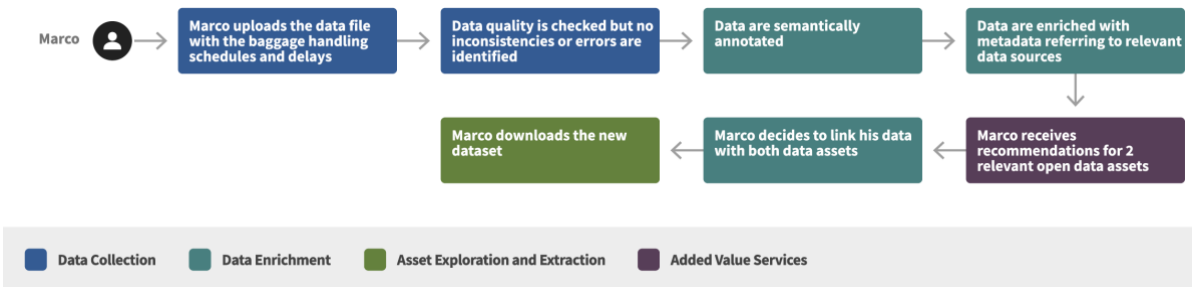
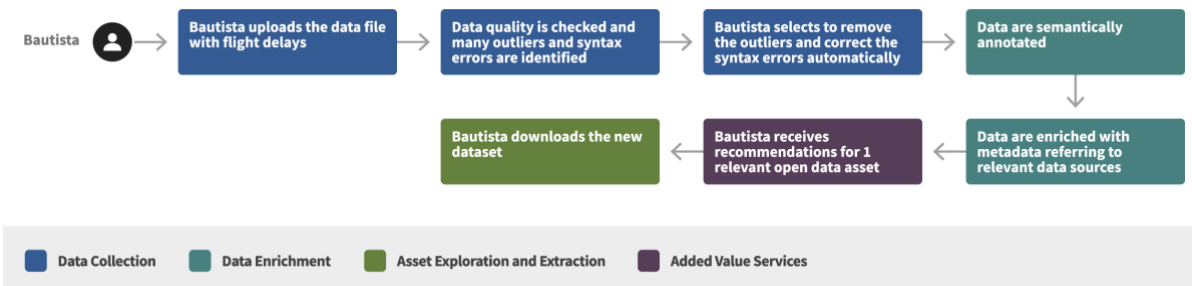
- **Data provider:** The user’s main objective is to share his/her data for processing or consuming in ICARUS;
- **Data consumer:** The user’s main objective is to consume and/or process data offered through ICARUS;
- **Service asset provider:** The user’s main objective is to create a service (e.g. custom-made algorithm) on top of ICARUS data value chain and make it available in ICARUS;
- **Service asset consumer:** The user’s main objective is to consume a service asset offered through ICARUS.

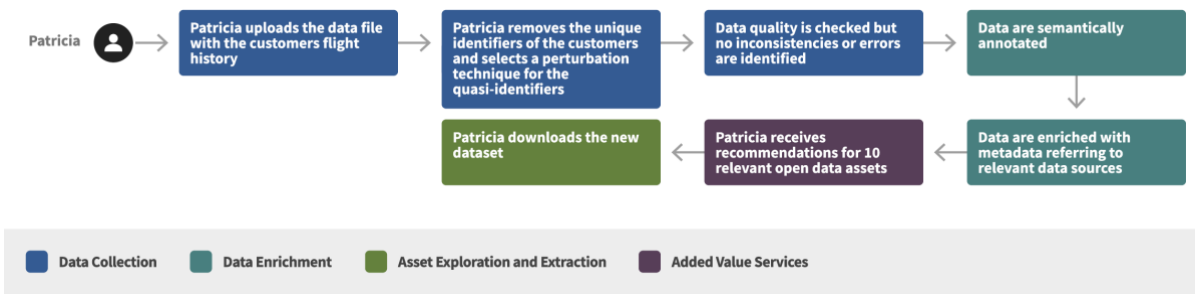
3.1 Scenario 1: Data Upload and Download

ICARUS Scenario - [SCE-1]			
Scenario ID	SCE-1	Scenario Name	Data Upload and Download
Scenario Overview	A data provider in the ICARUS data value chain requests to enhance the richness and value of his data assets without storing or sharing them in ICARUS platform. In particular, the data provider uploads his data and utilizes ICARUS mechanisms for data anonymization, data quality check, data curation, semantic annotation and data linking with relevant data sources. Finally, he downloads the new dataset without storing it to ICARUS.		
Triggers	A data provider wants to anonymize, curate and/or link his data with other relevant sources through ICARUS without storing or sharing them.		

Scenario Workflow Diagram	 <pre> graph LR DP[Data Provider] --> U[Upload data] U --> A[Anonymize data] A --> C[Check data quality] C --> Cur[Curate data] Cur --> SA[Data are semantically annotated] SA --> E[Enrich with metadata for linking with similar datasets] E --> R[Get recommendations for data linking] R --> L[Link data] L --> D[Download data] </pre> <p>Legend:</p> <ul style="list-style-type: none"> Data Collection (Blue) Data Enrichment (Teal) Asset Exploration and Extraction (Green) Added Value Services (Purple)
Scenario Sequence	<ol style="list-style-type: none"> 1. The data provider selects the option of uploading a data asset and then he selects and uploads the data asset to ICARUS. (Phase I: Data Collection, Step I.1: Data Retrieval, Section 2.1.1) 2. The sensitive information in the data asset is anonymized with the data provider's interaction. This step may be optional in case the data are already anonymized or if the data provider does not wish to share the data with others. (Phase I: Data Collection, Step I.2: Data Anonymization, Section 2.1.2) 3. The data quality is checked for inconsistencies and errors and the results are presented to the data provider. This step may be optional in case the data provider does not desire to share the data with others. (Phase I: Data Collection, Step I.3: Data Quality Check, Section 2.1.3) 4. Based on the result of the data quality check, the data are filtered and cleaned from inconsistencies and errors with the data provider interaction. This step may be optional in case the step of data quality check ensures that the data is of high quality or if the data provider does not wish to change the data. (Phase I: Data Collection, Step I.4: Data Curation, Section 2.1.4) 5. Additional information to various concepts (e.g. people, locations, organizations, etc.) is attached to the dataset in order to enrich the data assets with machine-processable information by linking background information from aviation-related vocabularies, ontologies and semantic models. (Phase II: Data Enrichment, Step II.1: Semantic Enrichment and Annotation, Section 2.2.1) 6. The data are enriched with metadata, referring to relevant data sources in ICARUS that may be linked together either directly or indirectly. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) 7. The data provider receives recommendations for relevant data sources in ICARUS that may be linked with his data. This step may be optional in case the data provider does not wish to link his data. (Phase VI: Added Value Services, Recommendations, Section 2.6.2) 8. The data are linked with other relevant data sources in ICARUS based on the data provider's decision. This step may be optional in case the data provider does not wish to link his data. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) 9. The data provider downloads the new data without storing them in ICARUS. (Phase IV: Asset Exploration and Extraction, Step IV.2: Asset Export, Section 2.4.2)

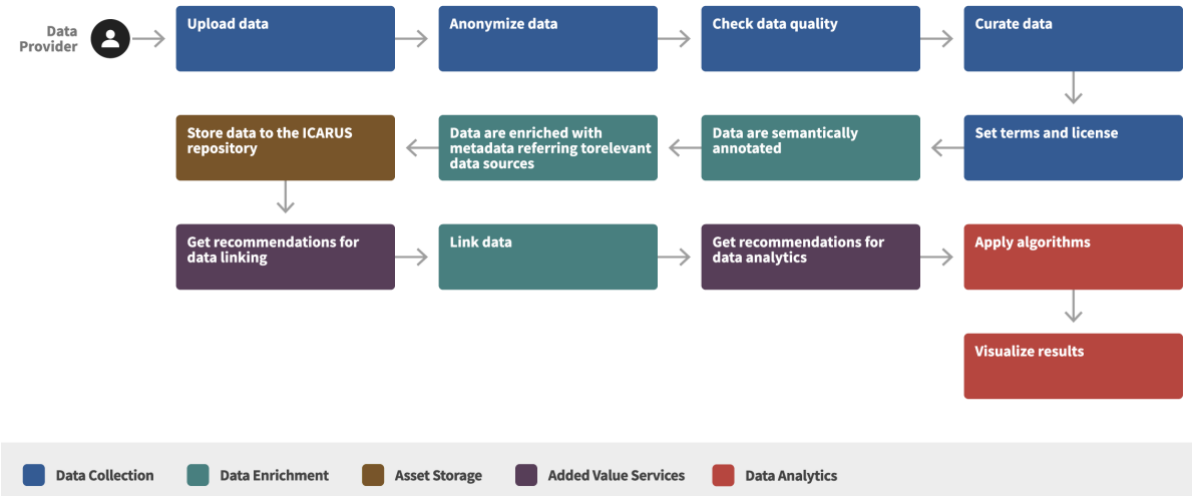
Users' Benefits	<ul style="list-style-type: none"> • Availability of a wide variety of state-of-the-art anonymization and curation techniques in order to ensure the privacy and quality of the data; • Easier discoverability and access to a variety of data sources that may be linked with the uploaded data; • Efficient mechanisms for processing data; • High level of automation in several parts of the workflow, thereby significantly reducing time required for tedious task like data anonymization, curation and linking; • Increase the value and the richness of the data through data curation and linking; • Reduced in-house development time and effort; • Private mechanisms for processing the data.
Challenges	<ul style="list-style-type: none"> • Ensure the anonymity and quality of data; • Discovery of other relevant data sources for data linking.
Exemplar Flow #1	
Example Name	An aircraft ground handling company desires to use the data linking mechanism of ICARUS to increase the value and richness of its data.
Example Description	<p>Marco is a head data scientist in "AERO Ground Handling", an aircraft ground handling company based in Germany with more than 280 employees. Currently, the company wants to increase the value and richness of its data in order to become more competitive. Thus, Marco suggested to enrich the data using other relevant information. However, although all team members found the idea interesting, nobody was aware of an easy way to find and acquire relevant data sources that could be linked with the company's data. Therefore, Marco proposed that all team members search for existing solutions that could be used. Through the specific process, one of his colleagues came across ICARUS. So, Marco decided to upload the data and experiment with the data linking mechanism of ICARUS.</p> <p>Marco accesses ICARUS and selects to upload a dataset. The dataset that Marco wishes to upload is a batch file which contains structured information about baggage handling schedules and delays. Furthermore, the dataset contains 150.000 records and was collected over the period 2016-2017. After uploading the dataset, Marco is asked to anonymize the data but ignores this step as the data are already anonymized. Afterwards, he selects to check the quality of the data using the data quality mechanism of ICARUS, but no inconsistencies or errors are identified. Hence, he ignores the step of data curation and proceeds with the semantic enrichment and metadata enrichment of relevant data sources that are executed by ICARUS. Then, based on Marco's dataset, ICARUS recommendation mechanism suggests two relevant open data assets that could be linked with his dataset. The first data asset contains weather data with linking compatibility 92%, while the second data asset contains information about the expected passengers per flight with linking compatibility 86%. The resulted linking compatibility indicates that the probability of the data assets to refer in the same object is very high. Thus, Marco decides to link both of these data assets with his data and accepts the terms of use of the public data assets. Finally, he downloads the new dataset without storing it inside ICARUS repository.</p>

Example Workflow Diagram	 <pre> graph LR Marco((Marco)) --> A[Marco uploads the data file with the baggage handling schedules and delays] A --> B[Data quality is checked but no inconsistencies or errors are identified] B --> C[Data are semantically annotated] C --> D[Data are enriched with metadata referring to relevant data sources] D --> E[Marco receives recommendations for 2 relevant open data assets] E --> F[Marco decides to link his data with both data assets] F --> G[Marco downloads the new dataset] </pre> <p> ■ Data Collection ■ Data Enrichment ■ Asset Exploration and Extraction ■ Added Value Services </p>
Exemplar Flow #2	
Example Name	An airport authority wishes to experiment with the data curation mechanism of ICARUS.
Example Description	<p>Bautista is a senior data analyst and he is working in the marketing department at an airport. The airport is located in Spain and employs more than 700 employees. Currently, Bautista is working on a project in order to analyze the flights' delays of the airport. However, he detected a lot of errors and inconsistencies in the dataset that downgrade his analysis. Furthermore, because the dataset is huge, he will need a lot of time and effort to find all the errors. Thus, he decided to search for existing solutions that could be utilized. Through this process, he came across ICARUS and decided to upload the dataset so as to experiment with the data curation mechanism of ICARUS.</p> <p>Bautista accesses ICARUS and selects to upload a dataset. The dataset that Bautista wishes to upload is a batch file which contains structured information about flights delays. Furthermore, the dataset contains 50.000.000 records and was collected over the period 2013-2018. After uploading the dataset, Bautista is asked to anonymize the data but bypasses the specific step as the data are already anonymized. Afterwards, he chooses to check the quality of the data using the data quality mechanism of ICARUS. The mechanism pinpointed many inconsistencies and errors like outliers and syntax errors and thus it provides him with potential suitable cleaning techniques and guidelines. Based on the identified errors and inconsistencies, Bautista selects to remove the outliers and correct the syntax errors. Once finished, the semantic enrichment and metadata enrichment of relevant data sources are automatically performed by ICARUS. Subsequently, based on Bautista's dataset, the ICARUS recommendation mechanism suggests one relevant open data asset that could be linked with his dataset. The recommended data asset contains public transportation schedules in Spain with linking compatibility 60%. Since the data linking compatibility is not high, Bautista decides to bypass the step of data linking and he downloads the curated dataset without storing it in ICARUS repository.</p>
Example Workflow Diagram	 <pre> graph LR Bautista((Bautista)) --> A[Bautista uploads the data file with flight delays] A --> B[Data quality is checked and many outliers and syntax errors are identified] B --> C[Bautista selects to remove the outliers and correct the syntax errors automatically] C --> D[Data are semantically annotated] D --> E[Data are enriched with metadata referring to relevant data sources] E --> F[Bautista receives recommendations for 1 relevant open data asset] F --> G[Bautista downloads the new dataset] </pre> <p> ■ Data Collection ■ Data Enrichment ■ Asset Exploration and Extraction ■ Added Value Services </p>
Exemplar Flow #3	
Example Name	A travel agency wishes to use the data anonymization mechanism of ICARUS in order to comply with the European data protection regulations.

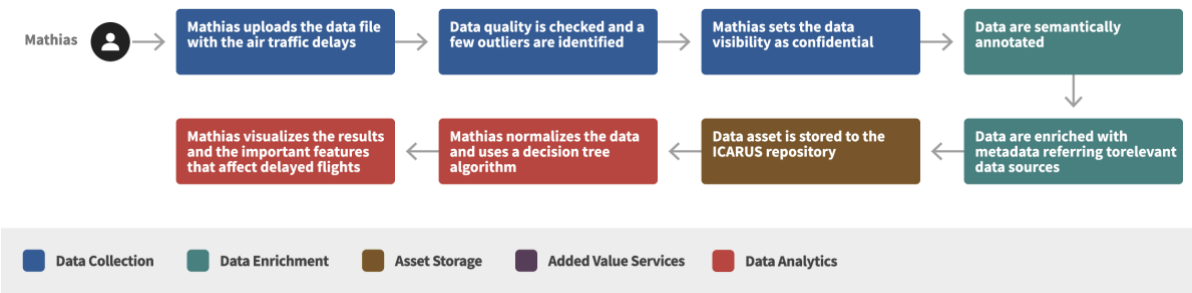
Example Description	<p>Patricia is a junior software developer in "Oasis Vacations", a travel agency based in United Kingdom with more than 150 employees. Currently, the company wants to comply with the European data protection regulations. Therefore, Patricia was assigned with the task to anonymize the data of the travel agency. However, Patricia finds it quite challenging to understand and apply all the new regulations. Thus, she decided to search for existing solutions that could be utilized. Through this process, she came across ICARUS and decided to upload the data in order to use the mechanism for data anonymization of ICARUS.</p> <p>Patricia accesses ICARUS and selects to upload a dataset. The dataset that Patricia wishes to upload is a batch file which contains semi-structured information about customers flight history. Furthermore, the dataset contains 80.000 records and was collected over the period 2010-2017. After uploading the dataset, the ICARUS mechanism of data anonymization identifies many unique identifiers and quasi-identifiers of the customers and thus it provides the user with potential suitable anonymization techniques. Therefore, Patricia selects to remove the unique attributes and use a perturbation technique for the quasi-identifiers in order to remove and hide sensitive information about the customers. Afterwards, she selects to check the quality of the data using the data quality mechanism of ICARUS, but no inconsistencies or errors are identified. Afterwards, the semantic enrichment and metadata enrichment of relevant data sources are performed by ICARUS. Subsequently, based on Patricia's dataset, the ICARUS recommendation mechanism suggests ten relevant open data assets that could be linked with her dataset, but she decides to bypass the step of data linking. Finally, she downloads the anonymized dataset without storing it in ICARUS repository.</p>
Example Workflow Diagram	 <pre> graph LR Patricia((Patricia)) --> A[Patricia uploads the data file with the customers flight history] A --> B[Patricia removes the unique identifiers of the customers and selects a perturbation technique for the quasi-identifiers] B --> C[Data quality is checked but no inconsistencies or errors are identified] C --> D[Data are semantically annotated] D --> E[Data are enriched with metadata referring to relevant data sources] E --> F[Patricia receives recommendations for 10 relevant open data assets] F --> G[Patricia downloads the new dataset] G --> Patricia </pre> <p>Legend:</p> <ul style="list-style-type: none"> Data Collection (Blue) Data Enrichment (Teal) Asset Exploration and Extraction (Green) Added Value Services (Purple)

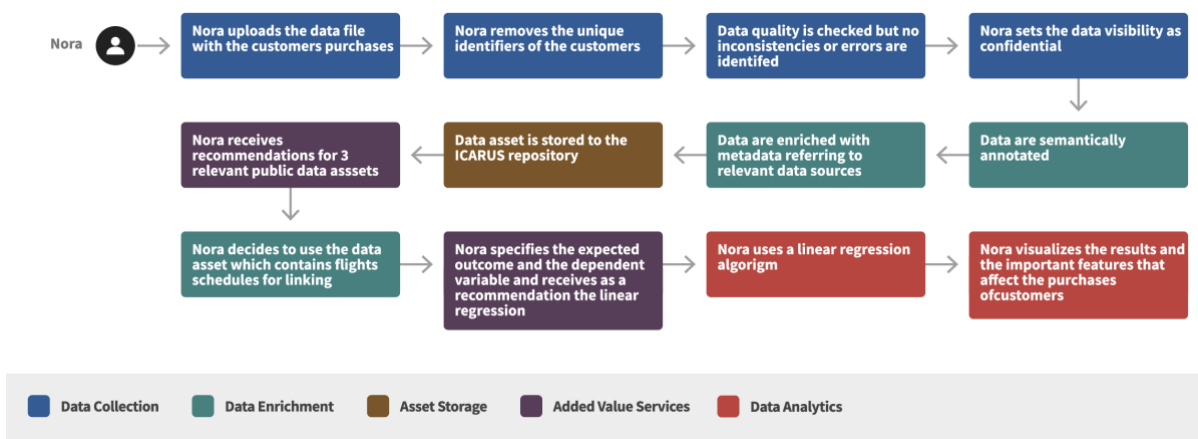
3.2 Scenario 2: Data Upload and Analysis

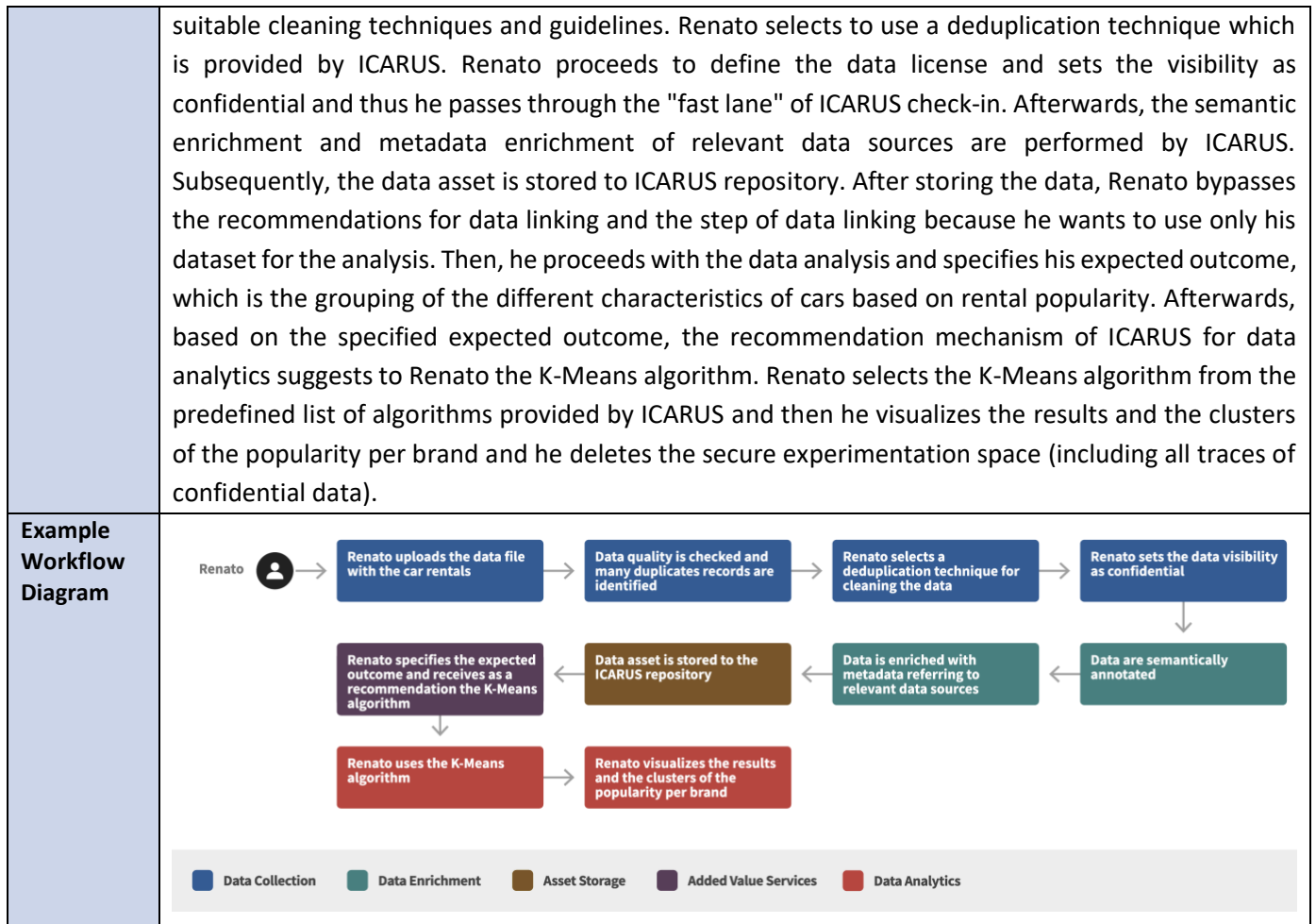
ICARUS Scenario - [SCE-2]			
Scenario ID	SCE-2	Scenario Name	Data Upload and Analysis
Scenario Overview	<p>A data provider in the ICARUS data value chain wants to privately analyze its data and extract insights about them. More specifically, the data provider uploads a dataset and makes all the required steps in order to store it inside ICARUS repository. These steps are the anonymization of the data to ensure privacy, quality check and curation of the data to remove inconsistencies and errors as well as data enrichment with semantic annotations and metadata referring to relevant sources that can be linked together. Afterwards, the data provider sets the data license and specifies that the data are not for sharing (restricted visibility). After the data is stored in the ICARUS repository, the data provider may take advantage of the data linking recommendations and the data linking mechanisms of ICARUS.</p>		

	Finally, the data provider can take advantage of the data analysis recommendations mechanism of ICARUS before proceeding to the analysis of data.
Triggers	A data provider wants to analyze his data through ICARUS without sharing them.
Scenario Workflow Diagram	 <pre> graph LR DP[Data Provider] --> U[Upload data] U --> A[Anonymize data] A --> C[Check data quality] C --> Cur[Curate data] Cur --> S[Set terms and license] S --> E[Data are semantically annotated] E --> M[Data are enriched with metadata referring to relevant data sources] M --> S2[Store data to the ICARUS repository] S2 --> R[Get recommendations for data linking] R --> L[Link data] L --> G[Get recommendations for data analytics] G --> A2[Apply algorithms] A2 --> V[Visualize results] </pre> <p>Legend: Data Collection (Blue), Data Enrichment (Green), Asset Storage (Brown), Added Value Services (Purple), Data Analytics (Red)</p>
Scenario Sequence	<ol style="list-style-type: none"> 1. The data provider selects the option of uploading a data asset to ICARUS either as a batch file or through an API. (Phase I: Data Collection, Step I.1: Data Retrieval, Section 2.1.1) 2. The sensitive information in the data asset is anonymized with the data provider interaction. This step may be bypassed if the data are already anonymized or if the data provider does not wish to share the data with others. (Phase I: Data Collection, Step I.2: Data Anonymization, Section 2.1.2) 3. The data quality is checked for inconsistencies and errors and the results are presented to the data provider. This step may be bypassed if the data provider does not desire to share the data with others. (Phase I: Data Collection, Step I.3: Data Quality Check, Section 2.1.3) 4. Based on the result of the data quality check, the data are filtered and cleaned from inconsistencies and errors with the data provider interaction. This step may be optional if the step of data quality check ensures that the data is of high quality or if the data provider does not wish to change the dataset. (Phase I: Data Collection, Step I.4: Data Curation, Section 2.1.4) 5. The data provider defines the data asset's metadata regarding the data profiling (e.g. type, format, language etc.) and the data asset rights (e.g. data visibility, license, pricing, updating policy, generate public data sample etc.). In this scenario, the data provider sets the data visibility to "confidential" (not to be shared). (Phase I: Data Collection, Step I.5: Data Check-in, Section 2.1.5) 6. Additional information to various concepts (e.g. people, locations, organizations, etc.) is attached to the data in order to enrich the data assets with machine-processable information by linking background information from aviation-related vocabularies, ontologies and semantic models. (Phase II: Data Enrichment, Step II.1: Semantic Enrichment and Annotation, Section 2.2.1) 7. The data are enriched with metadata, referring to relevant data sources in ICARUS that may be linked together either directly or indirectly. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) 8. The data asset is stored to the storage of ICARUS. (Phase III: Asset Storage, Section 2.3)

	<ol style="list-style-type: none"> 9. The data provider receives recommendations for relevant data sources in ICARUS that may be linked with its data. This step may be optional if the data provider does not wish to link his data. (Phase VI: Added Value Services, Recommendations, Section 2.6.2) 10. The data are linked with other relevant data sources in ICARUS based on the data provider's decision. This step may be optional if the data provider does not wish to link his data with other datasets. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) 11. The data provider receives recommendations about the methodology to follow for the data analysis (data pre-processing, data analytics algorithms, visualizations) based on the specified expected outcome. This step may be optional if the data provider does not wish to specify the expected outcome. (Phase VI: Added Value Services, Recommendations, Section 2.6.2) 12. The data provider applies different algorithms and statistical methods to the data. (Phase V: Data Analytics, Step V.1: Data Analysis, Section 2.5.1) 13. The data provider visualizes the results of the analysis. (Phase V: Data Analytics, Step V.2: Data Visualization, Section 2.5.2)
Users' Benefits	<ul style="list-style-type: none"> • Availability of a wide variety of state-of-the-art machine learning algorithms; • Availability of a wide variety of visualizations; • Easy experimentation and comparison of results of analysis; • Efficient mechanisms for processing and analyzing data; • Recommendations for algorithms and visualizations to utilize, that make it even easier for novice users to analyze their data; • Private mechanisms for processing the data; • High performance storage; • Easier discoverability and access to a variety of data sources that can be linked with the uploaded data; • High level of automation in several parts of the workflow, thereby significantly reducing time required for tedious task like anonymization, curation and linking; • Availability of a wide variety of state-of-the-art anonymization and curation techniques in order to ensure the privacy and quality of the data assets; • Increase the value and the richness of the data through data curation and linking; and • Reduced in-house development time and effort and less investments in local IT infrastructure and experts.
Challenges	<ul style="list-style-type: none"> • Ensure the anonymity and quality of data; • Discovery of other relevant data sources for data linking; • Use of the appropriate analytical methodology (data preparation, data analysis algorithm and visualizations) in order to achieve the expected outcome; • Real-time requirements for data analysis; • Configurable visualizations and visual analytics features for data
Exemplar Flow #1	
Example Name	An air navigation service provider wishes to analyze its data (air traffic delays) privately and extract knowledge, using the data analytics tools of ICARUS.

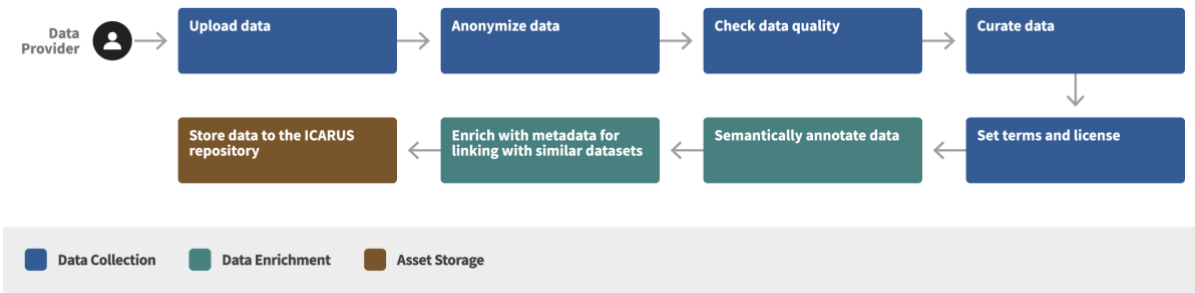
Example Description	<p>Mathias is a junior data scientist and he is working in an air navigation service provider. The air navigation service provider is located in Austria and employs more than 260 employees. Mathias was assigned to analyze the data and find patterns for the air traffic delays. However, he finds it quite challenging to implement an algorithm for data analytics. Thus, Mathias searched for existing solutions that could be utilized. Through this process, he came across ICARUS. So, he decided to upload and analyze the data using ICARUS platform.</p> <p>Mathias accesses ICARUS and selects to upload a dataset. The dataset that Mathias wishes to upload is a batch file which contains structured information about air traffic delays. Furthermore, the dataset contains 550.000 records and was collected over the period 2017-2018. After uploading the dataset, Mathias is asked to anonymize the data, but bypasses this step as the data are already anonymized. Afterwards, he selects to check the quality of the data using the data quality mechanism of ICARUS. The mechanism identified a few outliers; however, Mathias decides not to clean his data. Then, he proceeds to define the data license and sets the visibility as confidential and thus, he passes through the "fast lane" of ICARUS check-in. Subsequently, the semantic enrichment and metadata enrichment of relevant data sources are performed by ICARUS. Afterwards, the data asset is stored to ICARUS repository. After storing the data, Mathias bypasses the recommendations for data linking and the step of data linking because he wants to use only his dataset for the analysis. Then, he proceeds with the data analysis and selects to normalize the data. Afterwards, based on a predefined list of algorithms provided by ICARUS, he selects to use a decision tree algorithm in order to identify the important features that affect delayed flights. Finally, Mathias visualizes the results and the important features that affect the delayed flights.</p>
Example Workflow Diagram	 <pre> graph LR Mathias((Mathias)) --> A[Mathias uploads the data file with the air traffic delays] A --> B[Data quality is checked and a few outliers are identified] B --> C[Mathias sets the data visibility as confidential] C --> D[Data are semantically annotated] D --> E[Data are enriched with metadata referring to relevant data sources] E --> F[Data asset is stored to the ICARUS repository] F --> G[Mathias normalizes the data and uses a decision tree algorithm] G --> H[Mathias visualizes the results and the important features that affect delayed flights] </pre> <p>Legend:</p> <ul style="list-style-type: none"> Data Collection (Blue) Data Enrichment (Green) Asset Storage (Brown) Added Value Services (Purple) Data Analytics (Red)
Exemplar Flow #2	
Example Name	A duty-free store wishes to anonymize, link and analyze its data (customers purchases) privately and extract knowledge, using the data analytics tools of ICARUS.
Example Description	<p>Nora is a data analyst in "Emperor Power", a duty-free store based in Denmark with 15 employees. Currently, the manager of the store wants to anonymize and analyze the store's data and extract various customer trends. Therefore, Nora was assigned to analyze the data of the store. However, as the deadline was approaching, she decided to search for existing solutions that could help her analyze the data faster. Then, she came across ICARUS and decided to upload the data in order to use the mechanisms for data anonymization and analytics of ICARUS.</p> <p>Nora accesses ICARUS and selects to upload a dataset. The dataset that Nora wishes to upload is a batch file which contains semi-structured information about customers purchases. Furthermore, the dataset contains 50.000 records and was collected over the period 2015-2017. After uploading the dataset, the ICARUS mechanism of data anonymization identifies the unique identifiers of the</p>

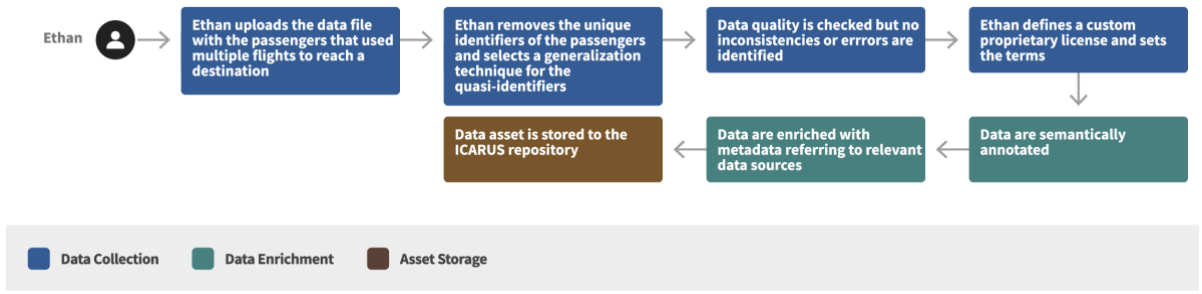
	<p>customers and thus it provides the user with potential suitable anonymization techniques. Therefore, Nora selects to remove the column with the sensitive information. Afterwards, she selects to check the quality of the data using the data quality mechanism of ICARUS, but no inconsistencies or errors are identified. Thus, Nora proceeds to define the data license and sets the visibility as confidential and thus she passes through the "fast lane" of ICARUS check-in. Afterwards, the semantic enrichment and metadata enrichment of relevant data sources are performed by ICARUS. Subsequently, the data asset is stored to ICARUS repository. Afterwards, based on Nora's dataset, ICARUS recommendation mechanism suggests three relevant open data assets that could be linked with her dataset. Based on the recommended data assets, Nora decides to use the data asset which contains information about flight schedules and accepts the terms of use of the public data asset. Then, she proceeds with the data analysis and specifies her expected outcome, which is to identify important features that may affect the purchases of the customers. Furthermore, she also specifies the amount spent by each customer as the dependent variable. Afterwards, based on the specified expected outcome and dependent variable, the recommendation mechanism of ICARUS for data analytics suggests to Nora the linear regression. Thus, Nora selects the linear regression from the predefined list of algorithms provided by ICARUS in order to identify the important features that affect the purchases of customers. Finally, Nora visualizes the results and the important features of the analysis.</p>
Example Workflow Diagram	 <pre> graph LR Nora((Nora)) --> A[Nora uploads the data file with the customers purchases] A --> B[Nora removes the unique identifiers of the customers] B --> C[Data quality is checked but no inconsistencies or errors are identified] C --> D[Nora sets the data visibility as confidential] D --> E[Data are semantically annotated] E --> F[Data are enriched with metadata referring to relevant data sources] F --> G[Data asset is stored to the ICARUS repository] G --> H[Nora receives recommendations for 3 relevant public data assets] H --> I[Nora decides to use the data asset which contains flights schedules for linking] I --> J[Nora specifies the expected outcome and the dependent variable and receives as a recommendation the linear regression] J --> K[Nora uses a linear regression algorithm] K --> L[Nora visualizes the results and the important features that affect the purchases of customers] </pre> <p>Legend: Data Collection (Blue), Data Enrichment (Green), Asset Storage (Brown), Added Value Services (Purple), Data Analytics (Red)</p>
Exemplar Flow #3	
Example Name	A car rental company desires to curate, link and analyze its data (car rentals) privately and extract knowledge, using the data analytics tools of ICARUS.
Example Description	<p>Renato is a software developer in "Auto Car Rental", a car rental company based in Italy with 20 employees. Currently, the manager of the company wants to analyze the customers' data in order to provide better offers. Thus, Renato was assigned to analyze the data of the store. However, Renato has no knowledge of data analytics. So, he decided to search for existing solutions that could help him analyze the data. After some searching, he came across ICARUS, a platform which ensures the privacy and security of the data by a blockchain-based asset sharing service. Thus, he decided to upload the data in order to use the mechanism for data analytics of ICARUS.</p> <p>Renato accesses ICARUS and selects to upload a dataset. The dataset that Renato wishes to upload will be uploaded through an API that responds to calls for specific geographic areas and time periods with semi-structured information about car rentals. After uploading the data for all geographic areas for 2014-2017, Renato selects to check the quality of the data using the data quality mechanism of ICARUS. The mechanism identified many duplicated records and thus it provides him with potential</p>

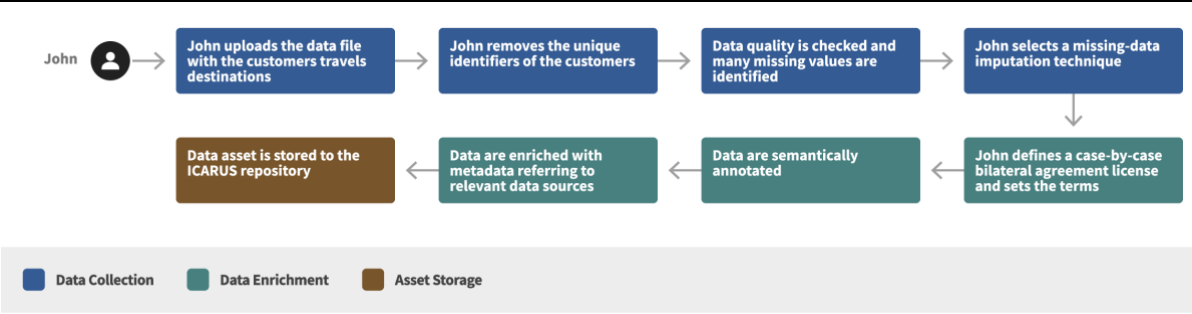


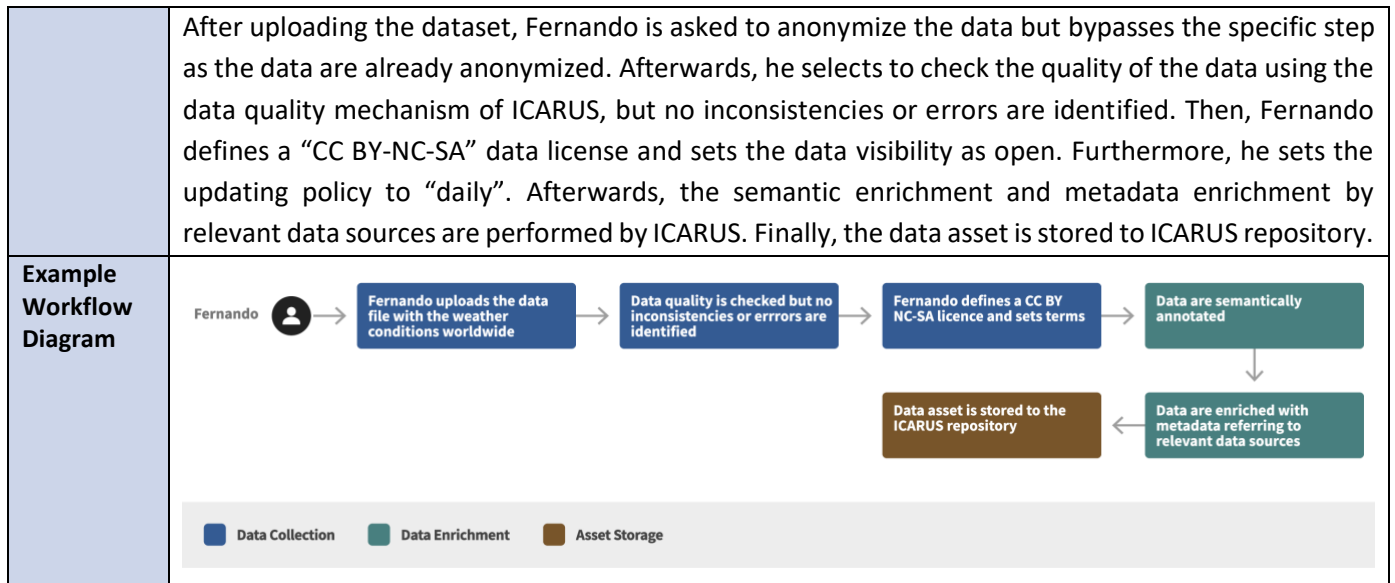
3.3 Scenario 3: Data Upload and Sharing

ICARUS Scenario - [SCE-3]			
Scenario ID	SCE-3	Scenario Name	Data Upload and Sharing
Scenario Overview	A data provider in the ICARUS data value chain wants to share his data as open to the public or to reach other stakeholders that are interested to purchase its data. In this process, the anonymity and quality of data need to be guaranteed, as well as the enrichment of the data with aviation-related information and metadata referring to relevant sources that can be linked together. The data provider's terms and data license are defined, before proceeding to the process of storing the data to ICARUS repository.		
Triggers	A data provider wants to share his data through ICARUS.		

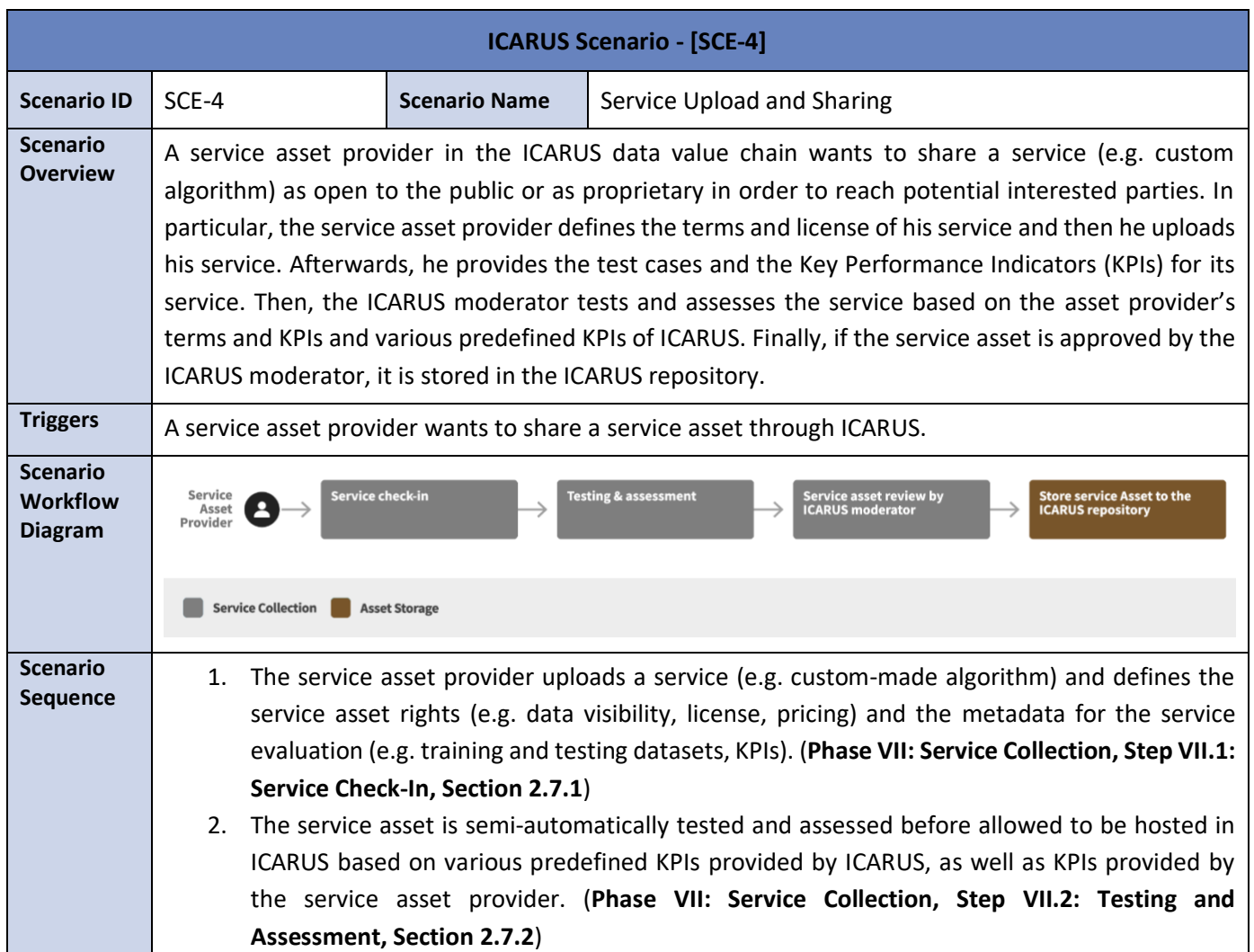
Scenario Workflow Diagram	 <pre> graph LR DP[Data Provider] --> U[Upload data] U --> A[Anonymize data] A --> C[Check data quality] C --> Cur[Curate data] Cur --> S[Set terms and license] S --> Ann[Semantically annotate data] Ann --> E[Enrich with metadata for linking with similar datasets] E --> S2[Store data to the ICARUS repository] </pre> <p>Legend: ■ Data Collection ■ Data Enrichment ■ Asset Storage</p>
Scenario Sequence	<ol style="list-style-type: none"> 1. The data provider selects the option of uploading a data asset and then he selects and uploads the data asset to ICARUS. (Phase I: Data Collection, Step I.1: Data Retrieval, Section 2.1.1) 2. The sensitive information in the data asset is anonymized with the data provider interaction. This step may be bypassed in case the data are already anonymized or if the data provider does not wish to share the data with others. (Phase I: Data Collection, Step I.2: Data Anonymization, Section 2.1.2) 3. The data quality is checked for inconsistencies and errors and the results are presented to the data provider. This step may be bypassed in case the data provider does not desire to share the data with other. (Phase I: Data Collection, Step I.3: Data Quality Check, Section 2.1.3) 4. Based on the result of the data quality check, the data are filtered and cleaned from inconsistencies and errors with the data provider interaction. This step may be bypassed in case the step of data quality check ensures that the data is of high quality or if the data provider does not wish to modify the data. (Phase I: Data Collection, Step I.4: Data Curation, Section 2.1.4) 5. The data provider defines the data asset's metadata regarding the data profiling (e.g. type, format, language etc.) and the data asset rights (e.g. data visibility, license, pricing, updating policy, generate public data sample etc.). In this scenario, the data provider selects to generate a data sample and sets the data visibility to "open" (public) or "proprietary" (can be shared with the appropriate licensing and/or after purchasing). (Phase I: Data Collection, Step I.5: Data Check-in, Section 2.1.5) 6. Additional information to various concepts (e.g. people, locations, organizations, etc.) are attached to the data in order to enrich the data assets with information by linking background information from aviation-related vocabularies, ontologies and semantic models. (Phase II: Data Enrichment, Step II.1: Semantic Enrichment and Annotation, Section 2.2.1) 7. The data are enriched with metadata, referring to relevant data sources in ICARUS that can be linked together either directly or indirectly. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) 8. The data asset is stored to the storage of ICARUS. (Phase III: Asset Storage, Section 2.3)
Users' Benefits	<ul style="list-style-type: none"> • Easier and safer sharing of open or proprietary data that reach a wider audience either for monetisation or for contributing to the aviation research society; • High performance storage; • High level of automation in several parts of the workflow, thereby significantly reducing time required for tedious task like anonymization and curation;

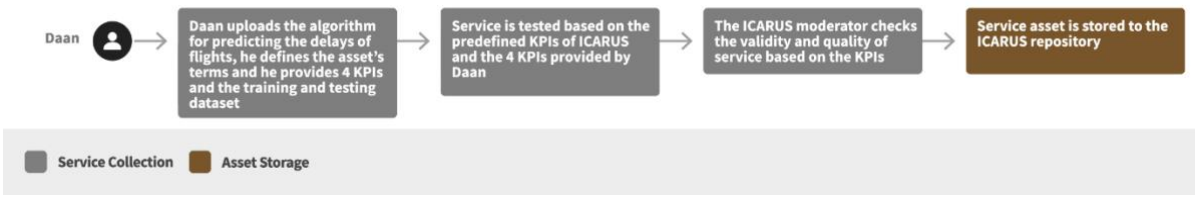
	<ul style="list-style-type: none"> • Increase the value and validity of the data through data curation; • Availability of a wide variety of state-of-the-art anonymization and curation techniques in order to ensure the privacy and quality of the data; • Efficient mechanisms for processing data.
Challenges	<ul style="list-style-type: none"> • Ensure the anonymity and quality of the data; • Define the right terms of usage and licensing schemes for data in order to reach a wider audience.
Exemplar Flow #1	
Example Name	An airline company wishes to share its non-critical data (as proprietary) to experiment with the ICARUS functionalities.
Example Description	<p>Ethan is a marketing manager in "Gamma Airlines", an airline company based in France with more than 1200 employees. Currently, the company wants to sell various of its data in order to increase the profit of the company. Hence, Ethan was assigned to find potential interested parties and negotiate with them. Thus, he decided to search for existing solutions that could help him negotiate and commercialize the data of the company. Then, he came across ICARUS and decided to upload and share the non-critical data of the company in order to experiment with the ICARUS functionalities.</p> <p>Ethan accesses ICARUS and selects to upload a dataset. Ethan wishes to upload a batch file which contains structured information about passengers that used multiple flights to reach a destination. Furthermore, the dataset contains 1.500.000 records and was collected over the period 2013-2018. After uploading the dataset, ICARUS mechanism of data anonymization identifies many unique identifiers and quasi-identifiers of the passengers and thus it provides the user with potential suitable anonymization techniques and guidelines. Thus, Ethan selects to remove the unique attributes and use a generalization technique for the quasi-identifiers in order to remove and hide sensitive information about the passengers. Afterwards, he chooses to check the quality of the data using the data quality mechanism of ICARUS, but no inconsistencies or errors are identified. Ethan proceeds to define his custom proprietary data license and sets the data visibility as proprietary. In particular, he sets the pricing to 200€ and selects to create a sample of the data with 2000 records. Furthermore, he sets the updating policy to "monthly" as the data will be updated per month. Afterwards, the semantic enrichment and metadata enrichment of relevant data sources are performed by ICARUS. Finally, the data asset is stored to ICARUS repository.</p>
Example Workflow Diagram	 <pre> graph LR Ethan((Ethan)) --> A[Ethan uploads the data file with the passengers that used multiple flights to reach a destination] A --> B[Ethan removes the unique identifiers of the passengers and selects a generalization technique for the quasi-identifiers] B --> C[Data quality is checked but no inconsistencies or errors are identified] C --> D[Ethan defines a custom proprietary license and sets the terms] D --> E[Data are semantically annotated] E --> F[Data are enriched with metadata referring to relevant data sources] F --> G[Data asset is stored to the ICARUS repository] G --> B </pre> <p>Legend: ■ Data Collection ■ Data Enrichment ■ Asset Storage</p>
Exemplar Flow #2	



Example Name	A travel agency desires to share its non-critical data (as proprietary) through ICARUS to reach potential interested parties.
Example Description	<p>John is a marketing assistant in "Dreamy Travelers", a travel agency based in Greece with 60 employees. Currently, the company wants to sell its data and reach more potential buyers. Thus, John was assigned to find potential interested parties and sell the data. Then, he decided to search for existing solutions that could help him commercialize the data in a systematic fashion. Then, he came across ICARUS and decided to upload and share the non-critical data through ICARUS to reach potential interested parties.</p> <p>John accesses ICARUS and selects to upload a dataset. The dataset that John wishes to upload is a batch file which contains structured information about customers' travels destinations. Furthermore, the dataset contains 30.000 records and was collected over the period 2015-2018. After uploading the dataset, ICARUS mechanism of data anonymization identifies the unique identifiers of the customers and thus it provides the user with potential suitable anonymization techniques. Therefore, John selects to remove the sensitive information about the customers. Afterwards, he selects to check the quality of the data using the data quality mechanism of ICARUS. The mechanism identified many missing values and provides the user with potential suitable cleaning techniques and guidelines. Based on the identified errors and inconsistencies, John selects to use a missing-data imputation technique. Then, John defines a "Case-by-Case Bilateral Agreement" data license and sets the data visibility as proprietary. In particular, he sets the pricing to "negotiable" and selects not to create a sample of the data. Furthermore, he defines that the data asset will have no updates. Afterwards, the semantic enrichment and metadata enrichment based on relevant data sources are performed by ICARUS. Finally, the data asset is stored to ICARUS repository.</p>
Example Workflow Diagram	 <pre> graph LR John((John)) --> A[John uploads the data file with the customers travels destinations] A --> B[John removes the unique identifiers of the customers] B --> C[Data quality is checked and many missing values are identified] C --> D[John selects a missing-data imputation technique] D --> E[John defines a case-by-case bilateral agreement license and sets the terms] E --> F[Data are semantically annotated] F --> G[Data are enriched with metadata referring to relevant data sources] G --> H[Data asset is stored to the ICARUS repository] </pre> <p>Legend: ■ Data Collection, ■ Data Enrichment, ■ Asset Storage</p>
Exemplar Flow #3	
Example Name	A weather forecasting organization wishes to share its non-critical data (as open) through ICARUS to contribute to the aviation research society.
Example Description	<p>Fernando is a research assistant in "Meteo-Analytics", a weather forecasting organization based in Portugal with 80 employees. Currently, Fernando was assigned to share the organization's non-critical data as open in order to increase the reputation of the organization by contributing to the aviation research society. Then, he decided to search for existing solutions that could help him share the data. He came across ICARUS and decided to upload and share the non-critical dataset through ICARUS.</p> <p>Fernando accesses ICARUS and selects to upload a dataset. The dataset that Fernando wishes to upload is a batch file which contains structured information about weather conditions worldwide. Furthermore, the dataset contains 400.000 records and was collected over the period 2016-2018.</p>



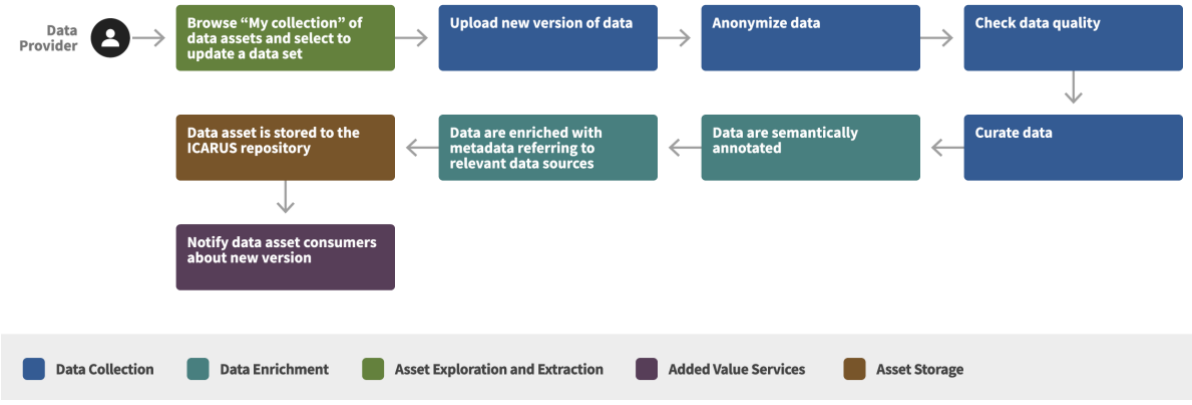
3.4 Scenario 4: Service Upload and Sharing



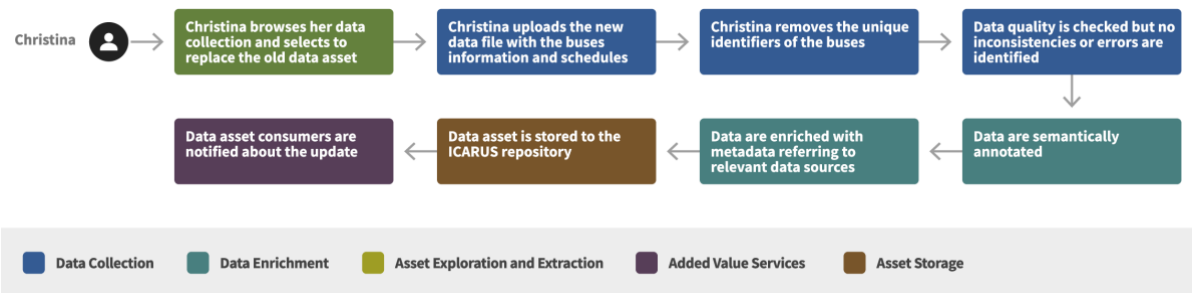
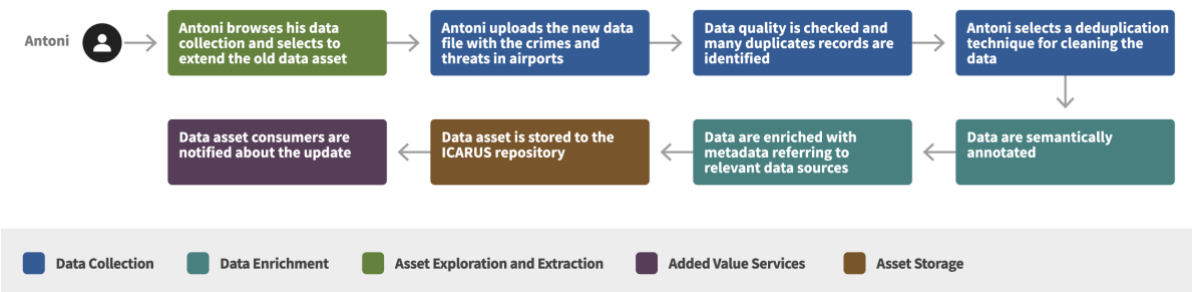
	<p>3. Based on the service's assessment and testing, the ICARUS moderator decides if the service asset will be approved and offered by ICARUS or if it will be rejected. If the service asset is rejected, then the service provider is informed about the decision and the reasons that led to the disapproval. (Phase VII: Service Collection, Step VII.3: Service Asset Review, Section 2.7.3)</p> <p>4. The service asset is stored to the storage of ICARUS. (Phase III: Asset Storage, Section 2.3)</p>
Users' Benefits	<ul style="list-style-type: none"> Easier sharing of innovative services that reach a wider audience for monetisation; Highly secure storage for assets; Usage of the custom-made services inside ICARUS.
Challenges	<ul style="list-style-type: none"> Test cases and KPIs that correspond to real-life scenarios; The service asset needs to be tested and assessed before it is approved.
Exemplar Flow #1	
Example Name	An airport authority wishes to share a custom-made service (as proprietary) through ICARUS to reach potential interested parties.
Example Description	<p>Daan is a senior data scientist and he is working in an airport. The airport is located in Netherlands and employs more than 1600 employees. Currently, Daan has created an efficient and accurate algorithm for predicting the delays of flights. Thus, he was assigned by his manager to reach potential interested parties in order to commercialize the algorithm. Then, he decided to search for existing solutions that could help him monetize the algorithm. Then, he came across ICARUS and decided to upload and share the algorithm through ICARUS.</p> <p>Daan accesses ICARUS and selects to upload a service. Then, Daan selects the specific file containing the custom-made service. Afterwards, he/she defines his/her custom proprietary license and sets the data visibility as proprietary and the pricing to 500€. Furthermore, he provides 4 KPIs (87% accuracy, 10 cores, 3GB RAM, 1 hour is required for training the algorithm) and a training and testing dataset about flight delays in order for his/her algorithm to be assessed. Subsequently, the service is tested based on the predefined KPIs of ICARUS and the KPIs provided by Daan. Afterwards, the ICARUS moderator checks the information of the service, the validity and the quality of the service based on the results of the assessment. Finally, the service is approved by the ICARUS moderator and it is stored in the ICARUS repository.</p>
Example Workflow Diagram	 <pre> graph LR Daan((Daan)) --> A[Daan uploads the algorithm for predicting the delays of flights, he defines the asset's terms and he provides 4 KPIs and the training and testing dataset] A --> B[Service is tested based on the predefined KPIs of ICARUS and the 4 KPIs provided by Daan] B --> C[The ICARUS moderator checks the validity and quality of service based on the KPIs] C --> D[Service asset is stored to the ICARUS repository] </pre> <p>Legend: ■ Service Collection ■ Asset Storage</p>
Exemplar Flow #2	
Example Name	An aircraft manufacturer wishes to upload a custom-made service (as confidential) in ICARUS for private use only.
Example Description	Justina is a senior software developer in "Advanced Aerospace Corporation", an aircraft manufacturer based in Lithuania with more than 400 employees. Currently, Justina has created an algorithm for predictive maintenance of the aircrafts. However, the company does not have all the

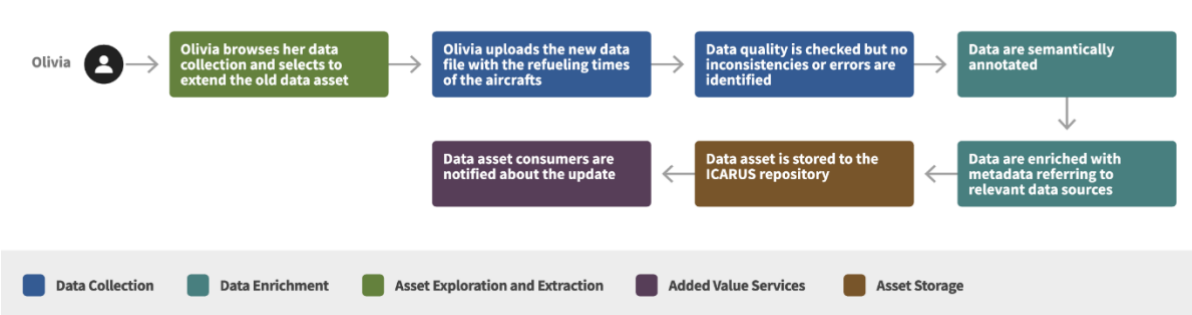
	<p>necessary resources to execute the algorithm. Thus, she was assigned by her manager to search for existing solutions in order to execute the algorithm. Through this process, Justina came across ICARUS and decided to upload her algorithm in ICARUS for private use only.</p> <p>Justina accesses ICARUS and selects to upload a service. Then, Justina selects the specific file containing the custom-made service. Afterwards, she sets the service visibility as confidential. Furthermore, she provides 3 KPIs (20 cores, 16GB RAM, 1 day for training) and a training and testing dataset with the aircraft maintenance data in order for her algorithm to be assessed. Subsequently, the service is tested based on the predefined KPIs of ICARUS and the KPIs provided by Justina. Afterwards, the ICARUS moderator checks the information of the service, the validity and the quality of the service based on the results of the assessment. Finally, the service is approved by the ICARUS moderator and it is stored in the ICARUS repository.</p>
Example Workflow Diagram	 <pre> graph LR Justina((Justina)) --> Step1[Justina uploads the algorithm for predictive maintenance of the aircrafts, she defines the asset's terms and she provides 3 KPIs and the training and testing dataset] Step1 --> Step2[Service is tested based on the predefined KPIs of ICARUS and the 3 KPIs provided by Justina] Step2 --> Step3[The ICARUS moderator checks the validity and quality of service based on the KPIs] Step3 --> Step4[Service asset is stored to the ICARUS repository] </pre> <p>Service Collection Asset Storage</p>
Exemplar Flow #3	
Example Name	A health research center uploads a custom-made service (as open) in ICARUS to contribute to the aviation research society.
Example Description	<p>Marie is a senior researcher in "European Health Center", a research organization based in Belgium with more than 200 employees. The organization has created an algorithm which predicts the spreading of epidemics and diseases. Marie was assigned by her supervisor to share the algorithm as open, in order to increase the reputation of the organization by contributing to the aviation research society. For that reason, she decided to search for existing solutions that could help her share the service. Then, she came across ICARUS and decided to upload and share the service through ICARUS.</p> <p>Marie accesses ICARUS and selects to upload a service. Then, Marie selects the specific file containing the custom-made service. Afterwards, she defines a "MIT" license and sets the service visibility as open. Furthermore, she provides 3 KPIs (5 cores, 12GB RAM, 5 days for training) and a training and testing dataset with spreading of epidemics and diseases data in order for her algorithm to be assessed. Subsequently, the service is tested based on the predefined KPIs of ICARUS and the KPIs provided by Marie. Afterwards, the ICARUS moderator checks the information of the service, the validity and the quality of the service based on the results of the assessment. Finally, the service is approved by the ICARUS moderator and it is stored in the ICARUS repository.</p>
Example Workflow Diagram	 <pre> graph LR Marie((Marie)) --> Step1[Marie uploads the algorithm for predicting the spreading of the epidemics and diseases, she defines the asset's terms and she provides 3 KPIs and the training and testing dataset] Step1 --> Step2[Service is tested based on the predefined KPIs of ICARUS and the 3 KPIs provided by Marie] Step2 --> Step3[The ICARUS moderator checks the validity and quality of service based on the KPIs] Step3 --> Step4[Service asset is stored to the ICARUS repository] </pre> <p>Service Collection Asset Storage</p>

3.5 Scenario 5: Data Update

ICARUS Scenario - [SCE-5]			
Scenario ID	SCE-5	Scenario Name	Data Update
Scenario Overview	<p>A data provider in the ICARUS data value chain wants to update a data asset and add a new version in ICARUS platform. Specifically, the data provider browses his collection of data assets and selects the data asset which needs to be updated. Then, he selects either to “extend” or “replace” the old data and uploads the new version of data. Afterwards, the data provider uses the ICARUS mechanisms for data anonymization, data quality check, data curation, semantic annotation and enrichment with metadata that refers to relevant sources. Subsequently, the new data are stored in the ICARUS repository. Finally, data consumers that have already purchased or are currently using an old version of these data, are notified about the new version in order to choose whether they want to receive the updated data or keep the old data (this depends on the updating policy specified by the data provider).</p>		
Triggers	A data provider wants to update one of his data asset stored in ICARUS.		
Scenario Workflow Diagram	 <pre> graph LR DP[Data Provider] --> B[Browse "My collection" of data assets and select to update a data set] B --> U[Upload new version of data] U --> A[Anonymize data] A --> C[Check data quality] C --> CU[Curate data] CU --> SA[Data are semantically annotated] SA --> EM[Data are enriched with metadata referring to relevant data sources] EM --> AS[Data asset is stored to the ICARUS repository] AS --> N[Notify data asset consumers about new version] </pre> <p>Legend: Data Collection (Blue), Data Enrichment (Teal), Asset Exploration and Extraction (Green), Added Value Services (Purple), Asset Storage (Brown)</p>		
Scenario Sequence	<ol style="list-style-type: none"> 1. The data provider browses his collection of data assets that exist in ICARUS repository and selects to update (“replace” or “extend”) one of the data assets. (Phase IV: Asset Exploration and Extraction, Step IV.1: Asset Indexing and Searching, Section 2.4.1) 2. The data provider uploads the new version of the data asset to ICARUS. (Phase I: Data Collection, Step I.1: Data Retrieval, Section 2.1.1) 3. The sensitive information in the data asset is anonymized with the data provider interaction. This step may be bypassed in case the data are already anonymized or if the data provider does not wish to share the data with others. (Phase I: Data Collection, Step I.2: Data Anonymization, Section 2.1.2) 4. The data quality is checked for inconsistencies and errors and the results are presented to the data provider. This step may be bypassed in case the data provider does not desire to share the data with other. (Phase I: Data Collection, Step I.3: Data Quality Check, Section 2.1.3) 5. Based on the result of the data quality check, the data are filtered and cleaned from inconsistencies and errors with the data provider interaction. This step may be bypassed in case the step of data quality check ensures that the data is of high quality or if the data 		

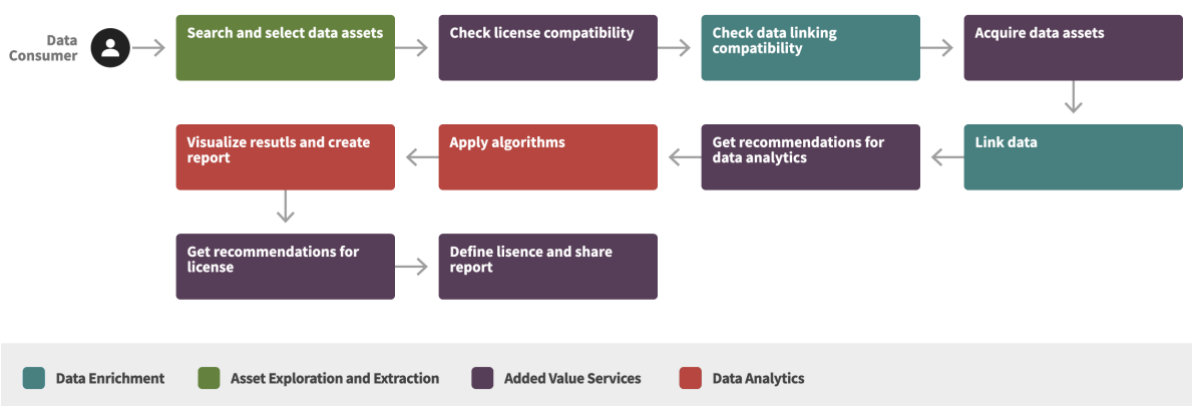
	<p>provider does not wish to change the data. (Phase I: Data Collection, Step I.4: Data Curation, Section 2.1.4)</p> <ol style="list-style-type: none"> Additional information to various concepts (e.g. people, locations, organizations, etc.) are attached to the data in order to enrich the data assets with machine-processable information by linking background information from aviation-related vocabularies, ontologies and semantic models. (Phase II: Data Enrichment, Step II.1: Semantic Enrichment and Annotation, Section 2.2.1) The data are enriched with metadata, referring to relevant data sources in ICARUS that may be linked together either directly or indirectly. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) The data asset is stored to the storage of ICARUS. (Phase III: Asset Storage, Section 2.3) The data consumers are notified about the updated data asset. (Phase VI: Added Value Services, Notifications, Section 2.6.3)
Users' Benefits	<ul style="list-style-type: none"> Easier updating and management of the data assets; Mechanism for notifying data consumers about the updated data asset; Easier and safer sharing of open or proprietary data that reach a wider audience either for monetisation or for contributing to the aviation research society; High performance storage; Availability of a wide variety of state-of-the-art anonymization and curation techniques in order to ensure the privacy and quality of the data; High level of automation in several parts of the workflow, thereby significantly reducing time required for tedious task like anonymization and curation; Increase the value and validity of the data through data curation.
Challenges	<ul style="list-style-type: none"> Ensure the anonymity and quality of the data; Connection of streaming data in a way that supports update of data with various processing tasks to be applied in real-time.
Exemplar Flow #1	
Example Name	A transport organization wishes to update its old data (buses and schedules) in ICARUS in order to increase the usefulness of the data asset.
Example Description	<p>Christina is a software developer in "Golden Bus", a transport organization based in Croatia with 40 employees. The organization has decided to update its data asset that already exists in ICARUS in order to increase its usefulness and validity. Thus, Christina was assigned by her manager to update the organization's old data in ICARUS.</p> <p>The dataset that Christina wishes to upload is a batch file which contains structured information about buses and schedules. Furthermore, the dataset contains 500 records and was collected over the period 2018. Christina accesses ICARUS and browses her "data collection" in order to select and replace the old data asset. Then, she selects and uploads the new data asset. After uploading the dataset, the ICARUS mechanism of data anonymization identifies unique identifiers of the buses. Thus, Christina selects to remove the sensitive information. Afterwards, she chooses to check the quality of the data using the data quality mechanism of ICARUS, but no inconsistencies or errors are identified. Afterwards, the semantic enrichment and metadata enrichment of relevant data sources</p>

	are performed by ICARUS. Finally, the new data asset is stored in ICARUS repository and the data consumers that are currently using the old data asset are notified about the update.
Example Workflow Diagram	 <pre> graph LR Christina((Christina)) --> A[Christina browses her data collection and selects to replace the old data asset] A --> B[Christina uploads the new data file with the buses information and schedules] B --> C[Christina removes the unique identifiers of the buses] C --> D[Data quality is checked but no inconsistencies or errors are identified] D --> E[Data are semantically annotated] E --> F[Data are enriched with metadata referring to relevant data sources] F --> G[Data asset is stored to the ICARUS repository] G --> H[Data asset consumers are notified about the update] </pre> <p>Legend: Data Collection (Blue), Data Enrichment (Teal), Asset Exploration and Extraction (Green), Added Value Services (Purple), Asset Storage (Brown)</p>
Exemplar Flow #2	
Example Name	A private security company wants to update its old data (crimes and threats in airports) in ICARUS in order to increase the outreach of the data asset.
Example Description	<p>Antoni is a junior data analyst in "AAA-Secure Patrol Services", a private security company based in Poland with 90 employees. The organization has decided to update its data asset that already exists in ICARUS in order to increase the outreach of the data asset. Thus, Antoni was assigned by his manager to update the organization's old data in ICARUS.</p> <p>The dataset that Antoni wishes to upload is a batch file which contains semi-structured information about crimes and threats in airports. Furthermore, the dataset contains 8.000 records and was collected over the period of 2017. Antoni accesses ICARUS and browses his "data collection" in order to select and "extend" the old data asset. Then, he selects and uploads the new data asset. After uploading the dataset, Antoni is asked to anonymize the data but bypasses the specific step since the data are already anonymized. Afterwards, he chooses to check the quality of the data using the data quality mechanism of ICARUS. The mechanism identified many duplicated records; thus, Antoni selects to use a deduplication technique which is provided by ICARUS. Afterwards, the semantic enrichment and metadata enrichment based on relevant data sources are performed by ICARUS. Finally, the new data asset is stored in ICARUS repository and the data consumers that had purchased the old data asset are notified about the update.</p>
Example Workflow Diagram	 <pre> graph LR Antoni((Antoni)) --> A[Antoni browses his data collection and selects to extend the old data asset] A --> B[Antoni uploads the new data file with the crimes and threats in airports] B --> C[Data quality is checked and many duplicates records are identified] C --> D[Antoni selects a deduplication technique for cleaning the data] D --> E[Data are semantically annotated] E --> F[Data are enriched with metadata referring to relevant data sources] F --> G[Data asset is stored to the ICARUS repository] G --> H[Data asset consumers are notified about the update] </pre> <p>Legend: Data Collection (Blue), Data Enrichment (Teal), Asset Exploration and Extraction (Green), Added Value Services (Purple), Asset Storage (Brown)</p>
Exemplar Flow #3	
Example Name	An aircraft refueller company desires to update its old data (refuelling times of the aircrafts) in ICARUS in order to increase the outreach of the data asset.
Example Description	Olivia is a data analyst in "AERO Fuel", an aircraft refueller company based in Sweden with more than 350 employees. The organization has decided to update its data asset that already exists in ICARUS

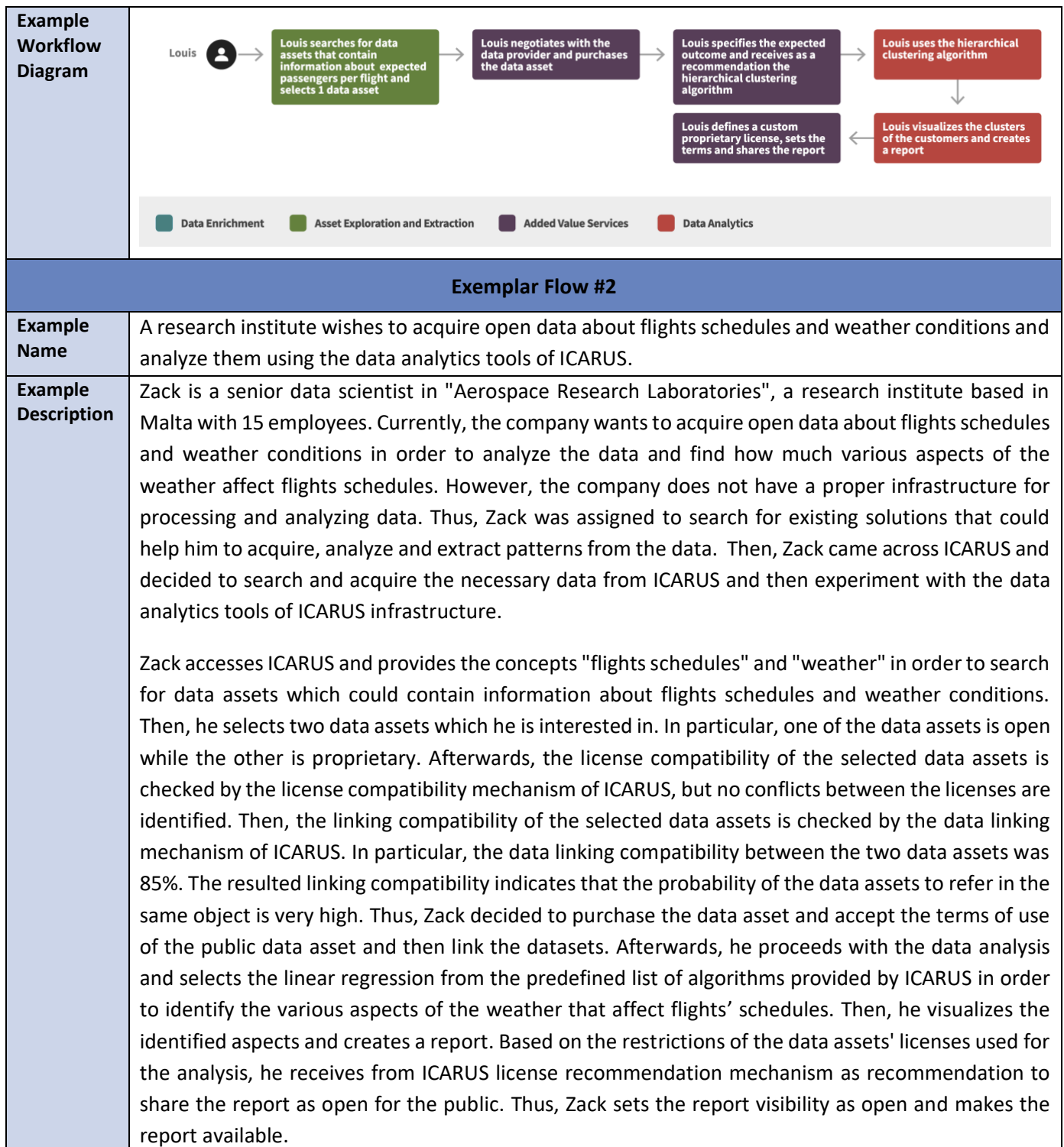
	<p>in order to increase the usefulness of the data asset. Thus, Olivia was assigned by her manager to update the organization's old data in ICARUS.</p> <p>The dataset that Olivia wishes to upload is a batch file which contains structured information about the refueling times of the aircrafts. Furthermore, the dataset contains 70.000 records and was collected over the period of 2017-2018. Olivia accesses ICARUS and browses her “data collection” in order to select and “extend” the old data asset. Then, she selects and uploads the new data asset. After uploading the dataset, Olivia is asked to anonymize the data but bypasses the specific step since the data are already anonymized. Afterwards, she chooses to check the quality of the data using the data quality mechanism of ICARUS, but no inconsistencies are identified. Afterwards, the semantic enrichment and metadata enrichment based on relevant data sources are performed by ICARUS. Finally, the new data asset is stored in ICARUS repository and the data consumers that had purchased the old data asset are notified about the update.</p>
Example Workflow Diagram	 <pre> graph LR Olivia((Olivia)) --> A[Olivia browses her data collection and selects to extend the old data asset] A --> B[Olivia uploads the new data file with the refueling times of the aircrafts] B --> C[Data quality is checked but no inconsistencies or errors are identified] C --> D[Data are semantically annotated] D --> E[Data are enriched with metadata referring to relevant data sources] E --> F[Data asset is stored to the ICARUS repository] F --> G[Data asset consumers are notified about the update] </pre> <p>Legend:</p> <ul style="list-style-type: none"> Data Collection (Blue) Data Enrichment (Teal) Asset Exploration and Extraction (Green) Added Value Services (Purple) Asset Storage (Brown)

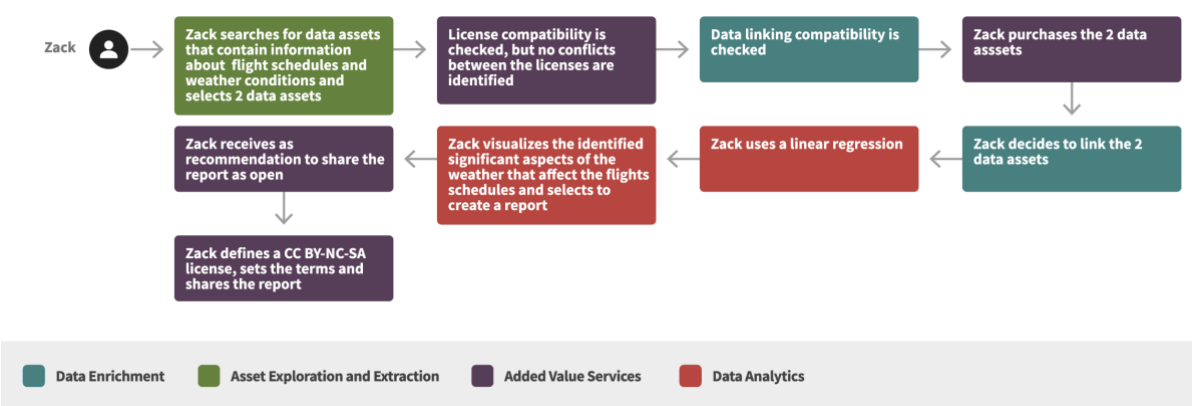
3.6 Scenario 6: Data Exploration, Analysis and Sharing

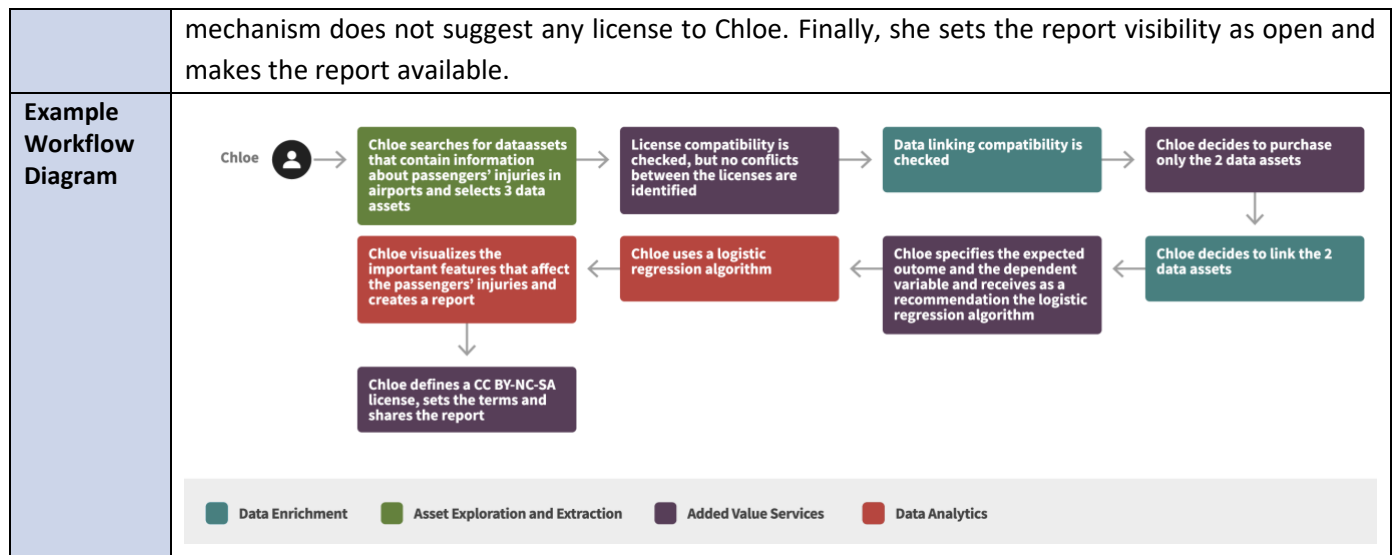
ICARUS Scenario - [SCE-6]			
Scenario ID	SCE-6	Scenario Name	Data Exploration, Analysis and Sharing
Scenario Overview	<p>A data consumer in the ICARUS data value chain wants to explore and acquire many data assets through ICARUS and then analyze them in order to extract insights. Specifically, the data consumer explores ICARUS repository and selects the data assets which he/she is interested in. Afterwards, a license compatibility check is performed so as to ensure the data providers' IPRs. Subsequently, a data compatibility check is performed to confirm the data linking compatibility. If the data assets are proprietary, then the data consumer must purchase the data assets or negotiate with the data provider before purchase. Afterwards, the data consumer can proceed with the data analytics to apply various algorithms and visualizations. In addition, he can specify the expected outcome prior to the analysis in order to receive recommendations about data analytics algorithms and visualizations. Finally, he is able to create a report, define a license for the report based on license recommendations and share it to potential stakeholders.</p>		
Triggers	<p>A data consumer wants to acquire some data assets, perform various analytics on them through ICARUS mechanisms and then share the outcomes.</p>		

Scenario Workflow Diagram	 <pre> graph TD DC((Data Consumer)) --> S1[Search and select data assets] S1 --> S2[Check license compatibility] S2 --> S3[Check data linking compatibility] S3 --> S4[Acquire data assets] S4 --> S5[Link data] S5 --> S6[Get recommendations for data analytics] S6 --> S7[Apply algorithms] S7 --> S8[Visualize results and create report] S8 --> S9[Get recommendations for license] S9 --> S10[Define licence and share report] </pre> <p>Legend: Data Enrichment (Teal), Asset Exploration and Extraction (Green), Added Value Services (Purple), Data Analytics (Red)</p>
Scenario Sequence	<ol style="list-style-type: none"> 1. The data consumer searches the ICARUS repository for open or proprietary data assets and selects the data assets in which she/he is interested in. (Phase IV: Asset Exploration and Extraction, Step IV.1: Asset Indexing and Searching, Section 2.4.1) 2. The license compatibility of the selected data assets is checked in order to avoid conflicts between the data assets' terms and licenses. This step may be bypassed in case the data provider does not wish to link the data assets. (Phase VI: Added Value Services, Asset Sharing, Section 2.6.1) 3. The data linking compatibility of the selected data assets is checked. This step may be bypassed in case the data provider does not wish to link the data assets. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) 4. The data consumer acquires the data assets. If a data asset is proprietary, then the data consumer must purchase and/or negotiate with the data provider. (Phase VI: Added Value Services, Asset Sharing, Section 2.6.1) 5. The acquired data assets are linked based on the data consumer's decision. This step may be bypasses in case the data provider does not wish to link his data. (Phase II: Data Enrichment, Step II.2: Data Linking, Section 2.2.2) 6. The data consumer receives recommendations about the methodology to follow for data analysis (data pre-processing, algorithms, visualizations) based on the specified expected outcome. This step may be bypassed in case the data provider does not wish to specify the expected outcome. (Phase VI: Added Value Services, Recommendations, Section 2.6.2) 7. The data consumer applies different algorithms and statistical methods to the data. (Phase V: Data Analytics, Step V.1: Data Analysis, Section 2.5.1) 8. The data consumer visualizes the results of the analysis and creates an aviation DVC (data value chain) report that contains the results of the analysis. (Phase V: Data Analytics, Step V.2: Data Visualization, Section 2.5.2) 9. The data consumer receives recommendations for licenses in order to share the report. (Phase VI: Added Value Services, Recommendations, Section 2.6.2) 10. The data consumer defines the terms (e.g. data visibility, license, pricing) which are more suitable for the specific asset and shares the aviation DVC report. (Phase VI: Added Value Services, Asset Sharing, Section 2.6.1)
Users' Benefits	<ul style="list-style-type: none"> • Easier discoverability and access to a wide variety of useful data assets; • Easier negotiation and purchase of data assets;

	<ul style="list-style-type: none"> • Availability of a wide variety of state-of-the-art machine learning algorithms and visualizations; • Efficient mechanisms for processing and analyzing data; • Reduced in-house development time and effort; • Recommendations for algorithms and visualizations that make it even easier for novice users to analyze data; • Easy experimentation and comparison of results of analysis; • High level of automation in several parts of the workflow, thereby significantly reducing time required; • Easier sharing of data analysis results by using reports.
Challenges	<ul style="list-style-type: none"> • Discover data assets with compatible licenses; • Discover data assets with high level of data linking compatibility; • Use of the appropriate analytical methodology (data preparation, data analysis algorithm and visualizations) in order to achieve the expected outcome; • Real-time requirements for data analysis; • Configurable visualizations and visual analytics features for data; • Define the right terms and license for sharing a report based on the data assets used in the analysis.
Exemplar Flow #1	
Example Name	A travel agency wishes to acquire data about the expected and checked passengers per flight and experiment with the data analytics tools of ICARUS.
Example Description	<p>Louis is a junior data scientist in "Oceanic Destinations", a travel agency based in Luxembourg with more than 150 employees. Currently, the company needs to acquire data about the expected and checked passengers per flight in order to analyze the data and provide new offers on the customers. Thus, Louis was assigned to find, analyze and extract patterns from the data. However, Louis finds it quite challenging to acquire such dataset, so he decided to search for existing solutions that could help him acquire the data. He came across ICARUS and decided to search and acquire the necessary datasets from ICARUS and then experiment with the data analytics tools of ICARUS.</p> <p>Louis accesses ICARUS and provides the concept "expected passengers" in order to search for data assets which could contain information about the expected passengers per flight. Then, he selects one data asset which he is interested in and negotiates with the data provider in order to purchase the data asset. Afterwards, he proceeds with the data analysis and specifies his expected outcome, which is to group the different characteristics of customers based on the flights destinations. Then, based on the specified expected outcome, the recommendation mechanism of ICARUS for data analytics suggests to Louis the hierarchical clustering algorithm. Louis selects the hierarchical clustering algorithm from the predefined list of algorithms provided by ICARUS and then, he visualizes the created clusters of the passengers. Afterwards, he creates a report that describes his findings and results. Since the license of the data asset used for the analysis does not impose any restrictions on the exchange of derivative data, the ICARUS license recommendation mechanism does not suggest any license to Louis. Thus, Louis defines a custom proprietary license. Finally, he sets the report visibility as proprietary and the pricing to 30€ and makes the report available.</p>



Example Workflow Diagram	 <pre> graph TD Zack((Zack)) --> A[Zack searches for data assets that contain information about flight schedules and weather conditions and selects 2 data assets] A --> B[License compatibility is checked, but no conflicts between the licenses are identified] B --> C[Data linking compatibility is checked] C --> D[Zack purchases the 2 data assets] D --> E[Zack decides to link the 2 data assets] E --> F[Zack uses a linear regression] F --> G[Zack visualizes the identified significant aspects of the weather that affect the flights schedules and selects to create a report] G --> H[Zack receives as recommendation to share the report as open] H --> I[Zack defines a CC BY-NC-SA license, sets the terms and shares the report] </pre> <p>Legend:</p> <ul style="list-style-type: none"> Data Enrichment (Teal) Asset Exploration and Extraction (Green) Added Value Services (Purple) Data Analytics (Red)
Exemplar Flow #3	
Example Name	An aviation insurance company wants to acquire data about passengers' injuries and aircraft accidents from various airports and analyze them using the data analytics tools of ICARUS.
Example Description	<p>Chloe is a senior data analyst in "Aero-Insurance", an aviation insurance company based in Ireland with more than 450 employees. Currently, the company wants to acquire data about passengers' injuries and aircraft accidents from various airports in order to analyze them and provide better contracts terms to the companies. Thus, Chloe was assigned to acquire, analyze and extract patterns from the data. However, Chloe finds it quite challenging to acquire a worldwide dataset from various airports. With limited choices, she decided to search for existing solutions that could help her acquire the data. Then, she came across ICARUS and decided to search and acquire the necessary data from ICARUS. Using the acquired data assets, she will experiment with the data analytics tools of ICARUS platform.</p> <p>Chloe accesses ICARUS and provides the concept "accidents in airports" in order to search for data assets which could contain information about passengers' injuries and aircraft accidents from various airports. Then, she selects three data assets (accidents in airports, airlines total passengers per flight, airports total passenger) which she is interested in. Particularly, all of the selected data assets are proprietary. Afterwards, the license compatibility of the selected data assets is checked by the license compatibility mechanism of ICARUS, but no conflicts between the licenses are identified. Then, the linking compatibility of the selected data assets is checked by the data linking mechanism of ICARUS. In particular, the data linking compatibility between the three data assets is 55%. The resulted linking compatibility indicates that the probability of the data assets to refer in the same object is low. Thus, she decides to remove the data asset that contains information about airlines total passengers per flight because the period collected is different than the other two data assets. Subsequently, Chloe decides to purchase the data assets and link them. Afterwards, she proceeds with the data analysis and specifies her expected outcome, which is to find the important features that may affect the passengers' injuries. Furthermore, she also specifies the dependent variable for the analysis (if a passenger has been involved in an accident). Then, based on the specified expected outcome and dependent variable, the recommendation mechanism of ICARUS for data analytics suggests to Chloe the logistic regression algorithm. Thus, she selects the logistic regression algorithm from the predefined list of algorithms provided by ICARUS. Afterwards, she visualizes the important features that affect the passengers' injuries. Since the license of the data assets utilized for the analysis does not impose any restrictions on the exchange of derivative data, ICARUS license recommendation</p>



4 ICARUS MVP Definition

In preparation of the design and development phases of the project, the ICARUS consortium has initiated the definition of a Minimum Viable Product (MVP) in order to ensure that the ICARUS platform is designed and viewed as a product with enough features to satisfy early customers, minimizing the risk of failure and improving the value generated. The MVP is directly linked with the lean start-up approach in the design and development, which is in full alignment with the concept and work plan of ICARUS, especially when taking into account the limited resources and the short project duration.

Typically, an MVP is a product which has the highest return on investment versus risk and is used in order to move fast towards the prototyping phase, without investing effort on features and functionalities which could hamper the overall development due to low user acceptance, high complexity, and lack of alignment to the actual users' needs. In contrast to the usual practices that view the MVP as a critical asset in the prototyping phase, the MVP in ICARUS is instrumental to guide the design and development activities throughout the project implementation and represents the platform release that will be delivered in the end of the project. In essence, the ICARUS MVP represents the overall mindset and process adopted for product development¹ to continuously test the customer reaction, to deliver customer value and validate the methodological ideas and hypothesis. In this context, it needs to be noted that even if the MVP pinpoints the minimum set of features that are necessary for a product to be deployed and validated, it does not dictate the ICARUS consortium to seize their work when reaching that state; on the contrary, the MVP is a strategy for cutting out unnecessary spending, but being able to quickly learn about the aviation data value chain and what can be sold to them, resulting in this manner into a more viable, profitable and successful platform that shall be gradually improved and populated with extra features. To this direction, as the MVP will be only the primer for the future exploitation activities and the future product development, it is viewed as the mechanism to collect early feedback and appropriately steer the design and development activities. For this reason, many features or offerings might be tested manually (in WP3 and WP4), by simulating a process through mock-ups and with humans partially replacing the algorithms.

The present section aims at extracting a concrete set of features and assessing their added value in order to perform a preliminary prioritization. Since this deliverable represents the initial release of the MVP definition activities, such an assessment is initially conducted within the consortium, with the plan to expand to external stakeholders from the aviation domain in the final release (in D1.3, which is expected in M15). The MVP shall take into consideration that the support of certain activities/features is mandatory for the implementation of others which are dependent on them and is expected to evolve into prioritized user stories in WP3. The MVP will be thus tested and updated from the different activities in WP1, WP3 and WP5 through a set of questionnaires following a light "market-research" study, which will further streamline the initial release. Through the feedback acquired by external stakeholders, the consortium will consistently work towards adding new features and functions, in order to improve the overall platform and deliver a final solution that facilitates to the highest possible degree the ICARUS methodology and concept. In summary, as depicted in the following figure, the approach that is applied for the MVP definition bears three core phases, namely Feature Definition, Feature Assessment and MVP Consolidation (cross-cutting the WP1 and WP3 activities) and runs over the two iterations of WP1.

¹ <https://dzone.com/articles/minimum-viable-product-is-not-a-product-but-a-mind>

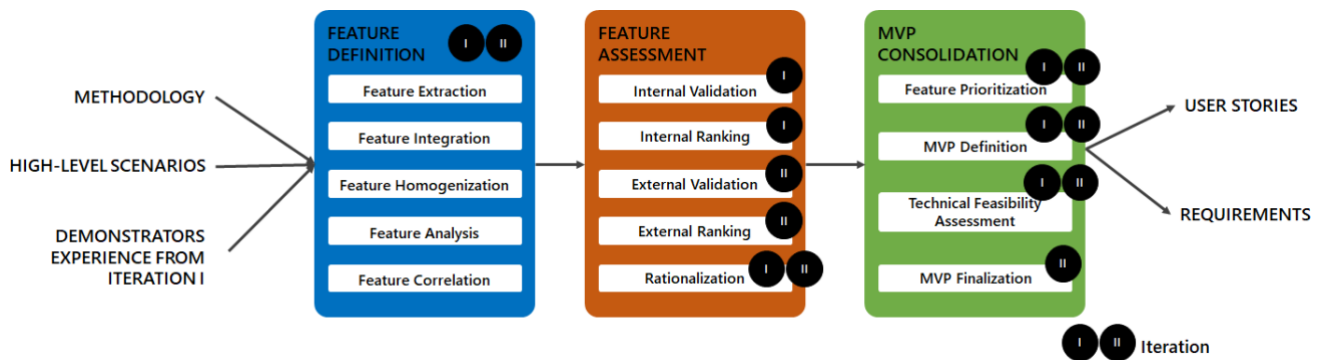


Figure 4-1: ICARUS MVP Approach

4.1 Features Extraction, Integration and Homogenization

Taking into account the methodology that has been elaborated in section 2 and the high-level scenarios that anticipate the ICARUS platform use from the stakeholders' point of view, the implied features that have been identified in section 4.1 are extracted, grouped and homogenized in the next lines. It needs to be noted that the relation to the methodology is defined at phase and step levels while the relation to the high-level scenarios is broadly defined at their different steps.

Due to the dual roles of the targeted aviation data stakeholders in ICARUS (i.e. data providers and data consumers), the MVP needs to be balanced in order to satisfy both their needs and this is why it has been described from the platform perspective.

PLATF_F_01. Retrieval of data directly from an aviation stakeholder's back-end system

Description	The ICARUS platform will directly retrieve data from the back-end system of any aviation data provider via well-defined APIs. As a data provider initiates the check in for the specific data asset, he/she needs to provide not only the API calls to retrieve the necessary information, but also the API restrictions and the link to the relevant API documentation.
Methodology Phase	I. Data Collection – I.1 Data Retrieval
Related Scenarios	SCE-2.1
Prerequisites	-

PLATF_F_02. Uploading of data assets as files extracted by the aviation stakeholder's back-end system

Description	The ICARUS platform will collect data from the aviation data providers as files (e.g. csv, json, xml, etc.) that are uploaded through easy-to-use interfaces.
Methodology Phase	I. Data Collection – I.1 Data Retrieval
Related Scenarios	SCE-1.1, SCE-2.1
Prerequisites	-

PLATF_F_03. Provision of links where open data assets are available to be uploaded in ICARUS

Description	The ICARUS platform will ingest data that are published in open data portals via links that can be provided by any aviation stakeholder.
Methodology Phase	I. Data Collection – I.1 Data Retrieval

Related Scenarios	SCE-6.1
Prerequisites	-

PLATF_F_04. Detailed data profiling of the data to be uploaded according to a specific metadata schema

Description	The ICARUS platform will require the profiling of the data assets to be uploaded to different levels of detail depending on whether they are intended to be public, private or confidential. Such a profiling will comply to the ICARUS metadata schema.
Methodology Phase	I. Data Collection - I.1 Data Retrieval, I.5 Data Check in
Related Scenarios	SCE-1.1, SCE-2.1, SCE-3.1
Prerequisites	PLATF_F_01-02-03

PLATF_F_05. "Fast lane" check-in for confidential data assets

Description	The ICARUS platform will allow for a quick uploading and check-in process for the confidential data assets that are only intended for use by the respective data provider. A minimum set of metadata needs to be provided in this case in order to facilitate the subsequent data enrichment and linking since the full range of metadata regarding the data asset (e.g. related to data sharing and its availability) are not necessary for data assets that are intended for internal consumption only (by their rightful owner).
Methodology Phase	I. Data Collection - I.1 Data Retrieval, I.5 Data Check in
Related Scenarios	SCE-1.1, SCE-2.1
Prerequisites	PLATF_F_01-02

PLATF_F_06. (Semi-)Automatic quality check of the data and assessment of quality level

Description	The ICARUS platform will provision for a semi-automatic quality check of the data assets that are uploaded by a data provider. Such a quality check results into assessing the quality level of an asset.
Methodology Phase	I. Data Collection – I.3 Data Quality Check
Related Scenarios	SCE-1.3, SCE-2.3, SCE-3.3, SCE-5.4
Prerequisites	PLATF_F_01-02-03

PLATF_F_07. Guidelines / recommendations for improving the data quality

Description	The ICARUS platform will provide a set of practical recommendations / guidelines for improving the quality of a data asset that is being uploaded, depending on the quality level it has classified an asset (in PLATF_F_06).
Methodology Phase	I. Data Collection – I.3 Data Quality Check
Related Scenarios	SCE-1.3, SCE-2.3, SCE-3.3, SCE-5.4
Prerequisites	PLATF_F_06

PLATF_F_08. Manual completion of an anonymization "check-list"

Description	The ICARUS platform will request from the aviation data providers to fill in a check-list related to the privacy of the data asset they are about to upload and its compliance with the EC GDPR.
Methodology Phase	I. Data Collection – I.2 Data Anonymization
Related Scenarios	SCE-1.2, SCE-2.2, SCE-3.2, SCE-5.3
Prerequisites	PLATF_F_01-02

PLATF_F_09. (Semi-)Automatic on-the-fly anonymization in ICARUS

Description	The ICARUS platform will allow the aviation data providers to semi-automatically mask and obfuscate the data asset they are about to upload in order to ensure the privacy of the data subjects (e.g. passengers).
Methodology Phase	I. Data Collection – I.2 Data Anonymization
Related Scenarios	SCE-1.2, SCE-2.2, SCE-3.2, SCE-5.3
Prerequisites	PLATF_F_01-02

PLATF_F_10. Offline masking (anonymization) of "shareable" (public and private) data prior to uploading in ICARUS

Description	The ICARUS platform will provide to the related data providers a set of offline tools they may utilize in order to anonymize the data assets they are about to upload, especially if they are to be shared.
Methodology Phase	I. Data Collection – I.2 Data Anonymization
Related Scenarios	SCE-1.2, SCE-2.2, SCE-3.2, SCE-5.3
Prerequisites	PLATF_F_01-02

PLATF_F_11. (Semi-)Automatic extraction of and navigation within the Data Model of a Data Asset

Description	The ICARUS platform will semi-automatically identify the data model to which a data asset complies and will allow the respective data provider to navigate to the extracted model.
Methodology Phase	II. Data Enrichment – II.1 Semantic Enrichment and Annotation
Related Scenarios	SCE-1.5, SCE-2.6, SCE-3.6, SCE-5.6
Prerequisites	PLATF_F_04-05

PLATF_F_12. (Semi-)Automatic extraction of concepts from within the Data Asset

Description	The ICARUS platform will semi-automatically extract concepts that appear in a data asset. Such concepts may refer to a variable (for example, the flight scheduled arrival time) or the value of a variable (e.g. that ATH stands for the Athens International Airport).
Methodology Phase	II. Data Enrichment – II.1 Semantic Enrichment and Annotation
Related Scenarios	SCE-1.5, SCE-2.6, SCE-3.6, SCE-5.6
Prerequisites	PLATF_F_11

PLATF_F_13. (Semi-) Automatic Transformation / Mapping of data assets and extracted concepts to the ICARUS common schema

Description	The ICARUS platform will semi-automatically map both the data asset and the extracted concepts that appear in the specific data asset to the ICARUS common aviation schema. Such a mapping will be instrumental to ensure the alignment and linking of the different aviation-related data assets that have been already checked in in ICARUS.
Methodology Phase	II. Data Enrichment – II.1 Semantic Enrichment and Annotation
Related Scenarios	SCE-1.6, SCE-2.7, SCE-3.7, SCE-5.7
Prerequisites	PLATF_F_11-12-14-15

PLATF_F_14. Easily applicable data manipulation / transformation methods

Description	The ICARUS platform will provision for easily applicable data manipulation and transformation methods. Such methods may range from: format transformation (any format to csv, provided that the data quality is ensured) and data measurement unit transformation (e.g. from metre to foot, that can happen automatically based on the associated metadata or manually based on the data provider's input) to typical data manipulation functions (like replace, reduce, groupby, etc).
Methodology Phase	I. Data Collection – I.4 Data Curation & II. Data Enrichment – II.1 Semantic Enrichment and Annotation
Related Scenarios	SCE-1.4, SCE-1.5, SCE-2.4, SCE-2.6, SCE-3.4, SCE-3.6, SCE-5.5
Prerequisites	PLATF_F_11

PLATF_F_15. Easily applicable data cleaning methods

Description	The ICARUS platform will provision for easily applicable data cleaning methods. Such methods may range from handling missing data values and conformance to specific data types to handling outliers.
Methodology Phase	I. Data Collection – I.4 Data Curation & Enrichment
Related Scenarios	SCE-1.4, SCE-2.4, SCE-3.4, SCE-5.5
Prerequisites	PLATF_F_11

PLATF_F_16. Semantic enrichment of the data assets and extracted concepts with machine-processable information

Description	The ICARUS platform will semantically enrich the data assets provided by their respective data providers at different granularity levels (at schema level and at actual data level) with the help of aviation-related vocabularies, ontologies and semantic models.
Methodology Phase	II. Data Enrichment – II.1 Semantic Enrichment and Annotation
Related Scenarios	SCE-1.6, SCE-2.7, SCE-3.7, SCE-5.7
Prerequisites	PLATF_F_13

PLATF_F_17. Provision of suggestions for the semantic enrichment of the data assets and the extracted concepts

Description	The ICARUS platform will provide recommendations for the semantic enrichment of each data asset (both as a whole and with regard to its individual, extracted concepts) to its respective provider in order to facilitate its subsequent linking to other data assets that have been already checked in.
Methodology Phase	II. Data Enrichment – II.1 Semantic Enrichment and Annotation
Related Scenarios	SCE-1.6, SCE-2.7, SCE-3.7, SCE-5.7
Prerequisites	PLATF_F_16-55

PLATF_F_18. Searchability and identification of related additional data assets

Description	The ICARUS platform will be able to search (both on request when a data asset is checked in and periodically, in scheduled intervals) and identify related data assets that have been provided either by the same data provider or by different data providers in order to link them and facilitate their subsequent analysis. Such related data assets will be identified on the basis of common variables (e.g. destination, airport, timestamp, etc.) that appear in their respective data schemas and in relation to the ICARUS aviation data model.
Methodology Phase	II. Data Enrichment – II.2 Data Linking
Related Scenarios	SCE-6.3
Prerequisites	PLATF_F_16-55

PLATF_F_19. Indication of the similarity probability, namely the level of confidence of the similarity so that linking is meaningful and efficient

Description	The ICARUS platform will define the level of confidence (also referred to as similarity probability) on the linking of different data assets that is automatically calculated according to different criteria (related to the data asset as a whole, its schema and its contents in terms of data and extracted concepts). Such an indicator will provide to the respective data providers a clear evidence of whether the proposed linking is expected to be meaningful and efficient.
Methodology Phase	II. Data Enrichment – II.2 Data Linking
Related Scenarios	SCE-6.3
Prerequisites	PLATF_F_18

PLATF_F_20. Indication of the linkable denominators, upon which linking of the data assets can be performed

Description	The ICARUS platform will highlight which are the potentially linkable denominators in terms of variables that can be directly linked across different related data assets (at 1 st level of linking). This linking process also indirectly inherits all the links that the related data assets already have with other data assets (2 nd level of linking).
Methodology Phase	II. Data Enrichment – II.2 Data Linking
Related Scenarios	SCE-6.3
Prerequisites	PLATF_F_18

PLATF_F_21. (Semi-)Automatic data asset linking

Description	The ICARUS platform will semi-automatically link the data assets that have been identified as related in accordance with the data providers' preferences. For example, a data provider may decide to link his data asset with another one to which the ICARUS platform has identified as highly linkable (with a high level of confidence) and skip linking with other data assets that are medium or low linkable. In the latter case, a data consumer is also able to link additional data assets at the analytics phases. It needs to be noted that the linking process does not result into the generation of a new data asset, but only into the creation of a permanent link (join) between the data assets.
Methodology Phase	II. Data Enrichment – II.2 Data Linking
Related Scenarios	SCE-1.8, SCE-2.10, SCE-6.5
Prerequisites	PLATF_F_20-55

PLATF_F_22. Definition of simple and advanced "information" queries

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to define simple and advanced "information" queries. Such queries will be defined with the help of appropriate query editors and search forms, and will result into providing a list of data assets that better correspond to the information searched.
Methodology Phase	IV. Asset Exploration & Searching – IV.1 Asset Indexing & Searching
Related Scenarios	SCE-6.1
Prerequisites	-

PLATF_F_23. Navigation to the list and "cover" of data assets that fit to a certain "information" query, instantly checking their similarity degree

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to navigate to the results of the "information" query they have defined in different ways, e.g. through a list of all the results, through a list of results filtered based on different criteria (e.g. their similarity degree to the query) and through a glimpse to their actual data contained whenever permitted by the respective data providers.
Methodology Phase	IV. Asset Exploration & Searching – IV.1 Asset Indexing & Searching
Related Scenarios	SCE-6.1
Prerequisites	PLATF_F_22

PLATF_F_24. Including pins and favourites among the ICARUS data assets

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to pin and favourite important data assets for easy reference.
Methodology Phase	IV. Asset Exploration & Searching – IV.1 Asset Indexing & Searching
Related Scenarios	SCE-6.1
Prerequisites	-

PLATF_F_25. Filtering of data assets based on different criteria

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to filter the list of data assets to which they are navigating (e.g. recent data assets or list from search results) according to different criteria (e.g. temporal or spatial coverage, linking to a specific data asset).
Methodology Phase	IV. Asset Exploration & Searching – IV.1 Asset Indexing & Searching
Related Scenarios	SCE-6.1
Prerequisites	PLATF_F_23

PLATF_F_26. Access and inspection of data assets "extracts" depending on their license

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to access the data assets of their selection and inspect their contents (in practice, specific data assets extracts as selected by their respective data providers). Such an inspection is dependent on the data license and whether it is allowed by the data provider.
Methodology Phase	IV. Asset Exploration & Searching – IV.1 Asset Indexing & Searching
Related Scenarios	SCE-6.1
Prerequisites	PLATF_F_23-55

PLATF_F_27. Transformation of a data asset to other supported data formats and export

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to transform a data asset to other data formats that are supported (e.g. from csv to xml) and export the outcome, as long as it is permitted by the respective data license which the data provider has defined.
Methodology Phase	IV. Asset Exploration & Searching – IV.2 Asset Export
Related Scenarios	SCE-1.9
Prerequisites	PLATF_F_13

PLATF_F_28. Navigation to preconfigured analytics

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to navigate in an intuitive manner to analytics that have been preconfigured for specific aviation stakeholders' profiles and operational purposes. For example, on the basis of the ICARUS demonstrators, different preconfigured analytics will be available for the airports' operational efficiency and any interested stakeholder can quickly understand the results that can be acquired along the related KPIs to indicate the improvements achieved. If such an interested stakeholder-airport wishes to reuse such analytics, the algorithms that are pre-trained in ICARUS will be available to use with their own data.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_33

PLATF_F_29. Definition of an analytic task that runs an individual algorithm

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to define the analytic task they want to run (either once or periodically) by selecting the data assets and the individual algorithms for execution. Such analytics may range from core statistics to different machine learning algorithms.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	-

PLATF_F_30. Definition of a workflow of analytic tasks that combine algorithms

Description	The ICARUS platform will allow the aviation stakeholders (both providers and consumers) to define the analytic task they want to run (either once or periodically) by selecting the related data assets and the workflow of the algorithms for execution (e.g. in sequence or in parallel).
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	-

PLATF_F_31. Automatic check whether the data asset is appropriate for a specific algorithm

Description	The ICARUS platform will automatically check whether the selected data assets are appropriate for the analysis an aviation stakeholder (both providers and consumers) wishes to perform (in PLATF_F_29 and PLATF_F_30). For example, if the data asset does not contain any time-series data and the data consumer wants to execute a time-series prediction, the ICARUS platform will directly inform him/her that such an analysis is not possible.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.11, SCE-2.12, SCE-6.6, SCE-6.7
Prerequisites	PLATF_F_29-30

PLATF_F_32. Automatic check for data licences compatibility to run under a specific algorithm

Description	The ICARUS platform will automatically check whether the selected data assets have compatible licences for analysis (to be performed in PLATF_F_29 and PLATF_F_20). If the data assets' licenses are not compatible, the ICARUS platform will directly inform the data consumer that the required analysis is not possible to be performed.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.11, SCE-2.12, SCE-6.6, SCE-6.7
Prerequisites	PLATF_F_29-30-55

PLATF_F_33. Support for pre-trained core analytics algorithms

Description	The ICARUS platform will feature specific core algorithms that have been pre-trained with available data assets. Depending on the data licenses, such pre-trained algorithms may not be available for free in the ICARUS platform.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_29

PLATF_F_34. Support for ICARUS variations of analytics algorithms, pre-trained for specific actions/insights

Description	The ICARUS platform will feature variations of analytics algorithms that have been created in the context of the project and have been pre-trained with available data assets. The performance and accuracy of such algorithms typically exceeds the performance and accuracy of the core algorithms (see PLATF_F_33). Depending on the data licenses, such pre-trained variations of algorithms may not be available for free in the ICARUS platform.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_33

PLATF_F_35. Execution of an analytics task / algorithm according to specific preferences and settings for computation resources

Description	The ICARUS platform will allow the execution of an analytics task and its corresponding algorithm(s) (as defined in PLATF_F_29 and PLATF_F_30) according to specific preferences the data consumer may have (e.g. sacrificing efficiency / performance for precision / accuracy or vice versa) and to the settings for computation resources that are available depending on the user's account.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_29-30

PLATF_F_36. Association of each analytics task with a level of confidence to the results

Description	The ICARUS platform will calculate the level of confidence that the analytics task's outcomes have, e.g. taking into account the performance of the training data assets.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_35

PLATF_F_37. Comparison of the results of different algorithms

Description	The ICARUS platform will allow for comparison of the results of different algorithms for the same data asset side-by-side and will highlight different performance-related parameters depending on the algorithms' family and specific characteristics.
Methodology Phase	V. Data Analytics – V.1 Data Analysis

Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_35

PLATF_F_38. Visualization of the analytics results to gain insights on the data and / or comparison how the same results are visualized in different diagrams

Description	The ICARUS platform will visualize the results of any analytics task in order to provide intuitive insights on the data. The ICARUS platform may also depict how the same results can be visualized in different diagrams in order to allow the data consumer to select the data diagram that is more attuned to its needs and communicates more clearly the results.
Methodology Phase	V. Data Analytics – V.2 Data Visualization
Related Scenarios	SCE-2.13, SCE-6.8
Prerequisites	PLATF_F_35

PLATF_F_39. Definition of customized dashboards by selecting which visualizations should appear

Description	The ICARUS platform will allow the aviation data stakeholders (data providers and data consumers) to define their own customized dashboards by selecting which visualizations and for which analytics task should appear. Such customized dashboards should be at their disposal in their future visits to the ICARUS platform.
Methodology Phase	V. Data Analytics – V.2 Data Visualization
Related Scenarios	SCE-2.13, SCE-6.8
Prerequisites	PLATF_F_38

PLATF_F_40. Definition of an end-to-end workflow / recipe

Description	The ICARUS platform will allow the aviation data stakeholders (data providers and data consumers) to define their end-to-end workflows / recipes that consist of specific data that comply with certain data schemas (under the ICARUS aviation data model), specific interwoven chains of algorithms and visualizations. Such workflows and recipes can become available in the ICARUS platform as templates that can be reused (under specific licences, so not necessarily for free) by other interested stakeholders.
Methodology Phase	V. Data Analytics – V.1 Data Analysis & V.2 Data Visualization
Related Scenarios	SCE-2.12, SCE-2.13, SCE-6.7, SCE-6.8
Prerequisites	PLATF_F_11-35-38-55

PLATF_F_41. Export of analytics results in machine-readable format (via an API, or as csv, json)

Description	The ICARUS platform will allow the aviation data stakeholders (data providers and data consumers) to export the analytics results they have obtained in a machine-readable format. Such an export may be obtained as a downloadable file (e.g. in csv or json format) or through calls to the ICARUS platform APIs (in json format).
Methodology Phase	V. Data Analytics – V.1 Data Analysis & IV. Asset Exploration & Searching – IV.2 Asset Export
Related Scenarios	SCE-2.12, SCE-6.7

Prerequisites	PLATF_F_35
----------------------	------------

PLATF_F_42. Export of analytics reports as a downloadable file

Description	The ICARUS platform will allow the aviation data stakeholders (data providers and data consumers) to export the analytics reports they have created in their customized dashboards as downloadable files (in pdf format).
Methodology Phase	V. Data Analytics – V.1 Data Analysis & IV. Asset Exploration & Searching – IV.2 Asset Export
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_35

PLATF_F_43. Saving your projects / analysis for future reference

Description	The ICARUS platform will allow the aviation data stakeholders (data providers and data consumers) to save their projects and the analysis they have performed for future reference (e.g. to execute the same analytics tasks again in the future for the same data assets or for up-to-date data assets).
Methodology Phase	V. Data Analytics – V.1 Data Analysis & V.2 Data Visualization
Related Scenarios	SCE-2.12, SCE-2.13, SCE-6.7, SCE-6.8
Prerequisites	PLATF_F_35

PLATF_F_44. Execution of scheduled analytics

Description	The ICARUS platform will allow the aviation data stakeholders (data providers and data consumers) to define scheduled analytics tasks that may run automatically in certain intervals (after x days/hours) or in specific dates (e.g. every 1 st of the month, every Monday, etc.). Such scheduled analytics shall be executed either in an incremental manner or in full and will be dependent on the exact user settings and the plan that a stakeholder has acquired in ICARUS.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_35

PLATF_F_45. Navigation to analytics on data asset usage

Description	The ICARUS platform will provide a dedicated data analytics interface to the aviation data providers in order to trace their data assets and get insights into how other people are using their data assets.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	N.A.
Prerequisites	PLATF_F_1-2

PLATF_F_46. Deployment and customization of a secure experimentation space in ICARUS

Description	The ICARUS platform will allow the aviation data stakeholders (providers and consumers) to deploy and customize their own secure experimentation space (within the platform) in which they may run the analysis they wish without getting concerned for their business-critical data visibility. In such a space, their data assets confidentiality is ensured.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.12, SCE-6.7
Prerequisites	PLATF_F_35

PLATF_F_47. Erasure of a secure experimentation space

Description	The ICARUS platform will allow the aviation data stakeholders (providers and consumers) to erase their own secure experimentation space (within the platform) as soon as they no longer need it (e.g. when the analysis is complete and its results have been exported). The deletion of such a space encompasses deleting all traces of confidential data that have been uploaded.
Methodology Phase	V. Data Analytics – V.1 Data Analysis
Related Scenarios	N.A.
Prerequisites	PLATF_F_46

PLATF_F_48. Delivery of notifications regarding new data assets checked in, and/or existing data assets updated, related to own data assets or to analysis and visualisations performed

Description	The ICARUS platform will deliver notifications to the data consumers regarding any data-related event: (a) new data assets that have been checked in, are relevant to their stakeholder profile and are potentially linkable to their own assets, and / or (b) existing data assets which they have pinned, utilized in the past or agreed on a data licence with their respective data providers and that have been updated. The ICARUS platform will deliver notifications to the data providers regarding their data assets and may range from notifications for new data assets that have been checked in and are linkable to their own to notifications for new types of analytics performed on the data assets they own.
Methodology Phase	VI. Added Value Services – Notifications
Related Scenarios	SCE-5.9
Prerequisites	PLATF_F_1-2-24-45

PLATF_F_49. Delivery of notifications regarding updates and modifications in the terms of use (e.g. licences) of data assets exploited through the platform

Description	The ICARUS platform will deliver notifications to the data consumers regarding updates on the data assets they have used, as well as any potential change in the data licences on which they have agreed with the data providers (e.g. regarding the time period in which a data asset is available, the pricing, etc.).
Methodology Phase	VI. Added Value Services – Notifications
Related Scenarios	SCE-5.9
Prerequisites	PLATF_F_58-60

PLATF_F_50. Delivery of notifications regarding the successful or unsuccessful execution of scheduled analytics

Description	The ICARUS platform will deliver notifications to the aviation data stakeholders (data providers and data consumers) that concern status updates on the execution of the scheduled analytics tasks they have defined. Typically, such notifications identify the successful vs failed run of such analytics and may even identify the cause for the unsuccessful execution.
Methodology Phase	VI. Added Value Services – Notifications & V. Data Analytics – V.1 Data Analysis
Related Scenarios	N.A.
Prerequisites	PLATF_F_44

PLATF_F_51. Proposition of additional data assets for the enrichment of existing data assets and / or for analysis and visualisation

Description	The ICARUS platform will recommend additional data assets, which either they own or come from different data providers, to both the data providers and the data consumers at the different moments they use it, based on their metadata and a set of common “variables in their data model (in terms of semantics). Such recommendations intend to assist the data providers in enriching their existing assets and the data consumers in performing more intricate analysis.
Methodology Phase	VI. Added Value Services – Recommendations & II. Data Enrichment – II.1 Semantic Enrichment and Annotation & V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-1.7, SCE-2.9, SCE-2.11, SCE-6.4
Prerequisites	PLATF_F_13-21-35-38

PLATF_F_52. Proposition of machine learning algorithms for the quicker and / or more efficient extraction of insights according to the data assets at hand

Description	The ICARUS platform will recommend to the data consumers specific machine learning algorithms which they may execute depending on their relevance and applicability to the data assets at hand in order to help them obtain insights to the data essence in a quicker and more efficient manner.
Methodology Phase	VI. Added Value Services –Recommendations & V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.11, SCE-6.6
Prerequisites	PLATF_F_29-30-31

PLATF_F_53. Proposition of optimal visualisations for the quicker and / or more efficient extraction of insights according to the data assets at hand

Description	The ICARUS platform will recommend to the data consumers the optimal visualizations depending on their relevance to the analysis performed in order to help them get a better and more intuitive understanding of data assets at hand.
Methodology Phase	VI. Added Value Services – Recommendations & V. Data Analytics – V.2 Data Visualization
Related Scenarios	SCE-2.13, SCE-6.8

Prerequisites	PLATF_F_31-38
----------------------	---------------

PLATF_F_54. Proposition of recommendations based on analyses run by people with similar interests (personas-based recommendations)

Description	The ICARUS platform will recommend to the data consumers different types of analysis that have been performed by other aviation data stakeholders and which they may also run for their own data assets or for the data assets of their selection. Such recommendations are characterized as personas-based, but they eventually coalesce into the different stakeholders in the aviation data value chain (i.e. airport, airline, ground handlers, etc.).
Methodology Phase	VI. Added Value Services – Recommendations & V. Data Analytics – V.1 Data Analysis
Related Scenarios	SCE-2.11, SCE-6.6
Prerequisites	PLATF_F_35

PLATF_F_55. Automatic license compatibility check for data assets that build on other assets

Description	The ICARUS platform will automatically check and ensure the compatibility of the data licences for data assets in different cases that will be enabled, e.g. when data assets are about to be linked, when data assets are about to be part of a common simple or complex analytics task, when data assets are about to be visualized together, when the deriving data asset is about to become an asset on its own.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	SCE-6.2
Prerequisites	-

PLATF_F_56. Step-by-step guidance on how to define the appropriate license of a data asset

Description	The ICARUS platform will provide step-by-step guidance to data providers before the check-in process concludes in order to practically help them select the appropriate data license or elaborate on the specific terms that are to be applied in their data asset (especially if the predefined data licenses are not to be adopted as-is or with minor modifications).
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	SCE-6.10
Prerequisites	PLATF_F_04

PLATF_F_57. Negotiation of a data sharing agreement

Description	The ICARUS platform will provision for a secure negotiation mechanism to allow the involved parties (data asset provider and consumer) to reach a mutually beneficiary, bilateral data sharing agreement, while ensuring the non-repudiation of the underlying terms.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	SCE-6.4
Prerequisites	PLATF_F_04

PLATF_F_58. Automatic renewal of a data sharing agreement

Description	The ICARUS platform will allow data consumers to automatically renew an active data sharing agreement when it is explicitly permitted by the terms of the agreement. In this way, the data consumers ensure that they have uninterrupted access to the data asset which is critical when they have defined scheduled analytics tasks that leverage it.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	SCE-6.4
Prerequisites	PLATF_F_57

PLATF_F_59. Cancellation of a data sharing agreement

Description	The ICARUS platform will in principle allow data providers and data consumers to cancel a data sharing agreement yet only the agreement itself foresees whether such an option is possible and which are the exact cancellation terms (e.g. to what extent penalties should be charged to the data consumers if they decided they no longer need a data asset or to data provider for deciding to withdraw a data asset after a specific time period). In any case, the reasons for cancellation of a data sharing agreement need to be explicitly defined as they may affect the reputation score of the defaulting party.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	N.A.
Prerequisites	PLATF_F_57

PLATF_F_60. Approval of a change in the terms of a data sharing agreement

Description	The ICARUS platform will in principle allow data providers and data consumers to change certain terms of their data sharing agreement (e.g. duration, frequency of updates) as long as it does not violate the principles of the original agreement. For relatively minor changes, the involved parties need to approve the changes requested by one of them while for major changes in the content of the data sharing agreement, it might be necessary to enter another round of negotiations.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	N.A.
Prerequisites	PLATF_F_57

PLATF_F_61. Acceptance of terms of use of a public data asset and availability to download

Description	Since the public data assets are also accompanied by specific data licences, the ICARUS platform will ensure that the respective terms are acknowledged and accepted by the data consumers. Such an acceptance will occur prior to allowing the data consumers to utilize, link, analyze or download the respective data asset.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	SCE-6.4
Prerequisites	-

PLATF_F_62. Closure" of a data sharing agreement/contract, proceeding to the analysis or to download a specific data asset

Description	The ICARUS platform will enable the involved data stakeholders (data provider and data consumer) to close a data sharing agreement for private data assets and will ensure that its terms are duly respected. Although extracts of such private data may be available in advance to the data consumer, such an agreement is a prerequisite in order to be able to properly link them with other data assets, run the analytics tasks and download the data in full scale.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	SCE-6.4
Prerequisites	PLATF_F_57

PLATF_F_63. Assessment of reputation of data assets by confirmed "byuers" / "users"

Description	The ICARUS platform will enable the data consumers who have evidently “purchased” a data asset, concluded a data sharing agreement for its use or actually used a data asset for analytics (especially in the case of a public data asset), to provide its assessment online. Such assessments may practically consist of a reputation score for the data asset and / or a comment provided by the data consumer.
Methodology Phase	VI. Added Value Services – Asset Sharing
Related Scenarios	N.A.
Prerequisites	PLATF_F_62

PLATF_F_64. Registration of a new service in the ICARUS platform according to a specific metadata schema

Description	The ICARUS platform will enable the aviation data stakeholders, as well as data scientists who are working with aviation data, to register a new service (e.g. code for a new custom analytics algorithm) according to the ICARUS metadata schema. A detailed profiling of the service, including its general information, the values in certain performance KPIs defined by ICARUS, the training data asset and the test data asset, is required by the service provider in order to allow ICARUS to perform a proper assessment.
Methodology Phase	VII. Service Collection – VII.1 Service Check-in
Related Scenarios	SCE-4.1
Prerequisites	-

PLATF_F_65. Semi-automatic assessment of a new service

Description	The ICARUS platform with the help of its administration team will examine the request for inclusion of a new service and semi-automatically assess it through in-depth verification and validation activities, e.g. in order to ensure that it does not contain any malicious code or code that is intrusive to the seamless platform operation, that the performance KPIs are accurately calculated and that the service delivers its intended purpose.
--------------------	--

Methodology Phase	VII. Service Collection – VII.2 Testing and Assessment
Related Scenarios	SCE-4.2
Prerequisites	PLATF_F_64

PLATF_F_66. Strict moderation of new services to strengthen trust to the ICARUS platform

Description	The ICARUS platform will anticipate the strict moderation of new services by its administration team through appropriate online procedures to ensure its seamless, secure and uninterrupted operation.
Methodology Phase	VII. Service Collection – VII.3 Service Asset Review
Related Scenarios	SCE-4.3
Prerequisites	PLATF_F_64-65

4.2 Initial Features Value Assessment

In order to assess the added value that the feature list that has been extracted and defined in section 4.2 brings to the different stakeholders in the aviation data value chain, a two-step assessment will be followed: (a) Internally within the ICARUS consortium in order to gauge the feedback of the 4 demonstrators (AIA, PACE/TXT, ISI, CELLOCK) and of the core data provider (OAG), and (b) Externally with key stakeholders that are approached either face-to-face in the content of the ICARUS stakeholder engagement activities or online through a targeted survey. Since the MVP definition is considered as a live, continuously evolving learning process that cross-cuts the design-development-piloting activities, only the internal assessment will be reported in the context of this deliverable and shall be complemented with the full, generalized assessment in the 2nd release of the WP1 activities.

Upon elaborating on the list of features that may constitute the ICARUS MVP, the relevant ICARUS stakeholders were requested to describe and rate online in a qualitative manner the business value of each feature, using a scale between 1 (Little or no impact on internal operations, Little or no competitive advantage for ICARUS in the aviation industry) and 5 (Extreme impact on internal operations, Critical competitive advantage for ICARUS in the aviation industry) points and taking into account their response to questions such as: How important is this feature for your internal operations? How crucial is this feature for the broader aviation industry? The scale that has been adopted builds on the proposition of Lant² regarding the assessment of the business value and adapts it to the broader context of features (rather than concrete user stories that will follow at later stages of the project).

Figure 4-2 presents the aggregated assessment of the ICARUS demonstrators for the own business value they identify in the different features. Since the ISI demonstrator belongs to the 3rd tier of the data value chain, the aggregated assessment for the other 3 demonstrators which are classified in the 1st and 2nd tiers is also calculated. As indicated in the figure, there are many features that are considered as important or very important (scoring above 3) while very few features were considered as somewhat important (scoring between 1.5 and 2).

² Michael Lant (2010). How to Easily Prioritize Your Agile Stories. Available at: <http://michaellant.com/2010/05/21/how-to-easily-prioritize-your-agile-stories/>

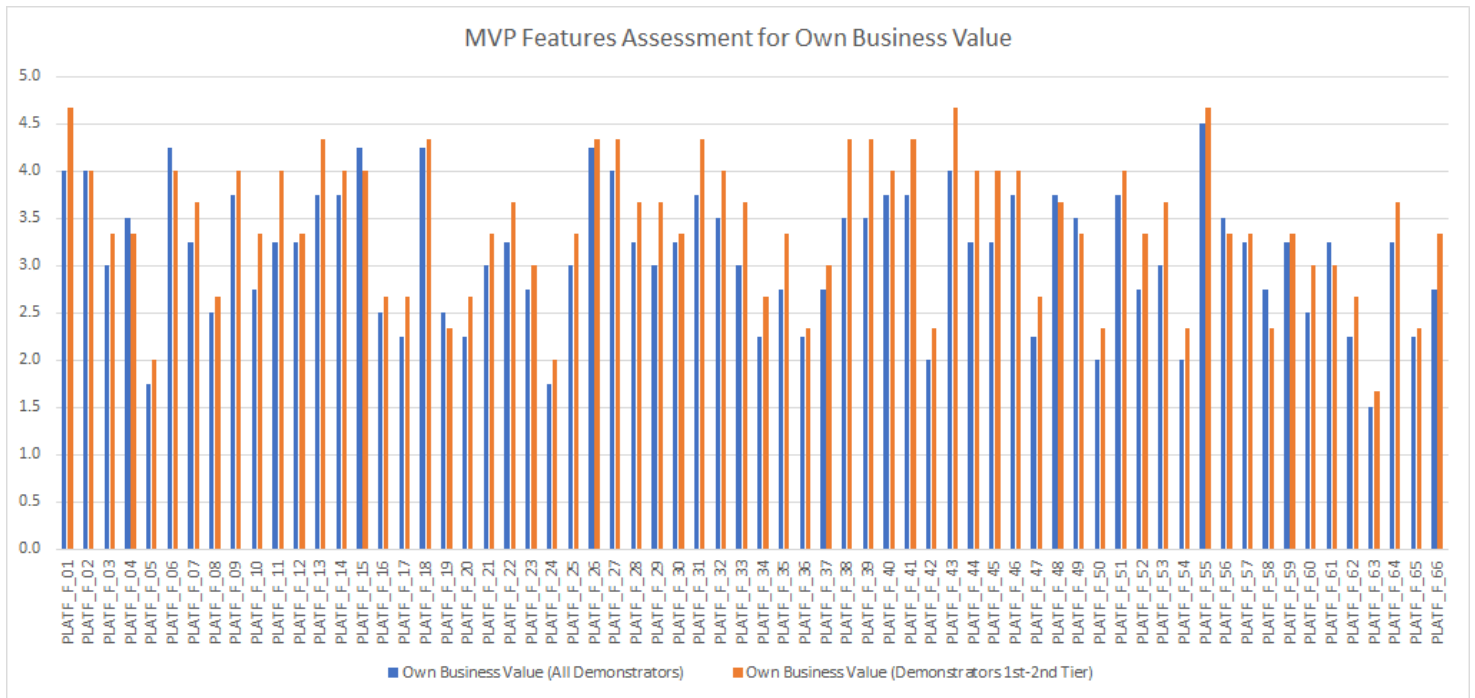


Figure 4-2: ICARUS MVP Features Assessment by the 4 Demonstrators for Own Business Value

With regard to the more generic assessment for the aviation industry in which 3 demonstrators (AIA, PACE/TXT, CELLOCK) and the core data provider (OAG) contributed, it can be easily noticed in Figure 4-3 that there are many features that distinguished (receiving a score above 3.5). Such an “industry” assessment is broadly consistent with the “own business value” dimension as depicted in Figure 4-4. The overall assessment of above 2 for all features in figure 4-3, though, suggests that the feature extraction (but also its preparatory steps, namely the methodology and the high-level scenarios elaboration) is in line with the industry needs, all features are relevant and bring added value to the aviation industry.

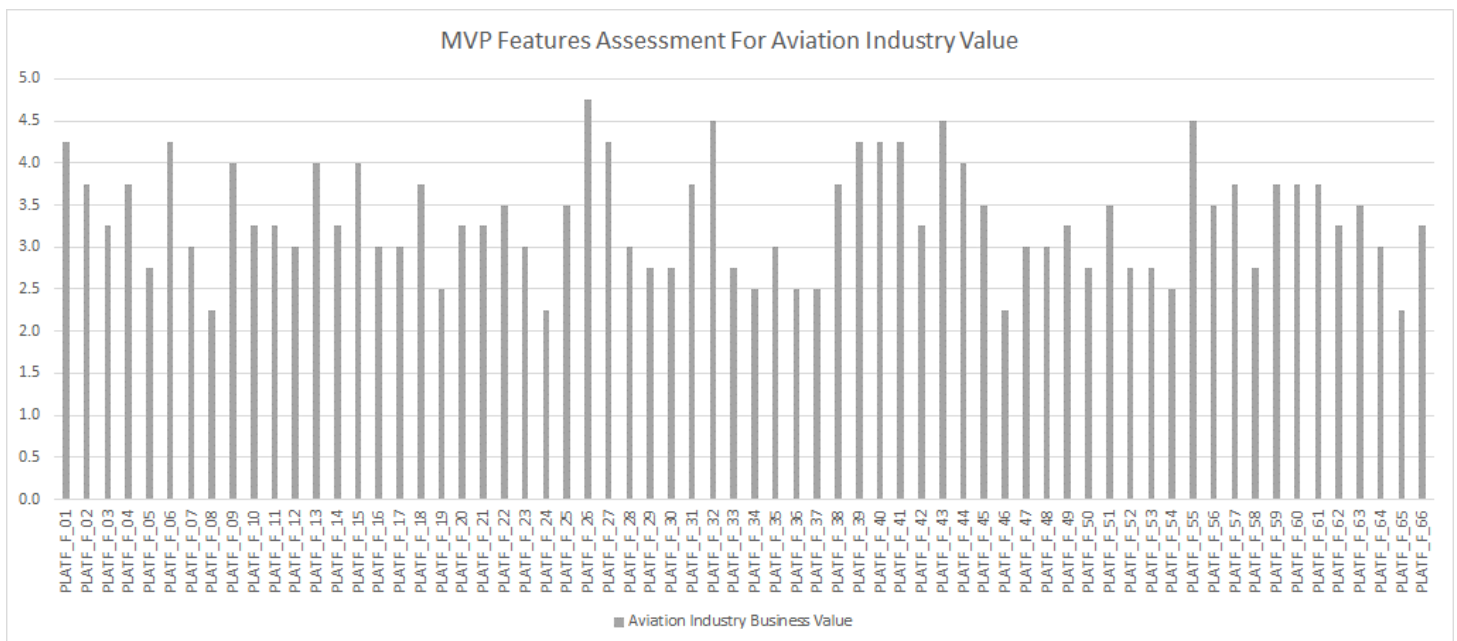


Figure 4-3: ICARUS MVP Features Assessment by 3 Demonstrators and OAG for Aviation Industry Business Value

In the ranking correlation in Figure 4.4, the only features in which there are notable differences (above 1.5 points) between the demonstrators' own value and the aviation industry business value are: PLATF_F_46 (regarding the secure experimentation space deployment that seems more important for aviation than the demonstrators) and PLATF_F_63 (for service check-in that is appreciated more by the demonstrators), whose added value will be further investigated in the 2nd iteration.

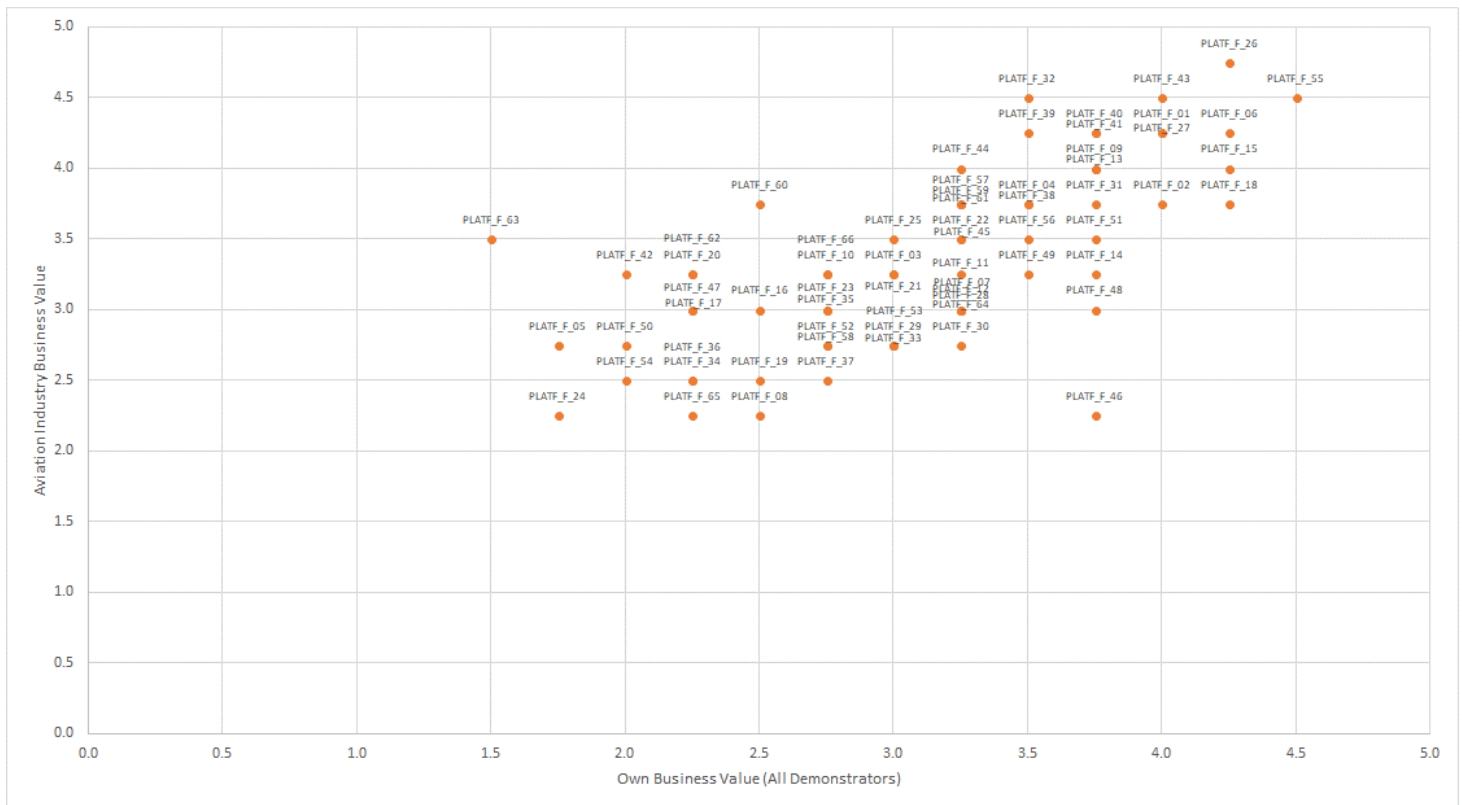


Figure 4-4: ICARUS MVP Features Assessment: Correlation between the Demonstrators' Own Business Value and the Aviation Industry Business Value

4.3 Preliminary MVP Consolidation

Taking into account the preliminary assessment reported in section 4.2, the ICARUS MVP consists of a set of features that have been selected depending on their combined rank (to be above 3) for the ICARUS demonstrators and the aviation industry, as depicted in the following table.

Overall, the ICARUS MVP is addressed to the aviation data value chain stakeholders and, in particular, to both data analysts and business users of the stakeholders in the 1st-2nd-3rd data tiers.

Table 4-1: Preliminary ICARUS MVP on M6

ID	Title	Assessment Comments
PLATF_F_01	Retrieval of data directly from an aviation stakeholder's back-end system	Reduction of the time needed to access potentially useful data assets, potentially allows for real time data sharing.
PLATF_F_02	Uploading of data assets as files extracted by the aviation stakeholder's back-end system	Standard way to enable data-driven collaboration between aviation stakeholders.

ID	Title	Assessment Comments
PLATF_F_06	(Semi-)Automatic quality check of the data and assessment of quality level	Quickly assess the level of maturity of the data sets to be integrated for analysis. Enable high-quality results.
PLATF_F_09	(Semi-)Automatic on-the-fly anonymization in ICARUS	Mitigation of compliance issues for sensitive data to be shared. Transparency regarding how anonymization happens to build trust.
PLATF_F_11	(Semi-)Automatic extraction of and navigation within the Data Model of a Data Asset	Better understanding of data.
PLATF_F_13	(Semi-) Automatic Transformation / Mapping of data assets and extracted concepts to the ICARUS common schema	More efficient data linking. Verify the consistency of a group of related data assets.
PLATF_F_14	Easily applicable data manipulation / transformation methods	Less effort required by the data consumer to bring the data to the required level for analysis.
PLATF_F_15	Easily applicable data cleaning methods	Increased data quality with less effort required on behalf of the data provider.
PLATF_F_18	Searchability and identification of related additional data assets	Complement existing data assets with additional ones that the data provider that was not aware they existed, producing even more analytics results and reaching conclusions that were impossible otherwise.
PLATF_F_20	Indication of the linkable denominators, upon which linking of the data assets can be performed	Quick and effortlessly identify related information.
PLATF_F_21	(Semi-)Automatic data asset linking	Visibility of the integration process is necessary for transparency and control. Original data assets should not be altered.
PLATF_F_22	Definition of simple and advanced "information" queries	Speed up data exploration but the query templates need to be thoughtfully constructed to be easily used and to quickly retrieve appropriate responses.
PLATF_F_25	Filtering of data assets based on different criteria	Useful for navigating and exploring relevant data assets according to different criteria.
PLATF_F_26	Access and inspection of data assets "extracts" depending on their license	Preview and validate the usefulness of data assets prior to purchasing them / creating a data license.
PLATF_F_27	Transformation of a data asset to other supported data formats and export	Easier to exploit data in the stakeholders' back-end systems. However, it leads to a significant risk: data that are exported from the platform can no longer be controlled according to the data sharing agreements and the related rules embedded in the platform.
PLATF_F_31	Automatic check whether the data asset is appropriate for a specific algorithm	Prevent obtaining wrong/unreasonable output results.
PLATF_F_32	Automatic check for data licences compatibility to run under a specific algorithm	Rightful use of data assets. Avoid potential licence infringements.
PLATF_F_35	Execution of an analytics task / algorithm according to specific preferences and settings for computation resources	Crucial to address specific analysis requirements and distribute resources in different analytics with different priority.
PLATF_F_38	Visualization of the analytics results to gain insights on the data and / or comparison how the same results are visualized in different diagrams	Easier understanding of results by business users (not competent analytics users). Quicker evaluation of results by data analysts.
PLATF_F_39	Definition of customized dashboards by selecting which visualizations should appear	Efficiently build different dashboards according to the needs each stakeholder has.
PLATF_F_40	Definition of an end-to-end workflow / recipe	Instant gratification from the platform operation. Reduced time to use (from familiarizing with the platform to obtaining actionable analytics results).
PLATF_F_41	Export of analytics results in machine-readable format	Easier interfacing of results to other stakeholders' back-end platforms.

ID	Title	Assessment Comments
PLATF_F_43	Saving your projects / analysis for future reference	Standard feature to avoid work repetition and save time for recurring analysis.
PLATF_F_44	Execution of scheduled analytics	Very important for data that are refreshed in defined time periods and for the periodic exchange of data and analytics results between industry stakeholders.
PLATF_F_45	Navigation to analytics on data asset usage	Auditing of data access and usage to confirm compliance with the data sharing agreement. Better understanding of the usage of data. More customer-oriented perspective of the data.
PLATF_F_48	Delivery of notifications regarding new data assets checked in, and/or existing data assets updated, related to own data assets or to analysis and visualisations performed	Instant action to adapt to new data assets or update existing ones. Risk of being ignored if there are too many notifications or if they are not relevant.
PLATF_F_49	Delivery of notifications regarding updates and modifications in the terms of use (e.g. licences) of data assets exploited through the platform	Keep track how licences are impacted, especially for data assets that are already exploited in order to check compliance and possibly take suitable actions (e.g. evaluate whether a renewal of the agreement is required or not).
PLATF_F_51	Proposition of additional data assets for the enrichment of existing data assets and / or for analysis and visualisation	Expedite the use of data assets that stakeholders would not normally employ in the analysis. Gain insights that would not be possible to reach otherwise, creating added value creation for the whole industry.
PLATF_F_55	Automatic license compatibility check for data assets that build on other assets	Safeguard the interests of the data providers. Highlight potential issues before acquiring input data assets. Build trust on the platform.
PLATF_F_56	Step-by-step guidance on how to define the appropriate license of a data asset	Crucial to bring together the right terms for the sharing agreement without ambiguity. Easier and more straightforward process.
PLATF_F_57	Negotiation of a data sharing agreement	Assistance to even non-competent personnel to create and close an agreement, especially if mediation for conflicts is provided.
PLATF_F_60	Approval of a change in the terms of a data sharing agreement	Easier change management. Consent for changing certain terms in the bilateral agreements ensures the data consumers' awareness and appropriate action.
PLATF_F_61	Acceptance of terms of use of a public data asset and availability to download	Ensure terms of use for public data assets (not only for private data assets for which licences are needed) are appropriately communicated to data consumers before use in the platform and / or download.

However, it needs to be noted that such features will be detailed into user stories and concrete requirements in WP3 and are subject to updates and prioritization depending on: (a) the generalized assessment provided by additional stakeholders in the aviation industry (beyond the ICARUS consortium), (b) the technical feasibility in terms of expected implementation effort, and (c) their dependency on other features which has not been considered in Table 4-1.

5 Conclusion

The present deliverable (D1.2) documents the results of T1.4 that had three main goals; to define the ICARUS methodology, high-level usage scenarios and MVP (Minimum Viable Product) that will guide the next implementation steps of the project. To derive the outcomes of T1.4, a clear and easily comprehensive approach was followed: 1) definition of ICARUS methodology, 2) scenarios definition, 3) features selection, 4) voting of feature importance and 5) MVP definition.

At first, ICARUS methodology was defined based on the key findings of D1.1 “Domain Landscape Review and Data Value Chain Definition”. In particular, the methodology was divided in 7 phases, with each phase having its own steps and aspects. These phases are: Phase I - Data Collection, Phase II - Data Enrichment, Phase III - Asset Storage, Phase IV - Asset Exploration and Extraction, Phase V - Data Analytics, Phase VI - Added Value Services, Phase VII - Service Collection.

The next step was to define the high-level usage scenarios of ICARUS based on the ICARUS methodology. In particular, the consortium defined several high-level usage scenarios. These scenarios took into account the key findings from D1.1. In this deliverable, six high-level scenarios (general workflow diagrams) were defined in detail, as representative scenarios of all core differentiated ICARUS stakeholders. Moreover, three different examples (technical sub-diagrams) were created for each scenario, referring to different potential stakeholders that may use ICARUS platform in different ways, depending on their needs and objectives.

The final step was to define the ICARUS MVP based on the methodology, the high-level scenarios and the demonstrators’ requirements as reflected in the brainstorming sessions of the plenary meeting in Nicosia in May 2018. In particular, 66 features were initially extracted and consist the basis for the ICARUS MVP definition. Such features were grouped based on the phases of the methodology and were assessed from the ICARUS demonstrators (AIA, PACE/TXT, ISI, CELLOCK) and the core data provider (OAG) with regard to their “business value and impact on business operations” and “business value and impact on the broader aviation industry”. The preliminary MVP has been formulated and will be further elaborated (e.g. in user stories and requirements in WP3 and through external validation activities in WP1), yet it needs to be underlined that the MVP in ICARUS represents the mentality of work to ensure that the ICARUS platform is appropriately validated by its end users and delivers the maximum added value, with the lowest possible risk, by the end of the project.

In the forthcoming steps, the outcomes of D1.2 will feed the use cases, architecture and specification tasks in WP2 and WP3. More precisely, this deliverable will feed the ICARUS deliverables D2.1 “Data Management and Value Enrichment Methods”, D2.2 “Intuitive Analytics Algorithms and Data Policy Framework”, D3.1 “ICARUS Architecture, APIs Specifications and Technical and User Requirements” and D7.1 “Initial Project Exploitation Plan–v1”. Finally, the task of this deliverable will be constantly monitored and the updates to the work and results will be presented in deliverable D1.3 “Updated ICARUS Methodology and MVP”, as it remains active until M15 of the project.

6 References

-
- [1] BDV SRIA, “European Big Data Value.” [Online]. Available: http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf. [Accessed: 29-Mar-2018].
 - [2] “GDPR Wiki.” [Online]. Available: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation. [Accessed: 15-Jun-2018].
 - [3] W3C, “Web API Design Cookbook.” [Online]. Available: <https://www.w3.org/TR/api-design/>. [Accessed: 06-Jun-2018].
 - [4] Apigee, “RESTful API design.” [Online]. Available: <https://apigee.com/about/tags/restful-api-design>. [Accessed: 06-Jun-2018].
 - [5] “GDPR.” [Online]. Available: <https://www.eugdpr.org/>. [Accessed: 15-May-2018].
 - [6] “China Cyber Security Law.” [Online]. Available: https://en.wikipedia.org/wiki/China_Internet_Security_Law.
 - [7] “Protecting the Privacy of Customers of Broadband and other Telecommunications Services.” [Online]. Available: <https://www.congress.gov/bill/115th-congress/senate-joint-resolution/34>. [Accessed: 15-May-2018].
 - [8] R. Matsunaga, I. Ricarte, T. Basso, and R. Moraes, “Towards an Ontology-Based Definition of Data Anonymization Policy for Cloud Computing and Big Data,” in *Dependable Systems and Networks Workshop (DSN-W), 2017 47th Annual IEEE/IFIP International Conference on*, 2017, pp. 75–82.
 - [9] J. Shao and H. Ong, “Semantic attack on anonymised transactions,” in *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIII*, Springer, 2016, pp. 75–99.
 - [10] P. Goswami and S. Madan, “Privacy preserving data publishing and data anonymization approaches: A review,” in *Computing, Communication and Automation (ICCCA), 2017 International Conference on*, 2017, pp. 139–142.
 - [11] A. Maydanchik, *Data quality assessment*. Technics publications, 2007.
 - [12] L. Cai and Y. Zhu, “The challenges of data quality and data quality assessment in the big data era,” *Data Sci. J.*, vol. 14, 2015.
 - [13] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
 - [14] “Data cleansing.” [Online]. Available: https://en.wikipedia.org/wiki/Data_cleansing.
 - [15] H. Müller and J.-C. Freytag, *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.
 - [16] “IATA’s Airline Industry Data Model.” [Online]. Available: <http://www.iata.org/whatwedo/passenger/Pages/industry-data-model.aspx>. [Accessed: 10-Jun-2018].
 - [17] “NASA’s Advanced Air Traffic Management Ontology.” [Online]. Available: <https://ti.arc.nasa.gov/news/atm-ontology/>. [Accessed: 10-Jun-2018].
 - [18] J. Peckham and F. Maryanski, “Semantic data models,” *ACM Comput. Surv.*, vol. 20, no. 3, pp. 153–189, 1988.

- [19] A. Ferraram, A. Nikolov, and F. Scharffe, "Data linking for the semantic web," *Semant. Web Ontol. Knowl. Base Enabled Tools, Serv. Appl.*, vol. 169, p. 326, 2013.
- [20] U. M. Fayyad, A. Wierse, and G. G. Grinstein, *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2002.
- [21] "Dimensionality Reduction." [Online]. Available: <https://github.com/niranjv/ml-notes/wiki/Dimensionality-Reduction>. [Accessed: 10-Jun-2018].
- [22] G. Zyskind, O. Nathan, and others, "Decentralizing privacy: Using blockchain to protect personal data," in *Security and Privacy Workshops (SPW), 2015 IEEE*, 2015, pp. 180–184.
- [23] M. Swan, *Blockchain: Blueprint for a new economy*. "O'Reilly Media, Inc.," 2015.
- [24] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.
- [25] D. W. Oard, J. Kim, and others, "Implicit feedback for recommender systems," in *Proceedings of the AAAI workshop on recommender systems*, 1998, vol. 83.
- [26] G. A. Sielis, A. Tzanavari, and G. A. Papadopoulos, "Recommender systems review of types, techniques, and applications," in *Encyclopedia of Information Science and Technology, Third Edition*, IGI Global, 2015, pp. 7260–7270.
- [27] "Recommender System." [Online]. Available: https://en.wikipedia.org/wiki/Recommender_system.
- [28] "Creative Commons Licenses." [Online]. Available: <https://creativecommons.org/share-your-work/licensing-types-examples/>. [Accessed: 10-Jun-2018].
- [29] "Creative Commons Licenses Wiki." [Online]. Available: https://en.wikipedia.org/wiki/Creative_Commons_license. [Accessed: 10-Jun-2018].