Ref. Ares(2019)5115234 - 06/08/2019

H2020 - INDUSTRIAL LEADERSHIP - Information and Communication Technologies (ICT)

ICT-14-2016-2017: Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation

ICARUS:

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

D2.3 – Updated ICARUS Data Management, Analytics and Data Policy Methods

Workpackage:	WP2 – ICARUS Big Data Framework Consolidation		
Authors:	Suite5, UBITECH, SILO, ENG, UCY, OAG, AIA, PACE, ISI, CELLOCK		
Status:	Final	**Classification:**	Public
Date:	06/08/2019	**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.

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 ubiquitous solutions	UBITECH (UBITECH)	Greece
ENGINEERING	ENGINEERING - INGEGNERIA INFORMATICA SPA (ENG)	Italy
PACE a TXT company	PACE Aerospace Engineering and Information Technology GmbH (PACE)	Germany
Suite5 We Deliver Intelligence	SUITE5 DATA INTELLIGENCE SOLUTIONS LIMITED (SUITE5)	Cyprus
University of Cyprus	UNIVERSITY OF CYPRUS (UCY)	Cyprus
CINECA	CINECA CONSORZIO INTERUNIVERSITARIO (CINECA)	Italy
OAG connecting the world of travel	OAG Aviation Worldwide LTD (OAG)	United Kingdom
SingularLogic	SingularLOGIC S.A. (SILO)	Greece
	ISTITUTO PER L'INTERSCAMBIO SCIENTIFICO (ISI)	Italy
CELLOCK	CELLOCK LTD (CELLOCK)	Cyprus
ATHENS INTERNATIONAL AIRPORT ELEFTHERIOS VENIZELOS	ATHENS INTERNATIONAL AIRPORT S.A (AIA)	Greece
TXT E-SOLUTIONS	TXT e-solutions SpA (TXT) – 3rd party of PACE	Italy

Document History

Version	Date	Author (Partner)	Remarks
0.10	10/05/2019	Evmorfia Biliri, Fenareti Lampathaki (Suite5)	Initial Table of Contents
0.20	11/06/2019	Evmorfia Biliri, Fenareti Lampathaki (Suite5)	Updates in Section 3.1.1, and templates for sections 3.1.3-3.1.6
0.21	20/06/2019	Corrado Gioannini (ISI)	Contribution to Section 3.1.5
0.22	24/06/2019	Minas Pertselakis (Suite5), Nikos Papagiannopoulos (AIA)	Updates in Section 3.1.3
0.23	25/06/2019	Susanna Bonura (ENG)	Contribution to Section 3.1.2
0.24	28/06/2019	Henry Puls (PACE), Stamatis Pitsios (UBITECH)	Contribution to Section 3.1.4
0.25	28/06/2019	Dimosthenis Stefanidis (UCY), Yiannis Diellas (CELLOCK)	Contribution to Section 3.1.6
0.30	05/07/2019	Dimitrios Miltiadou, Konstantinos Perakis (UBITECH)	Contribution to Sections 2.2.1, 2.3.1, 2.3.2, 4.1
0.40	12/07/2019	Evmorfia Biliri, Fenareti Lampathaki (Suite5)	Contribution to Sections 2.1, 2.2.2, 2.2.3, 4.1, 4.2 taking into account the feedback by OAG in section 2.2.3
0.50	15/07/2019	Marios Zacharias (SILO)	Contribution to Sections 2.3.3, 3.2, 4.2
0.60	17/07/2019	Evmorfia Biliri, Fenareti Lampathaki (Suite5)	Contribution to Sections 4.1, 4.2 and revisions in Section 3.2 taking into account the feedback by OAG
0.70	25/07/2019	Evmorfia Biliri, Fenareti Lampathaki (Suite5)	Contribution to Sections 1, 5, Annex II
0.80	29/07/2019	Fenareti Lampathaki (Suite5)	Updated full draft circulated for internal review
0.90	02/08/2019	Evmorfia Biliri, Fenareti Lampathaki (Suite5)	Updated version addressing feedback received during the plenary meeting and the internal review process
1.00	06/08/2019	Fenareti Lampathaki (Suite5), Dimitrios Alexandrou (UBITECH)	Final version for submission to the EC

Executive Summary

The present deliverable D2.3 "Updated ICARUS Data Management, Analytics and Data Policy Methods" aims at refining and finalizing the bundle of data management and data value enrichment methods that have been defined in WP2 and in particular in D2.1 and D2.2 to reflect the latest perspectives gained through the development activities and the preliminary demonstrators definition phase. Such methods have effectively laid the theoretical foundations of the different data bundles of the ICARUS platform taking into account the aviation industry needs, requirements and peculiarities while building on an extensive state-of-the art analysis and creating a compelling case for the aviation data value chain, vis-a-vis certain key considerations and open challenges.

In brief, the ICARUS data management and data value enrichment methods span over the following axes:

- Axis I: Data Collection, that considers the upstream, downstream, indirect and open data assets' collection from the supply-driven perspective of the data providers. The de facto data collection approach in ICARUS concerns files upload / exchange at the moment while the applicable processes for data check in and data update are elaborated and the supported data profiles in terms of formats and standards are put into context.
- Axis III: Data Curation, involving the data cleaning, data provenance and data mapping and linking perspectives to be applied in ICARUS. In particular:
  - The *ICARUS data cleaning process* aims at increasing the data quality by detecting and correcting (or removing) corrupt or inaccurate records, through its 5-step process.
  - The *ICARUS data provenance process* practically captures and manages trustworthy data asset trails that shall effectively track the lineage and the derivation of the data assets that have been checked in in ICARUS in a coarse-grained, light-weight manner at dataset level, considering the agent, artefact, process and timing perspectives. The ICARUS metadata schema has been also updated featuring core metadata, semantic metadata, distribution metadata, sharing metadata and preservation metadata, taking into account the ICARUS platform implementation feedback.
  - The 8 phases of the *ICARUS data mapping and linking approach* ensure effective data integration at data check-in time and at data query time and concur in creating compatible data assets at syntactic and semantic level based on the ICARUS common aviation data model that has been constructed taking into account the ICARUS ontology, 4 aviation data standards and 1 generic-purpose data standard. The emphasis laid by ICARUS on the data model lifecycle and in particular on its evolution needs to be also highlighted.
- Axis III: Data Safeguarding that sets different layers for data security and privacy assurance viewed under the perspectives of: (a) *attribute-based access control policies* that formally dictate the circumstances under which access requests to data assets should be granted, and are easily interpretable into policy enforcement rules; (b) *end-to-end symmetric key*

encryption for data assets (before they are uploaded in the ICARUS platform) and secure tunnels for direct key sharing to authorized data consumers with active data contracts, (c) *multiple data anonymization methods* and guidelines for data providers to achieve the right balance in the "privacy vs utility" trade-off.

- Axis VI: Data Analytics considering 31 algorithms classified under *Basic Analytics*, *Machine Learning*, *Deep Learning* and *Visual Analytics*. In ICARUS, *data analytics and visualization* follows a 6-step approach that is designed taking into account the key steps that are typical to any data analytics approach. The current data analytics practices in the ICARUS demonstrators are extensively discussed in order to understand the baseline in the different aviation stakeholders that ICARUS will need to overcome. Taking into consideration the demonstrator scenarios defined in WP5, the Data Analytics that are considered in ICARUS per demonstrator are investigated in detail, proposing specific algorithm types/ families and algorithms, as well as potential alternatives, per identified problem in each scenario, defining the purpose, the inputs and expected output and discussing the anticipated limitations / caveats / considerations. In total, over 25 algorithms will be tested with the most suitable input data and optimized for the different input features in the demonstrators which will contribute their domain knowledge to improve the results and the data analytics scope.
- Axis V: Data Sharing that initially presents the ICARUS positioning in respect to the 12 dimensions along which data marketplaces can be examined. Departing from the data-focused perspective, the *ICARUS Data Sharing Model* formalises all data attributes and qualities that affect, or are in any way relevant to, the ways in which data assets can be shared / traded, while taking into account 6 key decisions for effectively driving the ICARUS data sharing advancements. The *ICARUS Blockchain-enabled Data Policy and Assets Brokerage Framework* also elaborates on an advanced workflow that captures the complex provider-consumer interactions to demonstrate how ICARUS envisions to enable the creation of structured, machine-processable data contracts for the aviation industry, whilst maintaining the data owner in control of the provided data.

In D2.3, in total, 37 key considerations (that represent specific challenges to be investigated in ICARUS) were described and the current ICARUS positioning and perspectives were extensively discussed.

In conclusion, this deliverable reports the final outcomes of the ICARUS activities related to Tasks T2.1 "Data Collection, Provenance and Safeguarding Methods", T2.2 "Data Curation, Harmonization and Linking Frameworks", T2.3 "Deep Learning and Prescriptive Analytics Algorithms" and T2.4 "Data Policy and Assets Brokerage Frameworks". Although WP2 has ended, the data management and value enrichment methods will continue to progress through their application in the ICARUS platform and incorporate experiences and feedback gathered from the aviation data value chain stakeholders.

Table of Contents

1 Introduction 9
1.1 Purpose 9
1.2 Methodological Approach 10
1.3 Relation to other ICARUS Results 11
1.4 Structure of the Document 12
2 Data Management Methods 13
2.1 ICARUS Data Collection Methods 13
2.2 ICARUS Data Curation Methods 16
2.2.1 Data Cleaning 16
2.2.2 Data Provenance 19
2.2.3 Data Mapping and Linking 21
2.3 ICARUS Data Safeguarding Methods 26
2.3.1 Data Access Control 26
2.3.2 Data Encryption 31
2.3.3 Data Anonymization 33
3 Data Value Enrichment Methods 36
3.1 ICARUS Data Analytics Methods 36
3.1.1 Introduction 36
3.1.2 Supported Analytics Workflows, Algorithms and Visualizations 38
3.1.3 Perspectives for Demonstrator 1: AIA 40
3.1.4 Perspectives for Demonstrator 2: PACE 43
3.1.5 Perspectives for Demonstrator 3: ISI 48
3.1.6 Perspectives for Demonstrator 4: CELLOCK 50
3.2 ICARUS Data Policy and Assets Brokerage Framework 55
3.2.1 Aviation Marketplace Insights and ICARUS positioning 56
3.2.2 Data Sharing Model 63
3.2.3 Blockchain-enabled Asset Sharing Workflows 68
4 ICARUS Considerations and Positioning 74
4.1 Data Management Considerations and Open Questions 74
4.2 Data Value Enrichment Considerations and Open Questions 79
5 Conclusions & Next Steps 85
Annex I: References 86

Annex II: Updated ICARUS Metadata Schema 88

List of Figures

Figure 1-1: Approach bringing together the D2.1 and D2.2 Approaches 11
Figure 1-2: Relation to other ICARUS Work Packages 11
Figure 2-1: ICARUS Data Collection Approach 14
Figure 2-2: ICARUS Data Provenance Schema 20
Figure 2-3: ICARUS Data Mapping and Linking Approach...26
Figure 2-4: XACML Syntax which is applied in ICARUS (Ferraiolo et al, 2016)...................................27
Figure 2-5: ICARUS Data Access Control Basic Workflows...30
Figure 2-6: ICARUS Data Encryption-Decryption Workflows...32
Figure 3-1: Analytics algorithms workflow design process...38
Figure 3-2: CELLOCK’s BoB: bond loading page & sales page 50
Figure 3-3: CELLOCK’s BoB: sales visualization pages & discrepancies visualization graphs...............50
Figure 3-4: ICARUS Asset Sharing Model - High-Level View 65
Figure 3-5: Core Asset Trading Workflow 72
Figure II-1: ICARUS Metadata Schema Overview..89

List of Tables

Table 2-1: ICARUS Data Access Control Attributes 28
Table 3-1: ICARUS Data Analytics methods and algorithms 36
Table 3-2: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.1 40
Table 3-3: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.2 41
Table 3-4: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.3 42
Table 3-5: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.4 42
Table 3-6: ICARUS Data Analytics for the PACE Demonstrator Scenario 1.......................................45
Table 3-7: Table 3 2: ICARUS Data Analytics for the PACE Demonstrator Scenario 2 46
Table 3-8: ICARUS Data Analytics for the ISI Demonstrator Scenario 49

Table 3-9: ICARUS Data Analytics for the CELLOCK Demonstrator Scenario 1 51

Table 3-10: ICARUS Data Analytics for the CELLOCK Demonstrator Scenario 2 53

Table 3-11: ICARUS Smart Contract relevant to the data sharing model 67

Table 3-12: ICARUS Policies included in a valid contract 67

Table 4-1: Data Management Considerations 74

Table 4-2: Data Value Enrichment Considerations 80

Table 0-1: ICARUS Metadata Schema Details 90

1 Introduction

1.1 Purpose

ICARUS aspires to address critical data linking, analytics and sharing challenges that the aviation data value chain faces, and eliminate barriers hindering the adoption of Big Data in the aviation industry. In order to overcome data fragmentation and promote data sharing, concrete methods need to be put in place to handle all data management and analysis steps, from data collection to curation and safeguarding and subsequently to data analysis and well-defined brokerage schemes. State-of-the-art methods on big data management and powerful emerging technologies constitute the foundations on which the ICARUS data management and data value enrichment methods are built.

In this context, the ICARUS Deliverable D2.3 "Updated ICARUS Data Management, Analytics and Data Policy Methods" concludes the activities performed in WP2 "ICARUS Big Data Framework Consolidation" and provides the final definition of the theoretical foundations, from various perspectives, on which the data analytics and sharing methods and services provided by the ICARUS platform will be built.

In essence, D2.3 reports the final outcomes of Tasks T2.1 "Data Collection, Provenance and Safeguarding Methods", T2.2 "Data Curation, Harmonisation and Linking Frameworks", T2.3 "Deep Learning and Prescriptive Analytics Algorithms" and T2.4 "Data Policy and Assets Brokerage Frameworks". According to the ICARUS Description of Action, the main objective dictated for D2.3 is to provide an update of the data handling methods, the core data analytics and the final data policy and brokerage framework, based on feedback received during the initial development and piloting phases. Therefore, D2.3 effectively inherits all objectives defined for D2.1 and D2.2 which are as follows: (a) prescribe methods for the collection and safeguarding data both in terms of provenance, storage as well as secure information exchange, (b) describe the appropriate patterns for harmonising and processing the data that will be used in the ICARUS platform, (c) define the semantics, the data handling algorithms and the overall logic that will combine data from various sources and deliver value for the engaged stakeholders, (d) study semantic vocabularies and design the ICARUS metadata repository, (e) suggest algorithms for knowledge extraction, business intelligence and usage analytics deriving from big cross-sectorial data, using advanced deep learning and prescriptive analytics, and (f) elaborate on the Data policy and Business Brokerage methods. These objectives have been appropriately investigated in D2.1 and D2.2, which have documented the initial ICARUS activities towards achieving the described targets and have also enriched the aforementioned objectives with more concrete challenges that need to be addressed.

The current deliverable builds upon the work presented in the first two WP2 deliverables in order to design and describe the final ICARUS Big Data Consolidation Framework. In detail, the scope of this deliverable is:

- To investigate whether the insights gained through the state-of-play review spanning all data management, analysis, policy and brokerage aspects (presented in D2.1 and D2.2) remain relevant, and update and extend them as needed.
- To collect feedback from the initial application of the defined methods during the development of the ICARUS beta platform and leverage it to refine the proposed approach.
- To extract additional requirements and needs for data collection, curation, safeguarding, analytics and sharing from the latest definition and scoping of the ICARUS demonstrators as reflected in their demonstrator scenarios (in D5.2).
- To report on the updates for the applicable data analysis algorithms in the ICARUS demonstrators.
- To refine the definition of the ICARUS data management methods, spanning collection, curation and safeguarding.
- To revisit the definition of the ICARUS asset sharing model which extends the previously defined data sharing model.
- To finalize the ICARUS data license and assets brokerage framework and present the project's positioning on how novel flexible data-enabled value chains will be realised in the aviation industry.
- To provide guidelines for the subsequent steps of the ICARUS platform design and implementation in the form of clearly defined data and metadata models, detailed methods to be applied and workflows to be enabled for the interactions among the ICARUS stakeholders and the platform.

1.2 Methodological Approach

As explained, D2.3 concludes the WP2 activities and therefore constitutes a report on all the work performed in the scope of the defined objectives of the work package. However, D2.3 is by definition expected to provide updates on the outcomes of the preceding WP2 deliverables and as such, the methodological approach that was followed here was, to a large extent, a repetition of the steps foreseen by the methodological approach presented and used in D2.1 and D2.2.

Leveraging the feedback gained from WP3 and WP4, the updates from WP1 and the insights from the initial demonstrators' activities (WP5), the work performed in the context of D2.3 was more targeted to specific data handling aspects. Thus, the performed state of the art review of the data management and data value enrichment methods was significantly reduced to very targeted cases. The most challenging part of the activities reported in the current deliverable was to seamlessly synthesize insights from all aforementioned sources in order to define a detailed and robust Big Data Consolidation Framework, built on rich data and metadata models and comprising specific data management, analysis and brokerage methods.

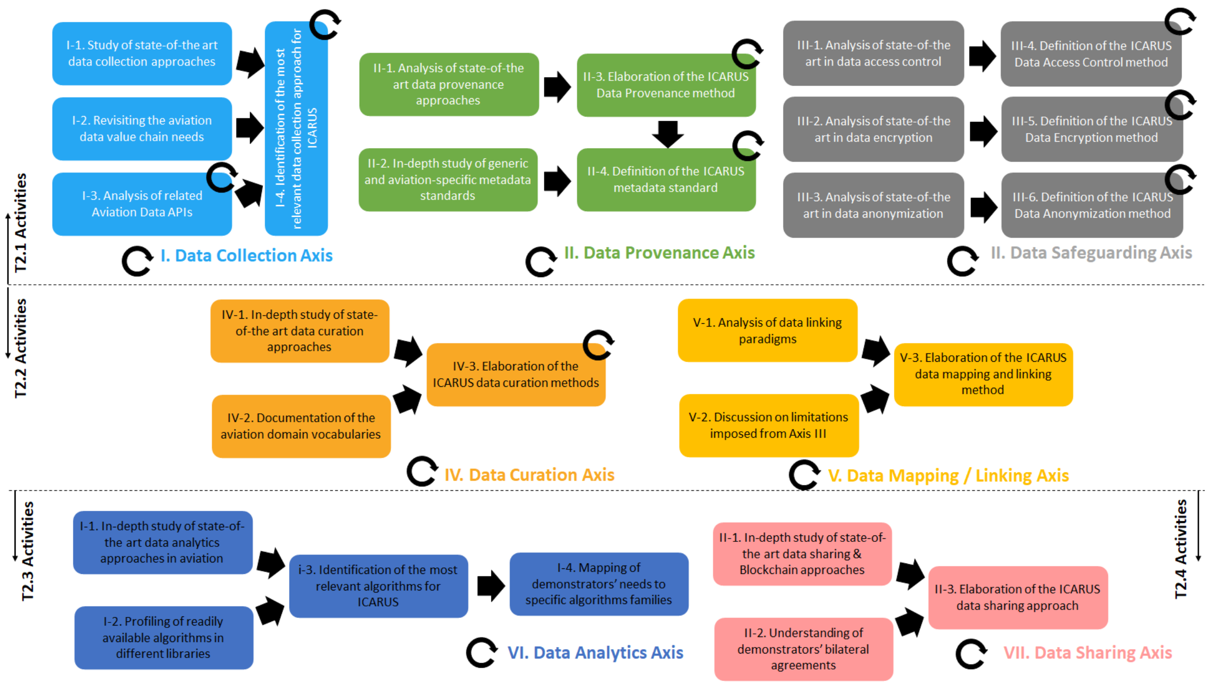

T2.1 Activities
I-1. Study of state-of-the art data collection approaches
I-2. Revisiting the aviation data value chain needs
I-3. Analysis of related Aviation Data APIs
I-4. Identification of the most relevant data collection approach for ICARUS
I. Data Collection Axis
II-1. Analysis of state-of-the art data provenance approaches
II-2. In-depth study of generic and aviation-specific metadata standards
II-3. Elaboration of the ICARUS Data Provenance method
II-4. Definition of the ICARUS metadata standard
II. Data Provenance Axis
III-1. Analysis of state-of-the art in data access control
III-2. Analysis of state-of-the art in data encryption
III-3. Analysis of state-of-the art in data anonymization
III-4. Definition of the ICARUS Data Access Control method
III-5. Definition of the ICARUS Data Encryption method
III-6. Definition of the ICARUS Data Anonymization method
II. Data Safeguarding Axis
T2.2 Activities
IV-1. In-depth study of state-of-the art data curation approaches
IV-2. Documentation of the aviation domain vocabularies
IV-3. Elaboration of the ICARUS data curation methods
IV. Data Curation Axis
V-1. Analysis of data linking paradigms
V-2. Discussion on limitations imposed from Axis III
V-3. Elaboration of the ICARUS data mapping and linking method
V. Data Mapping / Linking Axis
T2.3 Activities
I-1. In-depth study of state-of-the art data analytics approaches in aviation
I-2. Profiling of readily available algorithms in different libraries
I-3. Identification of the most relevant algorithms for ICARUS
I-4. Mapping of demonstrators' needs to specific algorithms families
VI. Data Analytics Axis
II-1. In-depth study of state-of-the art data sharing & Blockchain approaches
II-2. Understanding of demonstrators' bilateral agreements
II-3. Elaboration of the ICARUS data sharing approach
VII. Data Sharing Axis
T2.4 Activities

Figure 1-1: Approach bringing together the D2.1 and D2.2 Approaches

1.3 Relation to other ICARUS Results

As depicted in Figure 1-2, D2.3 is released in the scope of the WP2 "ICARUS Big Data Framework Consolidation" activities as the final outcome of all WP2 tasks and specifically T2.1 "Data Collection, Provenance and Safeguarding Methods", T2.2 "Data Curation, Harmonisation and Linking Frameworks", T2.3 "Deep Learning and Prescriptive Analytics Algorithms" and T2.4 "Data Policy and Assets Brokerage Frameworks".

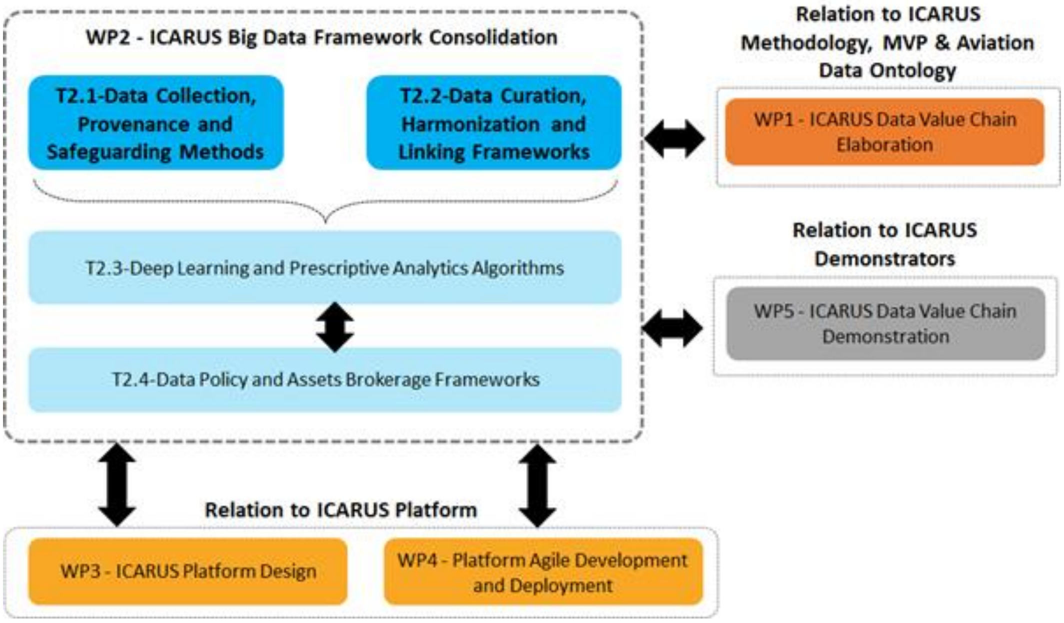

WP2 - ICARUS Big Data Framework Consolidation
T2.1-Data Collection, Provenance and Safeguarding Methods
T2.2-Data Curation, Harmonisation and Linking Frameworks
T2.3-Deep Learning and Prescriptive Analytics Algorithms
T2.4-Data Policy and Assets Brokerage Frameworks
Relation to ICARUS Methodology, MVP & Aviation Data Ontology
WP1 - ICARUS Data Value Chain Elaboration
Relation to ICARUS Demonstrators
WP5 - ICARUS Data Value Chain Demonstration
Relation to ICARUS Platform
WP3 - ICARUS Platform Design
WP4 - Platform Agile Development and Deployment

Figure 1-2: Relation to other ICARUS Work Packages

The work performed in the context of WP2 is strongly dependent on the outcomes of WP1 "ICARUS Data Value Chain Elaboration" with regard mainly to the final ICARUS methodology, the final ICARUS Minimum Viable Product (MVP) and the aviation data ontology. Insights from the activities performed in WP1, the final outcomes of which are documented in D1.3, have significantly affected the complete ICARUS data value chain, shaping the data management perspectives and outlining requirements, limitations and considerations for the data value enrichment methods.

The approach specified in the initial WP2 deliverables (D2.1 and D2.2) has been elaborated and, to an extent, applied during the design and development of the ICARUS platform in WP3 (in D3.1, D3.2 and D3.3) and WP4 (in D4.1 and D4.2), respectively. Feedback gained through this process, together with initial insights from work performed in the context of WP5 regarding the demonstrators' scenarios (in D5.2), has helped in refining the final ICARUS Big Data Framework, which is presented in the current deliverable (D2.3). The models and the methods presented in D2.3, but also the discussions around the identified challenges and important considerations, will feed subsequent activities in WP3 and WP4 and will also serve as facilitators for the WP5 activities through the provision of data management, analysis and sharing guidelines.

1.4 Structure of the Document

The structure of the document is as follows:

- Section 2 presents and discusses the ICARUS approach across all data management methods which comprise data collection, data curation (further detailed into cleansing, provenance and mapping aspects) and data safeguarding (further detailed into access control, encryption and anonymisation).
- Section 3 presents and discusses the ICARUS approach across the two main aspects of the data value enrichment methods foreseen in ICARUS, namely data analytics and data sharing. The data analytics aspect extends over the demonstrator-related perspectives, whereas data sharing involves the description of the ICARUS asset sharing model and the final definition of the Data Policy and Assets Brokerage Framework.
- Section 4 presents a discussion around the challenges that emerge from the designed data management and data value enrichment methods. The ways these challenges are handled and/or addressed are also described in detail, thus the aforementioned discussion can also serve as an outline of the ICARUS positioning regarding the different data-related challenges the aviation data value chain faces.
- Section 5 reports on the conclusions deriving from the work performed and documented in the deliverable at hand, as well as the directions and recommendations for the next steps.
- Annex I lists the references included in the present deliverable.
- Annex II presents the updated ICARUS metadata schema.

2 Data Management Methods

2.1 ICARUS Data Collection Methods

Data Collection is a broad term that refers to the population of the ICARUS platform with high-quality data from distributed information sources at proper granularity levels, in a timely manner. In general, the data collection activities are often mentioned in the international bibliography with diverse terms, such as data discovery, data harvesting, data ingestion and data acquisition that are practically differentiated in the method, frequency and origin of data that are collected.

As presented in D2.1, the ICARUS Data Collection Methods are designed taking into account the data assets profiling (by the ICARUS demonstrators and the core data provider, OAG) in the ICARUS Deliverable D1.1 and are revised in this document to consider both the updates in D1.3, but also the experience and lessons learnt that the consortium has gained from the beta release of the ICARUS platform. The ICARUS Data Collection Methods are based on a supply-driven mentality (since the data providers are responsible for collecting, ensuring the quality and checking in their data assets) and are presented under 3 core axes: Stakeholders, Applicable Processes and Data Profile, as explained in the following paragraphs and summarized in Figure 2-1.

Related Stakeholders. The ICARUS data collection method involves different aviation data value chain stakeholders that act as data providers in different modalities:

- Upstream data collection modality referring to the direct data discovery and gathering from their rightful source, that refers to aviation data stakeholders at the 1st and 2nd level. In the ICARUS context, the upstream dimension is covered by the ICARUS demonstrators.
- Downstream data collection modality encompassing the aggregation and distribution of data through an intermediary which typically acts as a data broker between supply (from data owners) and demand (from data consumers). In ICARUS, the downstream aspects are dealt with by the ICARUS aviation data provider, OAG.
- Indirect data collection modality. Thanks to the ICARUS scope of facilitating data linking and secure analytics, derivative data and intelligence as emerging from an analysis performed may also be indirectly collected to the platform according to the preferences of the data consumer and the licenses of the initial data assets upon which the analysis was performed. In ICARUS, the indirect data collection may originate from any of the 3 data tiers to which the aviation data value chain is classified.
- Open data collection modality embracing the open data repositories and open data sources that were considered as highly relevant for the aviation data value chain.

Applicable Processes. In order to become available in ICARUS, all data assets need to officially undergo the ICARUS checkin process that properly prepares the instructions to be applied prior to uploading any data asset in the ICARUS platform and records their associated metadata. The **data checkin**

process for a data asset that is uploaded for the first time to the ICARUS platform bears the following steps:

I. Uploading a representative **data sample** of the dataset that includes around 10-15 rows of data in the ICARUS platform. Such data may be even tampered with, since the purpose is only to identify the structure and define the processing instructions for the whole dataset.

II. Definition of the **data curation methods** that are to be applied on the complete data asset locally. Such methods include: (a) **Mapping the data structure to the ICARUS common aviation model** as elaborated in section 2.2.3, and (b) Design of the **cleansing rules** that are to be applied on each column of the dataset as described in section 2.2.1.

III. Identification of the **anonymization rules** that need to be employed on specific columns of the dataset that the data provider considers as sensitive as described in section 2.3.3.

IV. Selection of whether the **encryption** techniques that are described in section 2.3.2 are applicable in the whole dataset or on specific columns, and whether certain columns are to become searchable / indexed (by extracting their collective values prior to their encryption).

V. Verification of the **check-in configuration** in order to send the processing instructions (for steps II-IV) to be executed locally in the OnPremise Worker (as described in the ICARUS deliverables D3.1, D3.2).

VI. Provision of the **dataset metadata** in order to create its thorough data profile in the ICARUS platform. Such a metadata entry is compliant with the ICARUS metadata schema (defined in Annex II) and may be either minimal ("fast-lane") for confidential data that are not to be shared with other stakeholders or thorough: (a) for private data that are to be traded in the ICARUS platform, and (b) for open data that are to be shared through the ICARUS platform.

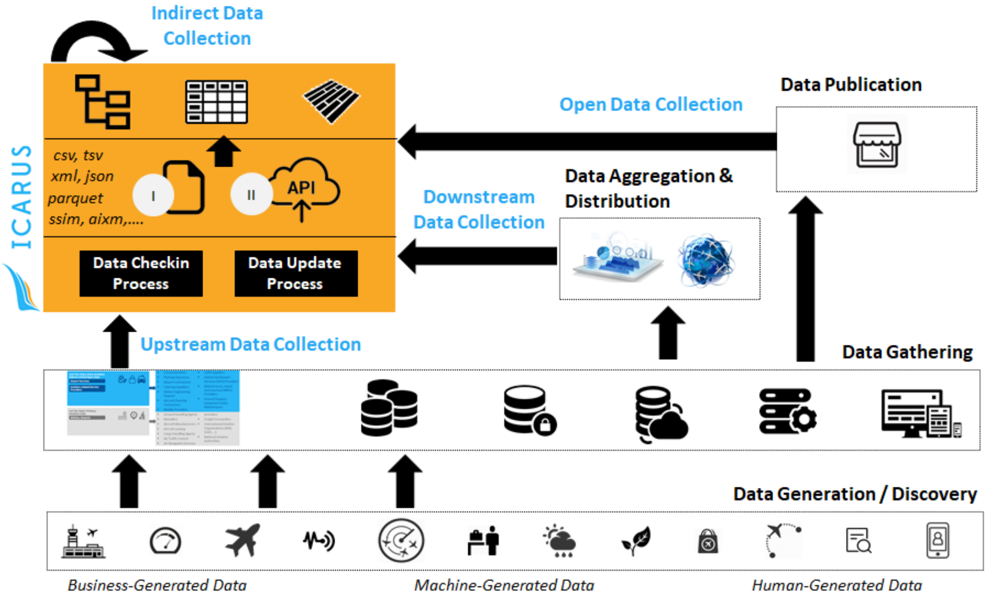


Figure 2-1: ICARUS Data Collection Approach

VII. Definition of the **data access policies** that are to be enforced in the ICARUS platform every time data consumers search for data as explained in section 2.3.1.

VIII. Preparation of the data locally and **uploading the data payload to the ICARUS platform** in a secure and efficient manner.

IX. **Transformation of the data sample** to comply with the rules II-IV.

In cases of frequent updates to the data that have been already collected in ICARUS, a flexible approach to append data to existing data assets shall be adopted, without compromising the data security and without requiring repeating any pre-processing tasks that have already been performed on the data. In this context, the **data update process** for a data asset that is already available in the ICARUS platform bears the following steps:

A. Selection and verification of the **check-in configuration** in order to send the stored processing instructions (that were defined in the data checkin process, steps II-IV) to be executed locally in the OnPremise Worker (as described in the ICARUS deliverables D3.1, D3.2).

B. Decision on the **data update strategy** that is applicable, i.e. append new data to the existing dataset or replace specific data within the dataset that are not up-to-date anymore.

C. **Updates on the dataset metadata and the data access policies** whenever necessary.

D. Preparation of the data locally and **uploading the data payload to the ICARUS platform** in a secure and efficient manner.

It needs to be noted that backward compatible changes in the check-in configuration are generally allowed, but are discouraged since they need to be propagated on the data that are already uploaded in the ICARUS platform. Non-backward compatible changes (e.g. change of data sample, different mapping of the data) require the full check-in process to be repeated for a new dataset, without though deleting the existing dataset since it may be already bound with data contracts.

Data Profiling. The ICARUS Data Collection Methods are properly designed to target only **data at rest**, namely "historical" data that are uploaded to the ICARUS platform in batches, due to the strong security requirements that are enforced in the aviation domain and prohibit the platform from effectively handling both real-time and batch data (as it would require totally different architectures). As explained in D2.1, the de facto data approach that is provided in ICARUS concerns Level 2: Files Upload / Exchange in which the data assets are generally pre-processed by their respective providers at the server side where multiple data tables (extracted from a legacy system) can be combined into a unique asset / file. In addition, reaching Level 3: APIs Release is also desirable in ICARUS in order to automate up to an extent the data collection process, even though the demonstrators do not already provide any APIs to effectively expose their data. With regard to the optional data formats that are to be supported in ICARUS, they include text data in: (a) tabular-like formats with delimiter-separated values, such as csv (comma-separated values) and tsv (tab-separated values), (b) data interchange formats like XML and JSON, (c) aviation-specific formats, such as IATA's SSIM (Standard Schedules Information Manual), and AIXM (Aeronautical Information Exchange Model) which are though typically based on either tabular-like formats or XML, and (d) flat columnar formats such as Parquet. The beta release of the ICARUS platform supports only tabular-like formats, but the next platform releases are expected to support additional formats depending on the feedback that will be received by the aviation data providers.

In order for ICARUS to effectively handle any data assets that may be checked in, they need to be transformed to a common format prior to their storage (in addition to their mapping to the ICARUS common aviation model). ICARUS shall adopt a tabular format due to the familiarity and popularity it has gained with data scientists of any data tier, taking into account the data profiling that has been reported in D1.1-D1.3 and the requirements of the ICARUS data value chain that derive from WP1 and WP3. However, this format may be reconsidered in case: (a) any performance issues are noticed in the ICARUS platform beta release, so alternative formats, such as Parquet which is a columnar storage format, shall be further examined in future platform releases, and (b) a tree data structure is deemed more appropriate for the data that will eventually populate the ICARUS platform.

It needs to be noted that the latest aviation data profiling is provided in D1.3 while the profiling of the aviation data APIs that was documented in D2.1 is maintained online, yet no updates have been recorded at the moment D2.3 was prepared so the respective section is not repeated in this deliverable. However, the APIs documentation, as well as the data profiling, will continue to be monitored and enriched throughout the project implementation.

2.2 ICARUS Data Curation Methods

The ICARUS Data Curation Methods consist of techniques and approaches for data cleaning, data provenance and data mapping and linking. The respective methods that had been originally defined in D2.1 are revisited and refined with minor or major improvements that are explained in the next paragraphs in order to reflect the latest advancements and perspectives in alignment with the ongoing ICARUS platform development activities.

2.2.1 Data Cleaning

Data Cleaning (or Data Cleansing) is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data (Wu, 2013). Hence, the main purpose of the Data Cleaning process is to improve the overall quality and usability of the data asset by employing: (a) a set of the validation rules, that are covering several aspects of the data quality dimensions and the data validation practises, in order to identify the possible errors or inconsistencies, and (b) a set of data cleaning and data completion techniques in order to correct or eliminate these identified errors or inconsistencies.

In the ICARUS perspective, the scope of the data cleaning process is to safeguard the quality of the produced results of a conducted data analysis by removing or correcting erroneous data that would lead to incorrect, inaccurate or even invalid results or conclusions.

The Data Cleaning process contains a series of steps related to the assessment and analysis of the data, as well as the refinement or removal of parts of the data as a result of the corrective actions that are performed based on the initial assessment and analysis. As such, the Data Cleaning process includes, among others, the definition and determination of the error types, the search and identification of the error instances and finally the correction of the uncovered errors (Maletic & Marcus, 2000).

Overall, the ICARUS Data Cleaning approach includes the following steps:

- **Preliminary Data analysis:** The purpose of the preliminary data analysis is to inspect and identify the data elements and their corresponding data characteristics. An analysis is performed from which a set of characteristics derives for the data elements such as the data type, the value format, the value pattern and the distinct values of the data elements. The extracted information is facilitating the assessment and the analysis of the data in the next steps.
- **Definition of the validation rules:** In order to detect inconsistencies, erroneous entries and any missing entries, a set of validation rules are defined. These validation rules include a list of constraints tailored to each data element. The errors are identified by evaluating the conformance to these constraints. The data provider is responsible for providing his/her input in the process by selecting the suitable constraint for each data element from the list of constraints that are offered by the data cleaning process. In this way, a level of customisation is offered by the process based on the nature of the data that will be cleaned, but also based on the needs of the data provider. The list of constraints can be grouped into technical validation and logical validation checks. The technical validation checks include, but are not limited to, the following rules:
    - Data type conformance (integer, string, etc.)
    - Value representation conformance (e.g. dates should be in the format "yyyymmdd")
    - Acceptable value range conformance (minimum and maximum acceptable value)
    - List of acceptable values conformance (e.g. airport codes, airline codes)
    - Uniformity conformance (e.g. all time-stamps are in UTC, weight values is in KGs)
    - Uniqueness conformance (i.e. no duplicate values are acceptable)
    - Required attributes conformance (i.e. mandatory fields should have values)

  Additionally, the logical validation checks include, among others, the following rules:

    - Cross-field validity conformance (e.g. the sum of fields with percentage values must be equal to 100)
    - Cross-field dependency conformance (e.g. if field A is set to zero then field B is a mandatory field)
    - Logical errors compliance (e.g. if field A is set with a value, then field B should not have the same value in the same record)

  The validation errors are identified by evaluating the conformance to these constraints as set by the data provider during the data preparation phase before they are uploaded to the ICARUS platform.
- **Definition of the cleansing workflow with the cleansing and missing value handling rules:**

In addition to the validation rules, a set of cleansing rules are defined and are bound to the validation rules. The cleansing rules define the corrective actions that are performed if a validation rule is violated and an error is identified. The corrective or removal actions are dependent on the nature of the identified error and the data provider's needs. The list of corrective actions includes, but is not limited to, the following:

- Rejection and removal of an inconsistent value based on a data validation rule (such as data type conformance).
- Replacement of an inconsistent value based on a data validation rule by applying a variety of methods such as the median, mean or most frequent value, the Linear Regression or the k-Nearest Neighbours algorithm, among others.
- Rejection and removal of an inconsistent record (set of values) based on a data validation rule (such as the required attributes conformance).

Moreover, a set of missing value handling (data completion) rules are defined and are also bound to the validation rules. The missing value handling rules define the corrective actions that are applied in order to perform automatic filling of the missing values based on the required attributes conformance errors. The list of techniques utilised for the automatic filling is the following:

- Basic statistic methods such as mean, median, most frequent value
- Linear Regression
- Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB) methods
- k-Nearest Neighbours algorithm
- Moving Average method
- Replacement of the empty value with a predefined value.

As with the validation rules, these rules are defined by the data provider during the data preparation phase before they are uploaded to the ICARUS platform.

- **Cleansing workflow execution:** Following the workflow specification step, the execution of the designed workflow is performed. In this step, the validation rules are evaluated and the identified errors are eliminated and corrected based on the corrective actions defined in the cleansing and missing value handling rules. During the execution of the workflow, detailed records are maintained (in line with the data provenance principles of section 2.3.3), containing information for the identified errors and the actions performed. These records can be provided for inspection and verification towards the assessment of the designed workflow. The execution of the designed cleansing workflow is performed in the background and the results of the execution are provided to the next step of the data preparation phase. Moreover, the detailed records are provided in the next (optional) step of the Data Cleaning process.
- **Verification:** In this optional step, the designed cleansing workflow and the results of its execution can be verified, and an assessment can be performed by the data provider on the

correctness and effectiveness of the workflow by inspecting the detailed records of the performed cleaning process.

2.2.2 Data Provenance

Data provenance is typically associated with the evidence-based detection of the origin and the evolution over time of a data asset, as well as of all its related processes, while contributing to determine any controversial data ownership aspects. As explained in D2.1, data provenance is considered in ICARUS from a rather coarse-grained perspective at data asset level, during its whole lifecycle (from its check-in to its sharing and disposal) in accordance with the final ICARUS methodology defined in the ICARUS Deliverable D1.3.

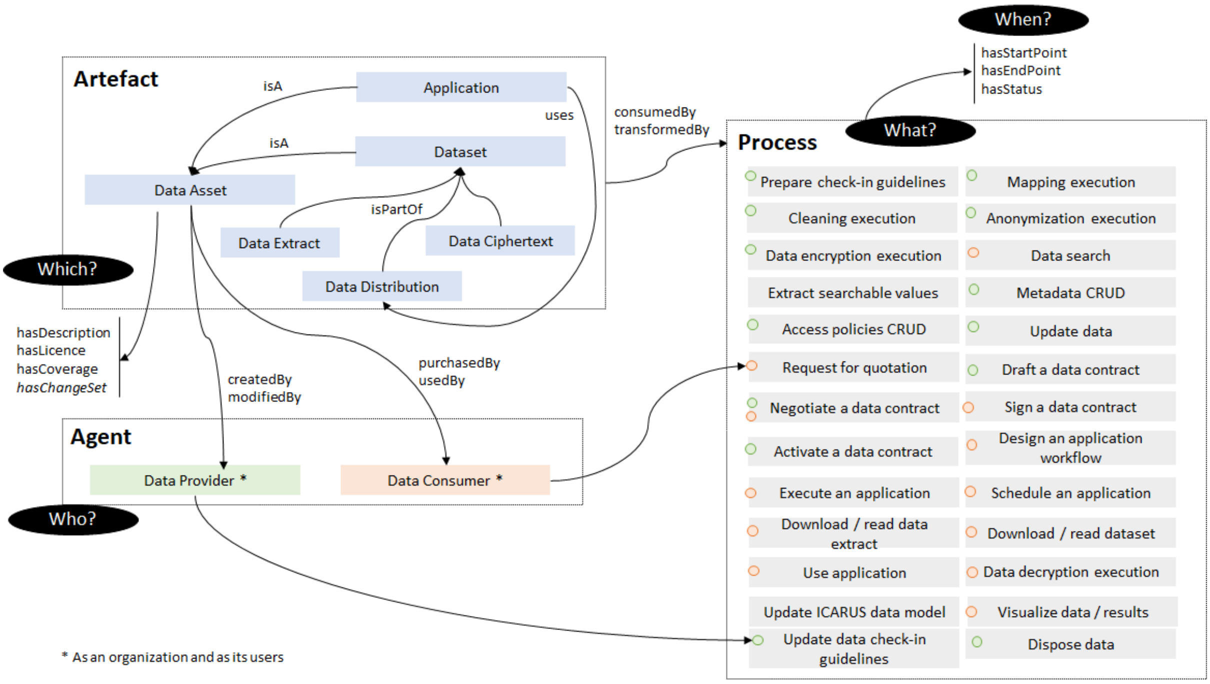
The ICARUS data provenance method aims at capturing and managing trustworthy data asset trails that shall effectively detect the lineage and the derivation of data (esp. private data) in an immutable manner in the background. As depicted in the following figure, the ICARUS-relevant provenance information complies with the W3C PROV Data Model and spans over 4 core axes:

- **The Agent Perspective: Who?** Information about who published (and practically owns) a data asset and who has consumed a free data asset or signed a data contract to obtain a private, paid data asset needs to be formally kept in order to map the identity of all actors (i.e. organizations and their members / users) in the data asset lifecycle.
- **The Artefact Perspective: Which?** In order to avoid confusion in the terminology, ICARUS adopts the following data-related terms:
    - Data Asset, an umbrella term for any dataset, its distributions and its extracts.
    - Dataset that refers to an identifiable collection of data, provided by a data provider and available for access or download in one or more data distributions.
    - Multiple Data Distributions as specific forms (e.g. different formats) in which a dataset is available, typically limited by some constraint such as spatial extent or temporal coverage.
    - Data Extract that contains sample data (up to 15 rows) of a dataset in a specific indicative distribution.
    - Data Ciphertexts which encapsulate the data that are encrypted on a column basis in accordance to the preferences of the data provider.
    - Data Applications (referred to as bundles in D2.1) containing a combination of data assets needed to run a specific workflow of algorithms and visualizations that are appropriate for gaining insights on the algorithms' outcomes.
- **The Process Perspective: What?** The potential "operations" and activities that are applicable on a data asset, e.g. data check-in (along with the settings and status in its intermediate stages: mapping, cleaning, anonymization, encryption, storage, indexing, metadata definition, data access policies definition), data search, data contract preparation (i.e. request for quotation, draft a data contract, sign a data contract, negotiate a data contract, activate a data contract), download/read data (requesting the related decryption key whenever

applicable), design an application (define the application workflow, define the application metadata), run / execute an application, schedule an application, visualize data, visualize results, update / evolve data, update / evolve the ICARUS common aviation data model, and dispose data (in the condition they do not have any active data contract). Such activities may be performed by data providers and / or data consumers.

- **The Underlying Timing Perspective: When?** The time (start time, end time) at which any data operation occurred as a core provenance characteristic in order to trace irregularities related to data assets.

Since the provenance trails cannot refer to the actual data included in a data asset due to inherent restrictions imposed from the ICARUS encryption schemes (as explained in section 2.3.2), the data assets values cannot be monitored and actual reproducibility of the data cannot be achieved (e.g. to view intermediate data or to replay possibly alternative data processing steps on intermediate data), yet a full history log of the actions related to the whole data asset will be diligently maintained.



* As an organization and as its users

Figure 2-2: ICARUS Data Provenance Schema

As part of the ICARUS provenance method, the Dataset Usage Vocabulary (W3C, 2016) will be also considered, yet it will be clearly stated in the ICARUS platform terms of use that analytics related to the usage of data assets are provided to the data providers, explaining in detail why and how information is gathered from data consumers (without compromising their privacy and the analysis they perform on their secure experimentation spaces).

In terms of storage, the provenance trails are decoupled from the original data assets that are typically encrypted in the ICARUS platform, and may accompany the asset's metadata (in alignment with the ICARUS metadata schema described in Annex II).

In summary, the ICARUS provenance model and overall approach remains largely unchanged since D2.1, with the exception of the process perspective that has been further elaborated to reflect the increased number and complexity of the activities for which provenance metadata need to be considered.

2.2.3 Data Mapping and Linking

Data Mapping and Linking encompass methods and techniques to address the inherent semantic interoperability problem at syntactic, schematic and semantic heterogeneity levels, that appears in any data integration endeavour. In general, lack of global standardization through globally agreed semantic models and common aviation schemas that address the needs of the overall aviation value chain hinders any data sharing efforts. In aviation as in most industries today, the prevalent "standards dilemma", defined as the diversity of standards (such as the Airport Collaborative Decision Making Manual (A-CDM), the Standard Schedules Information Manual (SSIM), the Aeronautical Information Exchange Model (AIXM) and the ACI Airport Community Recommended Information Services (ACRIS)) that address particular data requirements, but are designed on such a different basis that make the choice of a specific standard to be adopted a new challenge, is compounding the problem.

In ICARUS, a common aviation data model reconciling the different aviation data standards is considered as instrumental to ensure effective data integration at data check-in time and at data query time. To this end, a data model has been meticulously "designed for change" with the purpose of efficiently managing its whole lifecycle and effectively anticipating its consistent evolution (e.g. how new concepts will be effectively incorporated, without disrupting the existing model, in a way that ensures backward compatibility) as it is unrealistic to consider that any data model, no matter how well designed, will be inclusive of all the aviation-related data from the whole aviation ecosystem from its early beginning and shall address all future data needs the aviation stakeholders may have. The ICARUS data model lifecycle thus consists of 8 phases that include:

- Phase I: Modelling during which certain preparatory activities have been performed and the ICARUS common aviation data model has been constructed. The preparatory activities included two parallel streams: (a) the study of the ICARUS aviation ontology, based on the NASA ATM Ontology and considering the data collection activities from the ICARUS demonstrators and OAG, that were conducted in WP1 and documented in D1.3, (b) the analysis of a set of aviation data standards that were prioritized, namely: A-CDM, ACRIS, AIXM, and partly SSIM (through the OAG data as the full standard was not available at the whole consortium at the time the initial version of the ICARUS common aviation data model was prepared), as well as a generic purpose data standard like UN/CEFACT CCTS (Core Components Technical Specification).

  The ICARUS common aviation data model currently contains 9 core entities, namely: Aircraft, Airport, Booking, Carrier, Flight, Flight Leg, Passenger, Product and Weather, that collectively contain over 190 properties. As depicted in the following extract, the core entities are described based on metadata like "definition", "related_terms", "standards", "data_added", "date_deprecated", "version", and "children" while their properties feature metadata such as "definition", "type", "related_terms", "standards", "data_added", "date_deprecated", "version", and "facet" (to enforce non-encryption on certain, non-business critical properties that will act as filters to facilitate their acquisition).

```
{
        "flightLeg": {
                "definition": "A single journey (flight segment) from origin to
destination, covering the aircraft movement from the departure at the originating
airport to the arrival at the destination airport.",
                "related_terms": [
                        "single journey",
                        "flight segment"
                ],
                "standards": [
                        "AIXM"
                ],
                "data_added": "24/05/2019",
                "date_deprecated": null,
                "version": 1.0,
                "children": {
                        "arrivalAirport": {
                                "definition": "Details for the actual arrival airport
for the flight.",
                                "type": {
                                        "$ref": "#/airport"
                                },
                                "related_terms": [
                                        "scheduled arrival airport",
                                        "destination airport",
                                        "ADES",
                                        "Aerodrome of Destination"
                                ],
                                "standards": [
                                        "A-CDM",
                                        "AIXM",
                                        "ACRIS"
                                ],
                                "data_added": "24/05/2019",
                                "date_deprecated": null,
                                "version": 1.0,
                                "facet": "encryption allowed"
                        },
                        "plannedDepartureTime": {
                                "definition": "The time that the flight is scheduled to
depart per the flight plan. The estimated time at which an aircraft will become
airborne.",
                                "type": "datetime",
                                "related_terms": [
                                        "STOD",
                                        "Scheduled Departure Time",
                                        "Estimated Take Off Time",
                                        "ETOT",
                                        "Scheduled Time of Aircraft Departure",
                                        "STD",
                                        "Scheduled Date of Departure",
                                        "Scheduled Time of Departure"
                                ],
                                "standards": [
                                        "NASA ATM Ontology",
                                        "A-CDM",
                                        "AIXM",
                                        "ACRIS"
                                ],
                                "data_added": "24/05/2019",
                                "date_deprecated": null,
                                "version": 1.0,
                                "facet": "encryption prohibited"
                        },
                …
                }
        }

}
```

- Phase II: Model Storage that properly and securely stores the model in its JSON representation in order to be easily accessible at run-time.
- Phase III: Mapping Algorithms Definition embracing the design of algorithms for effectively mapping the data that are checked in in ICARUS (source schema) to the underlying ICARUS common aviation data model (target schema). In ICARUS, such algorithms range from traditional schema matching algorithms (that leverage the domain knowledge) to supervised machine learning algorithms (which learn from the data that are mapped) that shall be employed to calculate the mappings between source and target schema, at run-time.
- Phase IV: Mapping Algorithms Training, referring to the "offline" use of specific small training datasets that have been created by ICARUS to fit and tune the mapping algorithms that have been created in Phase III.
- Phase V: Semi-automated Data Mapping that practically executes the mapping algorithms and proposes specific mappings between the data that are checked in and the ICARUS common aviation data model. The mapping strategy that ICARUS intends to follow is summarized as follows:
    - *Mapping Case 1 (1:1)* - Exact match for a property of the data that are checked in is found in the title of a property in the ICARUS common aviation data model.
    - *Mapping Case 2 (1:1)* - Exact match for a property of the data that are checked in is found in the related terms of a property in the ICARUS common aviation data model.
    - *Mapping Case 3 (1:1)* - Similar match for a property of the data that are checked in is found in the title and / or the related terms of a property in the ICARUS common aviation data model.
    - *Mapping Case 4 (1:N)* - Multiple matches for a property of the data that are checked in is found in the related terms of different (even though related) properties in the ICARUS common aviation data model.
    - *Mapping Case 5 (N:1)* - Multiple properties of the data that are checked in are mapped in the same property in the ICARUS common aviation data model.
    - *Mapping Case 6 (1:0)* - No matches for a property of the data that are checked in is found in the properties in the ICARUS common aviation data model.

As the international bibliography also concurs, the schema matching algorithms and mapping techniques have significantly improved over the years and broadly yield satisfactory results, yet they cannot achieve 100% accuracy and the human intervention is always necessary to confirm or update the automatically calculated mappings, especially for the mapping cases (4)-(6). Therefore, the data mapping techniques in ICARUS are characterized as semi-automated as the mappings must be confirmed by the data provider: (a) to correct any erroneously calculated mappings in any columns (in which many alternatives may be provided as in mapping case 4), (b) to provide the mapping in cases when it was not concluded by the mapping algorithms, even though it is foreseen in the ICARUS common aviation data model (as potentially in mapping case 6), or (c) to propose updates to the ICARUS common aviation data model when it does not support specific concepts (as in mapping case 6, again). Although cases (a) and (b) can be instantly handled, case (c) with proposals for changes in the ICARUS model requires the intervention of an administrator as described in Phase VI.

- Phase VI: Model Evolution which reflects the inevitable updates and changes that need to be performed on the ICARUS common aviation data model as time goes by, either spontaneously by the ICARUS administrators to anticipate new needs (e.g. an update of an existing aviation data standard or the emergence of a new data standard) or on demand to address specific proposals they have received by data providers who attempt to check in their data assets in ICARUS. The changes that are performed on the data model are classified as major or minor, and result into a new version of the data model that may be backward compatible (so no action is needed for data that are already checked in) or may be non-backward compatible (so certain action for propagating the changes need to be taken). In detail, all evolution events concern addition, update or deletion and are practically governed by the following rules:

```
ON ADDITION of a new CORE CONCEPT TO ICARUS CADM THEN PROPAGATE
ON ADDITION of a new PROPERTY TO an ICARUS CORE CONCEPT THEN PROPAGATE
ON ADDITION of [related_terms / standards] of a CORE CONCEPT TO ICARUS CADM
   THEN PROPAGATE
ON ADDITION of [related_terms / standards] of a PROPERTY TO an ICARUS CORE
   CONCEPT THEN PROPAGATE
ON UPDATE of the [title] of a CORE CONCEPT TO ICARUS CADM THEN BLOCK
ON UPDATE of the [definition / related_terms / standards] of a CORE CONCEPT
   TO ICARUS CADM THEN PROMPT
ON UPDATE of the [title] of a PROPERTY TO an ICARUS CORE CONCEPT THEN PROMPT
ON UPDATE of the [definition / related_terms / standards] of a PROPERTY TO an
   ICARUS CORE CONCEPT THEN PROMPT
ON UPDATE of the [type / facet] of a PROPERTY TO an ICARUS CORE CONCEPT THEN
   BLOCK
ON DELETION of a CORE CONCEPT TO ICARUS CADM THEN BLOCK
ON DELETION of the [related_terms / standards] of a CORE CONCEPT TO ICARUS
   CADM THEN PROMPT
ON DELETION of a PROPERTY TO an ICARUS CORE CONCEPT THEN PROMPT
ON DELETION of the [related_terms / standards] of a PROPERTY TO an ICARUS CORE
   CONCEPT THEN PROMPT
```

In general, the actions that are to be taken after an evolution event in alignment with the above evolution events include:

- *Propagate action* that signifies a backward compatible evolution event that can be adopted without creating any inconsistencies between the ICARUS common aviation data model and the data that are already checked in.
- *Prompt action* to embrace all evolution events that are typically non-backward compatible. Prior to such events being addressed in the ICARUS common aviation data

model, their impact on the data that are already checked in needs to be cross-checked and validated as it may require executing Phase VI (Data Transformation) again (e.g. to rename the properties' titles).

    - *Block action*, prohibiting the specific evolution event as it will create major problems on the data that have already populated the ICARUS platform.

  Although the model evolution can be generally handled in a semi-automated manner, all evolution rules need to be also manually checked to avoid any model inconsistencies. It needs to be noted though that the decision of not imposing any cardinality restrictions on the properties of the ICARUS common aviation data model, significantly simplifies the evolution phase.

- Phase VII: Data Transformation which is responsible for transforming the data structure of a dataset in accordance to the mapping rules. It needs to be noted that if the ICARUS common aviation data model imposes specific measurement units or code lists in specific properties, the specific phase can also undertake the responsibility for properly transforming the data, as well, as it is applied prior to any data cleaning, anonymization and encryption method (defined in sections 2.2.1, 2.3.3 and 2.3.2, respectively).
- Phase VIII: Data Linking that concerns how data that belong to different data assets can be potentially linked during query time, at schema level based on the ICARUS common aviation data model and their metadata (in accordance with the ICARUS metadata schema), to facilitate data consumers in exploring the data assets. In order for two data assets to be potentially linkable, they need to: (a) possess at least 1 common property in the ICARUS common aviation data model, and (b) have the same or overlapping values and / or temporal / spatial coverage in the specific property (i.e. to avoid cases in which the datasets display 1 common property, such as departureAirport, yet one dataset refers to airports in Germany and the other dataset to airports in Cyprus). Since the data are to be uploaded in ICARUS in an encrypted format, the precondition (b) will be ensured by extracting and listing their collective values in each column on the on-premise environment prior to encryption while the role of the respective spatial and temporal coverage metadata also supports identifying potential links among datasets in a more reliable manner. Multiple datasets may be also linked in different ways depending on what the data consumer is looking for, as long as they have common properties at least in pairs.

Such phases are practically interwoven to support 3 main workflows, namely Workflow I: At data model preparation time (based on the Related Data Model Lifecycle Phases: I-IV), Workflow II: At data check-in time (based on Related Data Model Lifecycle Phases: V-VII) and Workflow III: At data query time (based on Related Data Model Lifecycle Phases: VIII) that are summarized in the following figure.

It needs to be noted that overall, the data mapping and linking method has been significantly revised in respect to D2.1 since the data linking approach was initially tightly interconnected with the use of GraphQL. However, the early implementation of the specific approach proved ineffective in the ICARUS beta platform release due to the end-to-end data encryption that is enforced in the core platform, and might be eventually adopted only in the secure and private experimentation spaces, but it will be decided in due time as the development activities progress.

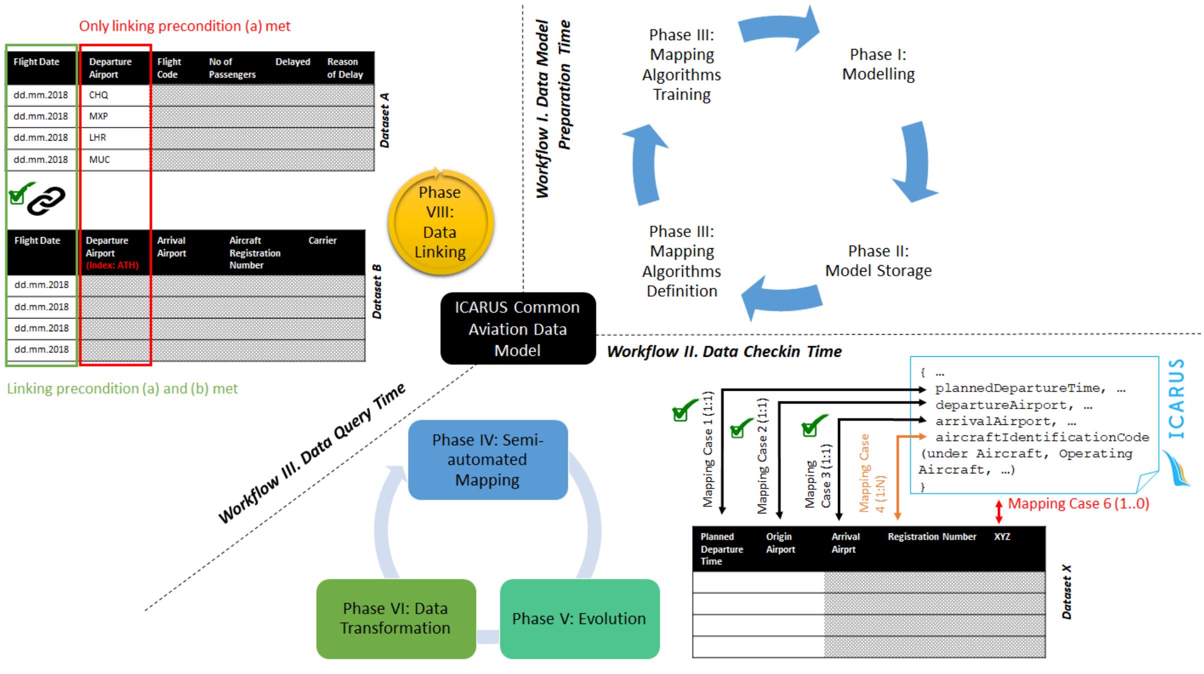


Figure 2-3: ICARUS Data Mapping and Linking Approach

2.3 ICARUS Data Safeguarding Methods

As explained in D2.1, the ICARUS Data Safeguarding Methods are viewed under the perspectives of: (a) data access control, (b) data encryption, and (c) data anonymization, that are discussed and refined to reflect the latest ICARUS implementation experience and advancements.

2.3.1 Data Access Control

The purpose of the ICARUS Data Access Control method is to define declaratively and deterministically the authorization policies for permitting or denying access requests to any data asset available in the ICARUS platform, in real-time. By effectively managing the whole policy lifecycle, such a method serves a dual purpose: to prevent unauthorized disclosure to private data assets (confidentiality) and any intentional or accidental unauthorized changes to data assets (integrity).

In ICARUS, access to data assets is regulated through Attribute-Based Access Control (ABAC) policies, based on the XACML standard, that allow the data providers to protect and share their data assets, even when they do not have any prior knowledge of the potential individual data consumers in the aviation data value chain. A proper separation of concerns between policy specification and policy enforcement is effectively pursued, while arbitrary attributes in policies are dynamically enforced.

In general, all XACML policies are expressed through:

- A Policy that refers to single access control expressed through a set of Rules, or

- A PolicySet, which acts as a container that can hold other Policies or PolicySets, as well as references to policies found in remote locations.

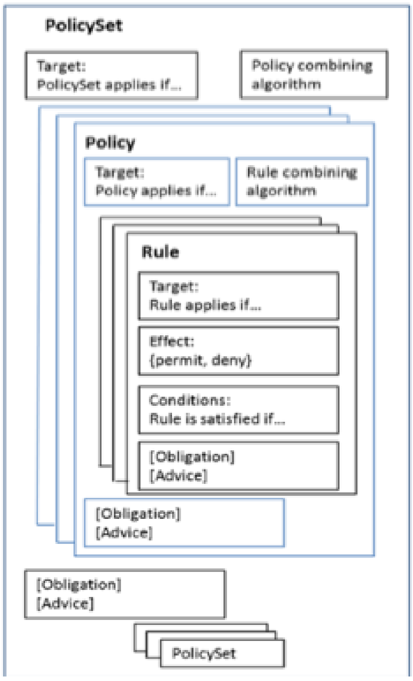


Figure 2-4: XACML Syntax which is applied in ICARUS (Ferraiolo et al, 2016)

Since a Policy or PolicySet may contain multiple policies or Rules, each of which may evaluate to different access control decisions (Permit, Deny, NotApplicable, or Indeterminate), XACML reconciles such individual decisions to arrive at an authorization decision through a collection of Combining Algorithms (Rule-combining algorithms or Policy-combining algorithms). In ICARUS, the Deny Overrides Algorithm (according to which if any policy / rule evaluation returns Deny, then the final result is also Deny) is employed for this purpose.

The rule template according to which the data access control policies are modelled complies with the following expression:

```
[subject] with [context expression] has [authorisation] for [action] on [controlled object]
```

Where:

`Subject` represents any representative of an organization in the aviation data value chain that has been uniquely identified and authenticated;

`Authorization` determines whether access is granted or not;

`Action` consists of the operations to be performed on the resource, such as Read-only in ICARUS or Download locally once or Always download locally, and needs to be always aligned with the data asset's license;

`Controlled object` refers to any data asset available in ICARUS (ranging from datasets and datasets extracts to algorithms and intelligence reports);

`Context expression` may include a combination of a number of attributes that are presented in the following table along the following categories:

- Subject Attributes for the user who issued a data asset request.
- Object Attributes related to the data assets for which access is sought, such as their metadata, their actual contents, or the existence of an active data contract between its provider and the subject.
- Environment Attributes derived from: (a) the current state of the system's environment, (b) the current session of a user, and (c) configuration settings applicable to the whole system which are either manually set by an administrator or by some automated process.

Table 2-1: ICARUS Data Access Control Attributes

Attributes Type	Indicative Attributes
Subject Attributes	Type of organization represented; Name of organization represented; Country of origin; Job title; Role; Security clearance; Trust level, IP of the Subject; User Agent of Subject; Geographic Region of the Subject
Object Attributes	Data Asset metadata according to Annex II; Actual contents (referring to temporal and spatial coverage for encrypted data assets); Data contract metadata
Environment Attributes	(a) Current time, day of the week, number of users logged in; (b) User's current session length, number of access requests made; (c) threat level (e.g., different policies could be used depending on whether or not the system was likely to be attacked), minimum trust level (e.g., the minimum amount of trust required for a user to access the system)

Such attributes are specified in XACML as name-value pairs, where attribute values can be of different types (e.g., integer, string). An attribute name/ID denotes the property or characteristic associated with a subject, resource, action, or environment.

An example of a valid XACML policy is presented below:

```
[subject]   with   [subject.country==GERMANY   &&   subject.organization==AirportX   &&
object.license=FREE && environment.session.authorized==TRUE]
has
[authorisation.ALLOW]
for
[action.READ]
on
[datasetx]
```

As presented in deliverable D2.1, the ICARUS data access control policy lifecycle consists of six phases in total. In brief, these phases are the following:

- Phase I: Definition during which all policies associated with a data asset are modelled at check-in time by the data provider.

- Phase II: Storage which is responsible for securely storing the data access policies.
- Phase III: Enforcement that evaluates the different access control policies associated with a data asset and takes decisions on whether to grant access to a data consumer.
- Phase IV: Reuse which potentially allows data providers to reuse the data access control policies defined for their data asset to other data assets they know, by packaging them as Policy Sets.
- Phase V: Evolution allowing for frequent updates of policies to address changing priorities and threats.
- Phase VI: Disposal which efficiently handles the removal of an access control policy (individually or as a batch at data asset level) taking into account the consistency of the related access policies (especially for the same data asset and for the same PolicySet).

Taking into account the XACML data flow defined in (OASIS, 2018), the ICARUS data access control method which is applied in ICARUS is built on the main functional points: the Policy Enforcement Point (PEP), the Policy Decision Point (PDP), the Policy Information Point (PIP), and the Policy Administration Point (PAP), which function together to provide access control decisions and policy enforcement. In practice, different workflows depending on the phase of the policy lifecycle are anticipated to be most frequently encountered in ICARUS as depicted in the following figure.

The basic workflows of the ICARUS Data Access Control were presented in D2.1. While no updates were introduced, for coherency reasons they are also presented below.

Workflow I: At data asset check-in time (Related Policy Lifecycle Phases: I and II)

In the PAP, the data provider defines the policies and policy sets that are related to the data asset that is checked in in the ICARUS platform at the given moment. For example, the data provider may define that “no airline will access the data asset” or that “only company X and Y can access the data asset” or that “only airlines from Greece or Cyprus can access the data asset”. The PAP is responsible for checking and ensuring the consistency of the access policies and policy sets defined for a data asset, for converting them into an XACML canonical form and for storing them in the PIP. PIP stores the arbitrary attributes of the policies and policy sets defined by the respective data provider and thus considered as trusted for the specific data asset.

Workflow II: At data query time (Related Policy Lifecycle Phase: III)

When a data consumer (representing company Z) implicitly requests access to a data asset through a query submitted to the ICARUS platform, e.g. he/she requests for “flight schedule data” and a potential result could be the “Europe flight schedule data for 2018” provided by OAG. Prior to returning the specific data asset in the list of results in the platform, the policies associated to it need to be resolved and access needs to be granted or denied.

The entire process involves the creation of a request for access to the respective data asset to the PEP that transforms it in an XACML canonical form and forwards it to the PDP. The PDP collects the policies / policy sets related to the specific data asset from the PIP and the values of the related attributes of the subjects, resource, action, and environment. At this stage, it is crucial that all attributes related to the affected policies and their values are globally accessible and tamper proof since the PDP may request to search the PIP for additional attributes if the attributes of the request are not sufficient for rule and policy evaluation. The PDP evaluates the policy and returns the response context (including the authorization decision) to the PEP. The PEP fulfils the obligations by permitting or denying access to the data asset.

For example, if the only rule that was evaluated for the "Europe flight schedule data for 2018" provided by OAG was that "no airline will access the data asset" and the data consumer was the Athens International Airport, then access would be permitted and the AIA representative would be able to navigate to the data asset, check its extract and eventually buy it.

It needs to be noted that XACML also includes the concepts of obligation and advice expressions: An obligation optionally specified in a Rule, Policy, or PolicySet is a directive from the PDP to the PEP on what must be carried out before or after an access request is approved or denied. Advice is similar to an obligation, except that advice may be ignored by the PEP. In the above workflow, thus, a potential obligation could be that the respective data provider is notified whenever a data consumer has tried to access a data asset he/she owns, but failed, while a potential advice to the data consumer could take the form of hints why access to an additional data asset (that could be potentially available) was denied.

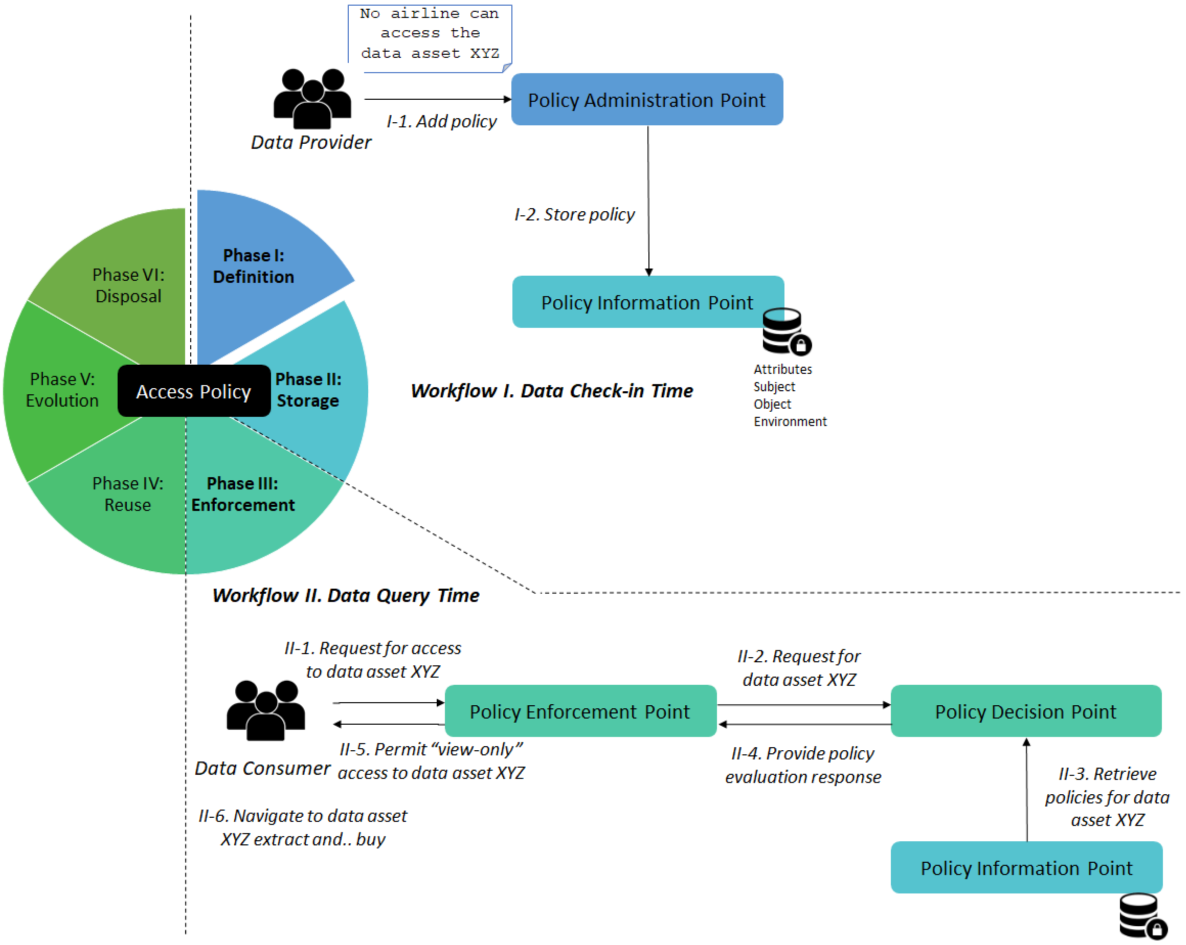

No airline can access the data asset XYZ
Data Provider
I-1. Add policy
Policy Administration Point
I-2. Store policy
Policy Information Point
Attributes
Subject
Object
Environment
Workflow I. Data Check-in Time
Phase I: Definition
Phase II: Storage
Phase III: Enforcement
Phase IV: Reuse
Phase V: Evolution
Phase VI: Disposal
Access Policy
Workflow II. Data Query Time
II-1. Request for access to data asset XYZ
Data Consumer
Policy Enforcement Point
II-5. Permit "view-only" access to data asset XYZ
II-6. Navigate to data asset XYZ extract and.. buy
II-2. Request for data asset XYZ
II-4. Provide policy evaluation response
Policy Decision Point
II-3. Retrieve policies for data asset XYZ
Policy Information Point

Figure 2-5: ICARUS Data Access Control Basic Workflows

Additional, similar workflows (mainly involving the PAP and the PIP) can be anticipated: (i) when a data provider packages the policies defined for a data asset to a Policy Set in order to be reused for other data assets that are to be uploaded (Related Policy Lifecycle Phases: IV and II), (ii) when a data provider updates the policies associated to his/her data asset in order to make more/less restrictive the data access, according to his/her preferences and commercial interests (Related Policy Lifecycle Phases: V and II), and (iii) when the policies related to a data asset are to be deleted, either 1 by 1 (if only a specific policy is not valid anymore) or at batch level (e.g. when a private data asset is no longer to be shared, but only for confidential use or when it is generally removed from the ICARUS platform).

Through the overall ICARUS data access control method, the separation of concerns between the access decision and the point of use is effectively ensured. By providing the opportunity to data providers to update the access policies on the fly and affect all potential data consumers immediately, it is also expected that the risk of unauthorized access to data assets will be significantly reduced, if not eliminated.

2.3.2 Data Encryption

The purpose of the ICARUS Data Encryption method is to ensure that the data assets shall be securely transmitted: (a) from the data providers' premises to the ICARUS platform and (b) from the ICARUS platform to the data consumer and the ICARUS secure experimentation spaces, as well as (c) stored in the ICARUS platform, without any alterations and only the authorized data consumers (who have an active data contract in the case of private data) shall be able to access and use the data asset. In ICARUS, a promising data encryption approach is put into use to protect the business confidentiality of the data assets and increase trust of the aviation stakeholders who are particularly sceptic to any data sharing approach to the ICARUS platform.

The ICARUS encryption method is based on a dual encryption approach bringing the best of breed from the symmetric encryption and the SSL worlds as follows:

- Symmetric key encryption for the data assets is considered as the most efficient solution for the aviation industry needs and the ICARUS scope, taking into account that huge amounts of data typically increase the security load and symmetric key encryption will not slow down the performance of the ICARUS platform as a whole.
- Secure SSL handshakes in order to securely share the symmetric key between: (a) the data provider and the data consumer, and / or (b) the data provider and the secure experimentation space of the data consumer in the ICARUS platform.

With the rapid advances in the cryptographic technology, it is generally challenging to choose which encryption algorithm is the most appropriate to use in each context, however the ICARUS consortium decided to utilize the AES256 symmetric key encryption algorithm.

The main phases of the ICARUS encryption-decryption workflow were presented in D2.1. While no updates were introduced, for coherency reasons they are also presented below.

In brief, the typical encryption-decryption workflow that is followed in ICARUS is described in the following three main phases:

<u>Phase I. Symmetric Key Encryption</u>

In practice, at data check-in time, a data provider decides whether and which columns of a data asset that shall become available in the ICARUS platform will be encrypted. It needs to be noted that in order for the ICARUS platform to seamlessly execute queries with a temporal and spatial restriction, it is mandatory that certain selected columns (that in any case do not reveal any business-critical information) will always remain unencrypted.

The data asset is actually encrypted at the premises of the data provider with the help of a locally generated symmetric key. The emerging data ciphertext is then transmitted and stored in an encrypted form in the ICARUS core platform to avoid both certain attacks and privacy breaches.

The ICARUS platform on its behalf cannot decrypt the data asset ciphertext (since it is not aware of the secret key that was used by the data provider for the encryption) and the actual data always remain private (even in the extremely unlike situation that the ICARUS platform was corrupted from internal and external attacks).

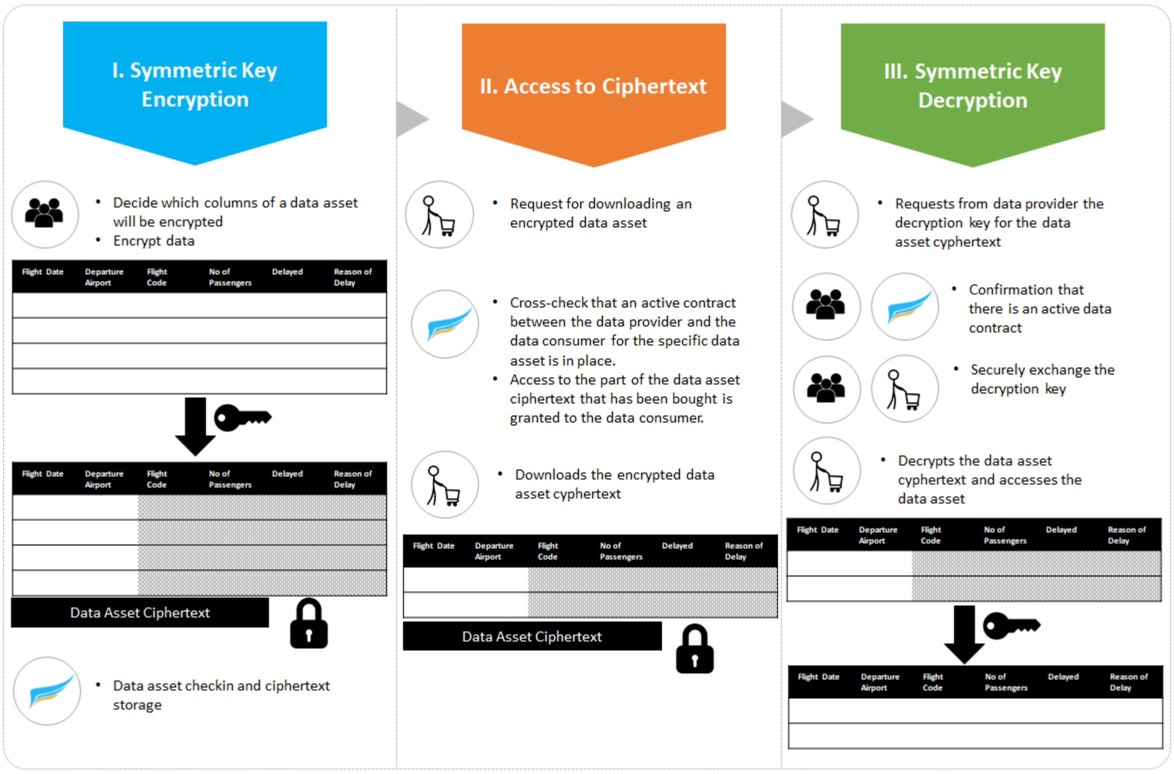

I. Symmetric Key Encryption
Decide which columns of a data asset will be encrypted
Encrypt data
Flight Date
Departure Airport
Flight Code
No of Passengers
Delayed
Reason of Delay
Data Asset Ciphertext
Data asset checkin and ciphertext storage
II. Access to Ciphertext
Request for downloading an encrypted data asset
Cross-check that an active contract between the data provider and the data consumer for the specific data asset is in place.
Access to the part of the data asset ciphertext that has been bought is granted to the data consumer.
Downloads the encrypted data asset cyphertext
Data Asset Ciphertext
III. Symmetric Key Decryption
Requests from data provider the decryption key for the data asset cyphertext
Confirmation that there is an active data contract
Securely exchange the decryption key
Decrypts the data asset cyphertext and accesses the data asset

Figure 2-6: ICARUS Data Encryption-Decryption Workflows

<u>Phase II. Access to Data Ciphertext</u>

A data consumer who is confirmed to be eligible to access and buy a data asset according to the ICARUS data access control method (described in section 2.4.1) expresses his interest and requests to download a specific data asset. The ICARUS platform cross-checks whether an active contract between the data provider and the data consumer for the specific data asset is in place: (a) If there is an active contract, then access to the part of the data asset ciphertext that has been bought is granted to the data consumer (e.g. the data consumer may have only bought data for a specific year or a specific location instead of the whole data asset). (b) If there is no active contract, then the data sharing workflow that is described in section 3.2.3 is practically followed in order for the data consumer to obtain access to a data ciphertext.

Phase III. Symmetric Key Decryption

In order for the data consumer who has downloaded the data asset ciphertext from the ICARUS platform to decrypt its contents, he/she needs to request from the data provider the decryption key for the specific data asset cyphertext. The data provider again validates that there is an active data contract with the data consumer for the specific data asset and then needs to proceed to securely share the symmetric key (for the data asset decryption) with him/her. In order to do so, though, a secure SSL-enabled connection / tunnel needs to be established with the data consumer, that encrypts all information exchanged and provides privacy and data integrity for the communication between the provider and consumer. In the end, the data consumer can use the decryption key to decrypt the data asset and properly access the underlying data.

It needs to be added that, for provenance purposes, the ICARUS platform tracks each time a data consumer attempts to access a ciphertext (from the core platform itself) or a decryption key (from the data provider).

2.3.3 Data Anonymization

As explained in D2.1, data anonymisation is a critical process, especially in the aviation domain, where sensitive data are included in the datasets that contain not only personal information but also confidential, business and / or private information regarding aircraft and airports management.

The ICARUS Anonymisation method aims at addressing the problem of data privacy protection by providing the customisable process that can be appropriately configured depending on the nature of the data to be anonymized, as well as on the privacy threat that needs to be properly eliminated. Hence, the ICARUS Anonymisation method supports a generic-enough anonymization workflow to cover the complete spectrum of aviation related data across a wide range of aviation-related data analysis cases, including -but not limited to- all demonstrators' needs in this aspect. As described also in D2.1, the ICARUS Anonymisation method is an adaptation of the workflow defined by the research team that is behind the ARX anonymization tool (Prasser & Kohlmayer, 2015).

The data provider, assumed to be the data owner, holds a key role in this method as he/she is the only one that both deeply comprehends the privacy concerns and vulnerabilities of the data and is accountable for selecting the appropriate parameters for anonymisation workflow to be followed prior to making a dataset available to other ICARUS stakeholders. The method provides the means to the data provider to configure and execute the tailored to his/her needs anonymisation process.

Thus, the actor in all steps of the anonymization process is expected to: (a) be a member of the organization that holds the role of the data provider, and (b) have strong data analysis background and deep understanding of the anonymization complexities and legal implications. The ICARUS Anonymisation method supports the data providers through an anonymization workflow aiming to ensure that no sensitive data can leave their premises as-is. However, this method does not aim to enforce usage of specific models and cannot provide any assertions regarding the information disclosure risks that may be caused by improperly anonymized data. Therefore, the responsibility of ensuring that sensitive information is not compromised lies, as expected, with the data provider.

Hence, in the ICARUS perspective, the process of Data Anonymisation includes the following steps:

Step 1: Define the attribute types

The first step after selecting the data asset to be anonymized is to define the attribute type of all the fields included in the data asset that will used in the anonymisation process. As it was described also in D2.1, there are four types of possible specifications for variables (attributes) in respect to privacy issues:

- *Insensitive variables*, which can be kept unmodified.
- *Identifying variables*, which are variables that must be removed from the data set as they pose a high risk of re-identification.
- *Quasi-identifying (QID) variables*, which are variables that can be used directly for re-identification, but they may in combination be used for linkage. These variables must be transformed, as it is assumed that they cannot be simply removed from the dataset as they may be required for analysis.
- *Sensitive variables* (or sensitive attributes (SA)), which can be kept as-is, but they can be protected using privacy models, such as T-closeness or L-diversity.

The data provider must set the fields of the data asset to one of the above specifications based on his knowledge about the data nature, as well as the content of the data asset. This categorisation of the fields of the data assets serves as the basis for the next step in which the privacy models will be configured.

Step 2: Selecting and configuring the Privacy Models

The following step after defining the attribute types of the fields of the data asset is to select and configure the privacy model(s) that will be used in the process. The privacy model is utilised in order to identity the privacy threat that needs to be eliminated. The ICARUS Anonymisation method supports a variety of privacy models that can be selected and utilised by the data provider depending on the nature, the structure and the actual content of the dataset. The privacy models can be grouped into:

- Syntactic privacy models that include the k-Anonymity, the ℓ-Diversity, t-Closeness, δ-Disclosure, β-Likeness and δ-Presence models
- Statistical privacy models that include the k-Map, Average Risk, Population Uniqueness and Sample Uniqueness models
- Semantic privacy models that include the Profitability and Differential Privacy models

The data provider is allowed to select the appropriate privacy model and different approaches for each of the included fields, as well as to customise how specific models are applied through the corresponding parameters of each model (if any).

Step 3: Selecting and configuring the Transformation Models

In this step, the transformation models that will be utilised in order to eliminate the privacy threat that were identified in the previous step are selected. A transformation model can be configured on a field-level for each field of the data asset that is marked as *Quasi-identifying variable.* As explained above, these fields are the ones that will be transformed, addressing also in this way the fields that are marked as sensitive variables. As the various transformation models can be customised and configurated through a list of parameters per transformation model, the resulting transformation varies. The ICARUS Anonymisation method supports a variety of transformation models, such as:

- Value Generalization
- Random Sampling
- Record, Attribute and Cell Suppression
- Microaggregation
- Top Top-and-Bottom-Coding
- Categorisation

As with the privacy models, the data provider is allowed to select the appropriate transformation model and set the corresponding parameters based on his knowledge and expertise.

Step 4 (optional): Assess re-identification risks

In the last step, once the data anonymization models have been applied, the data provider can optionally explore the outcome in terms of the privacy threat risks. The methods employed to assess the usefulness of the transformed (i.e. anonymized) data are highly data- and domain-dependent and involve various statistical comparisons between the input and output data. The risk assessment involves identifying the risks related to the quasi-identifiers mentioned earlier, as well as sample-based and population-based risk estimates and the data provider can obtain the results of the assessment in order to ensure the desired level of privacy risks has been achieved.

3 Data Value Enrichment Methods

3.1 ICARUS Data Analytics Methods

3.1.1 Introduction

This section is dedicated to the latest updates in the ICARUS data analytics methods. It can be considered as a follow-up to Section 3 of "D2.2 – Intuitive Analytics Algorithms and Data Policy Framework", where a large number of methods and algorithms have been already introduced, along with an initial plan of how the ICARUS data analytics approach can be applied to the project's demonstrators.

Table 3-1: ICARUS Data Analytics methods and algorithms

No	Algorithm Name	Algorithmic Family/Type
Axes I: Basic Analytics		
1	Summary Statistics (mean, std, etc.)	Statistical Analysis
2	Hypothesis Testing	Statistical Analysis
3	Sampling	Statistical Analysis
4	Pearson's and Spearman's Correlation	Feature Correlation
5	Linear Regression methods	Feature Correlation, Regression Analysis
6	Logistic Regression	Feature Correlation, Regression Analysis, Classification
7	Principal component analysis (PCA)	Dimensionality reduction, Feature extraction
8	Feature Selection	Dimensionality reduction
9	Autoregressive Integrated Moving Average (ARIMA)	Time series prediction
Axes II: Machine Learning Algorithms		
1	Self-Organising Map (SOM)	Clustering, Dimensionality Reduction
2	K-means	Clustering, Anomaly Detection
3	Streaming K-means	Clustering, Anomaly Detection
4	DBSCAN	Clustering, Anomaly Detection
5	Gaussian Mixture models	Clustering
6	Apriori	Association Rules
7	Collaborative Filtering (CF)	Recommendation Systems
8	Content-based Filtering (CBF)	Recommendation Systems
9	Support Vector Machines (SVM)	Classification, Regression, Outlier detection
10	Classification and Regression Tree (CART)	Classification, Regression
11	Random Forest (RF)	Classification, Regression, Outlier detection

No	Algorithm Name	Algorithmic Family/Type
12	Gradient Boosting Machines (GBM)	Classification, Regression
13	K-NN	Classification, Regression, Outlier detection
14	Naïve Bayes (NB)	Classification
15	Multi-Layer Perceptron (MLP)	Classification, Regression, Time Series Prediction
16	Adaptive Neuro-Fuzzy Inference System (ANFIS)	Classification, Regression, Time Series Prediction
17	Genetic Algorithms (GA)	Optimisation
Axes III: Deep Learning		
1	Deep Feedforward Networks (DFFN)	Classification, Regression, Deep Learning
2	Convolutional Neural Networks (CNN)	Classification, Regression, Deep Learning
3	Recurrent Neural Networks (RNN)	Classification, Regression, Time Series Prediction, Deep Learning
4	Deep Autoencoders	Dimensionality Reduction, Clustering, Data visualisation, Feature Learning, Deep Learning
5	Deep Q-Networks (DQN)	Reinforcement Learning, Deep Learning

The information provided in the following paragraphs is the outcome of the evaluation of the methods and algorithms up until this point, combined with the refinement of the demonstrators' requirements and perspectives, and the currently supported features given by the technical implementation progress of the ICARUS platform. Since the experimentation with the pilots' data is planned for the months to come and the implementation is still on-going, further improvements and updates can be expected for the ICARUS data analytics until the end of the project that will be documented in the demonstrator result reports.

Regarding the overall analytics approach, a minor change needs to be noted. As discussed in Section 3 of D2.2, the presented methods are divided into three axes:

- Axis I: Basic Analytics. This axis includes a number of diagnostic algorithms and statistical methods, useful to extract insights from data that help the analyst understand the underlying behaviour and foresee possible patterns.
- Axis II: Machine Learning Algorithms. This group contains the most widely accepted techniques from the field of machine learning, such as decision trees, support vector machines or random forest. These algorithms can be employed for descriptive, predictive and prescriptive analysis.
- Axis III: Deep Learning. The deep learning subset consists of advanced neural networks algorithms, such as convolutional or recurrent neural networks, especially designed to efficiently process big data by using multiple ("deep") internal processing layers. These networks are considered to be the next evolution of machine learning.

In this document, a fourth axis is introduced, named "Visual Analytics", which belongs exclusively to the descriptive analytics framework and aims to offer tangible insights through visuals. No algorithms are required for this process, as the outcome is usually a graphical representation of computations on selected data features and their relations. Nevertheless, the conclusions derived from this representation, can be of critical value to both the business user and the analyst, especially when the dataset at hand is characterized by high dimensionality and large volume. Thus, the visual analytics will be included in the lines to follow as an "algorithmic family" for the purposes of the present documentation.

3.1.2 Supported Analytics Workflows, Algorithms and Visualizations

One of the main goals of ICARUS is to support the analytics algorithms workflow design and execution, by making the best known and widely accepted algorithms in the aviation industry available, so as to allow all aviation-related stakeholders to analyse and visualize results downstream of big data applications and generate new knowledge and insights.

As explained in the deliverable D2.2, data analytics algorithms can be broken down into three key areas: descriptive analytics, predictive analytics and prescriptive analytics. If someone wants to know what happened, descriptive analytics are put into use. Descriptive analytics can be especially helpful in tracking trends to help plan for the future, handling data from multiple data sources to give valuable insights into the past. However, these findings simply warn that something is wrong or right, without explaining why. Predictive analytics tells what is likely to happen. It uses the findings of descriptive analytics to detect tendencies, clusters and exceptions, and to predict future trends, which makes it a valuable tool for forecasting. Finally, the purpose of prescriptive analytics is to literally prescribe what action to take to eliminate a future problem or take full advantage of a promising trend.

In order to support the ICARUS users in all these three kinds of analytics, a comprehensive set of algorithms were selected on the basis of the following criteria: *(i)* meeting the ICARUS platform requirements and be applicable to aviation specific tasks, *(ii)* proven ability and robustness in the research community through the years, and *(iii)* implementation in a commonly used software framework or library.

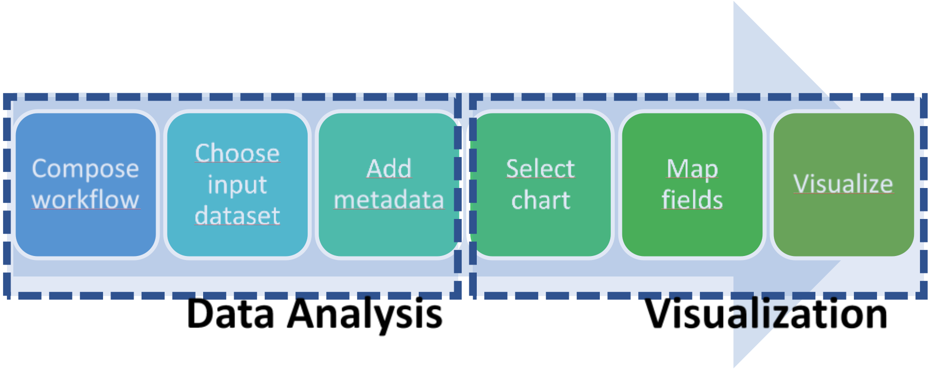


Figure 3-1: Analytics algorithms workflow design process

The analytics algorithms workflow design and execution form a procedure which involves several stages, from the creation of a workflow until the visualisation of the results (Figure 3-1).

Step 1: Workflow composition

The Workflow composition step consists of the design of a graph to automate the process involved in data analysis. Each node in the graph is a task, and edges define dependencies among the tasks. In general, a task is a basic algorithm. The list of available algorithms is organised in three categories according to the classification presented in the deliverable D2.2 and explained in section 3.1.1: basic analytics, machine learning, deep learning to make the browsing easier for users.

Step 2: Choice of data set for the workflow execution

The aim of this step is to connect the workflow to an input dataset so that it can be executed as a data pipeline: starting from the selected data set as an input, a task does its job and generates a target as a result, a second task takes the target file in input, performs some operations and outputs a second target file and so on.

Step 3: Metadata setting

Metadata setting is useful to save the couple (workflow, dataset) as an application, so that it can be stored, reused, updated, run, or shared to the other users by means of proper sharing policies. This step involves the setting of a set of information on the application, such as privacy level, target industry, payment method etc. in accordance with the ICARUS metadata schema (defined in Annex II).

Step 4: Chart selection

The main objective of the chart selection step is to represent knowledge more intuitively and effectively by using different graphs. To convey information easily by providing knowledge hidden in the complex and large-scale data sets, both the aesthetic form and the functionality are necessary.

So, besides displaying an extract of the resulting data obtained upon the workflow of algorithms involved in the application has completed its execution, it is very useful to visualize results by choosing a graph in the form of charts, tables, line graphs, column charts, and many other forms.

Step 5: Fields mapping

In this step, the user selects the dataset fields to be managed for the output visualization. The tuning of chart properties allows to get the final chart result in accordance with the preferences of the user.

Step 6: Visualization

The ultimate step is the visual representation of the results of data analysis, in the form of charts, tables, line graphs, column charts, and many other forms so as to have a comprehensive picture of the produced insights and predictions, making also explicit the trends and patterns inherent in the data. Such a visualization may be also stored in the application in order to facilitate quick access for the user.

3.1.3 Perspectives for Demonstrator 1: AIA

As described in the ICARUS Deliverable D5.2, the scenario to be explored by the Athens International Airport (AIA) demonstrator is entitled “Airport Capacity Planning” and tackles a very interesting, but complex, challenge in the field of aviation. Afterall, the optimisation of capacity planning for an airport is a multi-faceted procedure and involves not only the airport’s operations and infrastructure, but also several aviation stakeholders, such as airlines, ground handler companies and the air traffic control. Therefore, the data analytics results of this demonstrator can be beneficial not only towards the Athens Airport, but also to other, closely-related and cooperating, business actors in the aviation data value chain.

Current Data Analytics Practices in AIA

The Athens International Airport supports the collection of data from various sources into its own Data Warehouse and utilizes a number of Business Software tools for data management, analysis and reporting. For data integration, AIA uses the Informatica tool that is based on an ETL architecture. For Business Intelligence and data insights, AIA operates the SAP BusinessObjects suite. However, only few "power-users" have the ability to generate diagrams and custom reports. The majority of the users have access to these reports only for viewing purposes or to export them in the form of a PDF file.

The Management Information System (MIS) on the same Data Warehouse is integrated with the airport's ERP, the Billing system, as well as the Environmental and Operations Systems. MIS also supports one of the most critical tools for the airport's operations, the Universal Flight Information System (UFIS) which records all flight activities and operations.

Data Analytics to be applied in the AIA Demonstrator Scenario in ICARUS

For a more thorough examination and better results, the Airport Capacity problem is divided into four different, but interrelated, tasks related to data analysis: Capacity Modelling, Airport Traffic Forecasting, Flight Delay Prediction, and Position and Slot Allocation / Scheduling. In the following tables, each of these sub-scenarios is described, along with the proposed data analytics methods that accompany it.

Table 3-2: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.1

<table>
<tr><th colspan="6">Sub-scenario 1.1: Capacity Modelling</th></tr>
<tr><th colspan="6">Description</th></tr>
<tr><td colspan="6">The Capacity Modelling task aims to extract useful knowledge and insights from historical data regarding all possible aspects of capacity planning to the extent available. It will work as an entry point for all the following tasks (sub-scenario 1.2, 1.3 & 1.4), with a descriptive analysis performed on airport traffic, flight delays, slot and position allocation, among others, as well as their relations to external data, such as weather and economic, publicly available, data.</td></tr>
<tr><th>Algorithm Types/ Families</th><th>Purpose & expected output</th><th>Algorithms</th><th>Considered Alternatives</th><th>Input</th><th>Limitations/Caveats/ Considerations</th></tr>
<tr><td><b>Basic Analytics</b></td><td>Statistical view</td><td>Summary statistics</td><td>-</td><td>Flight data from airport</td><td></td></tr>
</table>

				operations[AIA_01, AIA_11]	
Visual Analytics	To provide descriptive analysis and initial insights. More specifically, to examine relationships among input features, as well as between different data sources	-	-	Flight data from airport operations[AIA_01, AIA_11], Weather data [AIA_DR_43], Flight schedules [AIA_07], Economic data	Domain knowledge is required to evaluate the visualisations.
Clustering	To provide descriptive analysis and initial insights. More specifically, to examine relationships among input features based on similarity.	SOM with k-means	DBSCAN	Flight data from airport operations [AIA_01, AIA_11], Weather data [AIA_DR_43], Flight schedules [AIA_07], Economic data	Categorical data need to be converted to numerical. Domain knowledge is required to interpret the results

Table 3-3: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.2

<table>
<tr><th colspan="6">Sub-scenario 1.2: Flight Delay Prediction</th></tr>
<tr><th colspan="6">Description</th></tr>
<tr><td colspan="6">The flight delay prediction is a very common subject in aviation-related research literature, since any such disruption to a flight may produce a significant negative impact on airlines, as well as airports. The core target of this task is to predict whether a future given flight will be delayed or not, either as True or False, which is a classification problem, or as an exact amount of time (e.g. in minutes), which is a regression problem. To make these predictions, a machine learning algorithm has to be trained with historical data that contain the appropriate information.</td></tr>
<tr><th>Algorithm Types/ Families</th><th>Purpose & expected output</th><th>Algorithm(s)</th><th>Considered Alternatives</th><th>Input</th><th>Limitations/Caveats/ Considerations</th></tr>
<tr><td><b>Classification</b></td><td>Classify flights of subsequent days into distinct classes based on their probable delay</td><td>Random Forest</td><td>SVM, Naïve Bayes, Decision tree, k-NN, MLP</td><td>Flight data from airport operations [AIA_01,AIA_09, AIA_11], Weather data [AIA_DR_43], Flight schedules[AIA_07]</td><td>Categorical data need to be converted to numerical.
Time-related variables need to use the same format and time zone (UTC).</td></tr>
<tr><td><b>Regression</b></td><td>Predict flight delay (in exact mins) based on next days'</td><td>SVM (rbf kernel)</td><td>Decision Trees, Random Forest</td><td>Flight data from airport operations[AIA_01,AIA_09,</td><td>Categorical data need to be converted to numerical.</td></tr>
</table>

	schedule and weather			AIA_11], Weather data [AIA_DR_43], Flight schedules[AIA_07]	Time-related variables need to use the same format and time zone (UTC).
Time Series prediction	Predict mean daily delay in airport based on previous days' data	RNN (LSTM)	ARIMA	Flight data from airport operations[AIA_01,AIA_09, AIA_11], Weather data [AIA_DR_43],	Categorical data need to be converted to numerical. Time-related variables need to use the same format and time zone (UTC).

Table 3-4: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.3

Sub-scenario 1.3: Airport Traffic Forecasting					
Description					
The main goal of this task is to forecast the incoming and outgoing traffic of the airport in a daily or hour-by-hour basis, in order to perform better planning of ground handling and operation services. The traffic forecasting should consider multiple external sources, such as weather forecasts, national economic factors, or ground-handling equipment availability.					
Algorithm Types/ Families	**Purpose & expected output**	**Algorithms**	**Considered Alternatives**	**Input**	**Limitations/Caveats/ Considerations**
Regression	Estimate number of incoming and outgoing flights per hour based on related airport and ground handling operations, seasonality, and weather conditions	SVM	Decision Trees, Random Forest	Flight data from airport operations[AIA_01, AIA_11], Weather data [AIA_DR_43]	Categorical data need to be converted to numerical. Time-related variables need to use the same format and time zone (UTC).
Time series prediction	Predict daily number of incoming and outgoing flights based on previous days traffic, weather and economic data.	RNN (LSTM)	ARIMA	Flight data from airport operations[AIA_01, AIA_11], Weather data [AIA_DR_43], Economic data.	Categorical data need to be converted to numerical. Time-related variables need to use the same format and time zone (UTC).

Table 3-5: ICARUS Data Analytics for the AIA Demonstrator Scenario – Sub-scenario 1.4

Sub-scenario 1.4: Position and Slot Allocation/Scheduling
Description

The position and slot allocation use case is, by nature, an optimisation problem. This means that out of a finite number of suggested solutions, the system should be directed to the optimal one with the best performance and cost-effectiveness. The optimisation algorithm should take into account how the positions (gates) are best allocated during a day and try to model the reasoning behind it.

Algorithm Types/ Families	Purpose & expected output	Algorithms	Considered Alternatives	Input	Limitations/Caveats/ Considerations
Optimisation	Optimal allocation of positions and time slots to aircrafts and airlines.	Genetic algorithms	-	Flight data from airport operations[AIA_01, AIA_11], Weather data [AIA_DR_43], Flight schedules[AIA_07], Airport infrastructure [AIA_10], Aircraft size[AIA_01]	The reasons behind a gate or slot change are usually not available. Domain knowledge is required to evaluate the results.

In the following months, the plan is to gradually explore all the aforementioned tasks by testing and optimizing algorithms based on the available data. The most suitable input data and input features will be selected and the most appropriate algorithms will be put to test. The feedback coming from the AIA domain experts is expected to help in this direction and improve the results, as well as the scope of the main data analytics scenario.

3.1.4 Perspectives for Demonstrator 2: PACE

Current Data Analytics Practices in PACE

PACE supplies numerous customers across different segments of the aerospace and aviation industry with innovative software products, tailor-made solutions, specialized services and new approaches for helping key departments and project teams to overcome their business challenges. Listening to such a highly diverse customer base for more than a decade led to a portfolio of Pacelab commercial off-the-shelf software products. Examples includes applications for the preliminary design of aircraft and systems, the modelling of aircraft cabins and configurations and the in-flight optimization of trajectories.

Thereof, the Pacelab Mission Suite is the one-stop solution for route and aircraft economic analysis and the only software of the Pacelab portfolio that is in scope of the ICARUS project. To date, the Pacelab Mission Suite supports only a limited number of data analytics tools or libraries. Below is an overview on the related tools and libraries.

Boeing PC WindTemp and Airport Temperatures

The Pacelab Mission Suite supports the use of statistical airport surface temperature data and statistical data of vertical wind and temperature profiles provided by Boeing Airport Temperatures and Boeing PC WindTemp in route definitions. The Boeing PC WindTemp is an external library that runs statistical analysis on wind and temperature data files based on a period, a probability and some flight specific data. The Boeing Airport Temperatures is an ASCII file lists airport temperature values based on a fixed pattern of period and probability. The Pacelab Mission Suite interfaces the Boeing PC WindTemp and Airport Temperatures at runtime.

Nevron Chart for .NET

The Pacelab Mission Suite supports Nevron Chart as a solution to business and scientific charting needs including a large set of one- and multi-dimensional charting types like bar and column, radar, mesh surface or heat map.

Xceed DataGrid

The Pacelab Mission Suite supports Exceed DataGrid as a solution for a table based asynchronous data virtualization and modern smooth scrolling mechanics including but not being limited to in-place editing, data grouping, input validation, data binding, search algorithms and Excel-like capabilities.

Data Analytics to be applied in the PACE Demonstrator Scenarios in ICARUS

In the scope of the ICARUS project, the Pacelab Mission Suite is to be considered as an aviation data provider and aviation data and service asset consumer. Two scenarios exist that expose a reasonable linkage between the Pacelab Mission Suite and the ICARUS platform, as follows:

- Scenario 1: Pollution Data Analysis
- Scenario 2: Massive Route Network Analysis and Evaluation

Scenario 1 comprises a set of activities aiming to support a more accurate analysis of pollution data and aircraft emissions. Typical use cases in this field involve the modelling of pollution data and the prediction of aircraft performance in relation to the environmental impact.

Scenario 2 comprises a set of activities aiming to analyse pollution data on a larger scale, that of a massive route network. Typical use case examples in this field involve the statistical evaluation of weather data, the modelling of aircraft payload capacity scenarios and the prediction of aircraft performance in relation to the underlying route network.

The motive behind both scenarios is to cut operational expenses while reducing environmental impact, which is one the biggest challenges that the aviation domain will face in next decade. The ability to provide more accurate pollution predictions can have great benefit and be of interest to many stakeholders. Results of such analyses can be shared with public authorities for calculating and checking CO2 emissions, with municipalities interested into allowing a better planning of approach paths and selection of the most suitable local airport under consideration of local pollution constraints, as well as airlines wishing to enhance their flight planning by complying with regulations and achieving fuel savings and safety enhancements. The following tables present the proposed analytics methods for each of the two scenarios.

Table 3-6: ICARUS Data Analytics for the PACE Demonstrator Scenario 1

Scenario 1: Pollution Data Analysis					
Description					
This scenario aims to support an accurate analysis of pollution data and aircraft emissions per flight legs, on the basis of actual flight routes. The goal is to predict the carbon emissions and fuel burn on a per-leg level of a specific flight, given the flight route information and the average weather conditions on that flight.					
Algorithm Types/ Families	**Purpose & expected output**	**Algorithms**	**Considered Alternatives**	**Input**	**Limitations/Caveats/ Considerations**
Basic Analytics	Perform some pre-processing on the original data in order to be ready to be used by some algorithms	One-Hot-Encoding	-	Aircraft performance data [PACE_02] Weather data [PACE_04, PACE_DR_02, PACE_DR_03 (if available)] Historical flight paths [PACE_DR_04, PACE_DR_05] Passenger load data [PACE_08] Historical taxi in/out times [PACE_DR_06] Historical Operation Costs (if available) [PACE_DR_07] Historical flight schedule data [PACE_DR_08]	-
Regression	Predict carbon emissions and fuel burn per flight leg	Gradient Boosted Trees	Linear Regression Decision Trees Random Forest	Aircraft performance data [PACE_02] Weather data [PACE_04, PACE_DR_02, PACE_DR_03 (if available)] Historical flight paths [PACE_DR_04, PACE_DR_05] Passenger load data [PACE_08]	Combining all the data needed in order to form a well-defined representation model ready to be used by the algorithms. Converting categorical features to numeric, using one-hot-encoding

				Historical taxi in/out times [PACE_DR_06] Historical Operation Costs (if available) [PACE_DR_07] Historical flight schedule data [PACE_DR_08]	
Dimensionality Reduction	Investigate the evaluate the trade-offs of using lower-dimensional datasets compared to the original ones	Principal Component Analysis (PCA)	Uniform Manifold Approximation and Projection (UMAP)	Aircraft performance data [PACE_02] Weather data [PACE_04, PACE_DR_02, PACE_DR_03 (if available)] Historical flight paths [PACE_DR_04, PACE_DR_05] Passenger load data [PACE_08] Historical taxi in/out times [PACE_DR_06] Historical Operation Costs (if available) [PACE_DR_07] Historical flight schedule data [PACE_DR_08]	Transforming the original datasets in way to maintain at least 85% of the original dataset variance.

Table 3-7: Table 3 2: ICARUS Data Analytics for the PACE Demonstrator Scenario 2

<table>
<tr><th colspan="6">Scenario 2: Massive Route Network Analysis and Evaluation</th></tr>
<tr><th colspan="6">Description</th></tr>
<tr><td colspan="6">This scenario can be considered as an extension and more advanced use-case of the first scenario. The goal here is again to predict carbon emissions and fuel burn compared to the aircraft performance but in a larger scale, that of a massive route network. That being said, weather and performance at specific times and geolocations (latitude, longitude, altitude) will be taken into account in order to form the underlying network, compared to the first scenario where only samples and averages are meant to be used.</td></tr>
<tr><th>Algorithm Types/ Families</th><th>Purpose & expected output</th><th>Algorithms</th><th>Considered Alternatives</th><th>Input</th><th>Limitations/Caveats/ Considerations</th></tr>
</table>

Basic Analytics	Perform some pre-processing on the original data in order to be ready to be used by some algorithms	One-Hot-Encoding Summary Statistics (min, max, avg, median)	-	Aircraft performance data [PACE_02], En-route weather data [PACE_05], Historical taxi in/out times for various airports [PACE_06], Historical operational costs for various routes [PACE_07], Historical flights paths [PACE_DR_04, PACE_DR_05], Historical airport weather data at ground level [PACE_DR_02], National and international airport specification data [PACE_DR_01]	-
Regression	Predict carbon emissions and fuel burn given the underlying route network	Gradient Boosted Trees	Linear Regression Decision Trees Random Forest	Aircraft performance data [PACE_02], En-route weather data [PACE_05], Historical taxi in/out times for various airports [PACE_06], Historical operational costs for various routes [PACE_07], Historical flights paths [PACE_DR_04, PACE_DR_05], Historical airport weather data	Combining all the data needed in order to form a well-defined representation model ready to be used by the algorithms. Converting categorical features to numeric, using one-hot-encoding. Find a suitable representation of the underlying network in a efficient format.

				at ground level [PACE_DR_02]	
Deep Learning	Predict carbon emissions and fuel burn given the underlying route network in a more advanced manner that can model non-linear and complex relationships	Long Short-Term Memory	Graph Convolutional Networks (GCN)	Aircraft performance data [PACE_02], En-route weather data [PACE_05], Historical taxi in/out times for various airports [PACE_06], Historical operational costs for various routes [PACE_07], Historical flights paths [PACE_DR_04, PACE_DR_05], Historical airport weather data at ground level [PACE_DR_02]	Modeling the available data into a deep neural network structure. Hyperparameter tuning of the deep neural network. Computational Intensiveness.

In the next steps, the described data analytics methods will be further explored and tested with the available data in order to extract useful knowledge that will be exploited in order to select the most suitable data analytics methods that will be utilised within the context of this demonstrator. In this process, the PACE domain experts will be also involved in the analysis and will provide useful feedback towards the formulation of the final solution. It should be noted that as the implementation of the demonstrator evolves, the list of data analytics methods are bound to change based on the results of the analysis and additional methods might be considered.

3.1.5 Perspectives for Demonstrator 3: ISI

Current Data Analytics Practices in ISI

The computational epidemiology framework exploits various standard techniques to perform the simulations and analyze the results. The GLEAM numerical engine uses multinomial distributions for generating random deviates to simulate long distance travels by flight and to perform the epidemic dynamics worldwide; such multinomial procedure currently employs a pseudorandom number generator belonging to the family of the so-called Well Equidistributed Long-period Linear (WELL) generators. Moreover, to take into account the daily commuting patterns among subpopulations, needed for computing the force-of-infection, ISI applies a method consisting in a time-scale separation of the commuting flows with respect to the characteristic timescales of the other processes (air travels, disease dynamics, ...) in the simulation.

Since the main implementation language of the GLEAM engine is Python (due to its flexibility and high-level data structures), there is extensive use of Numpy, a package for scientific computing that provides efficient multidimensional array objects and associated powerful functions to manipulate them. Being the most time-critical parts of the engine implemented in Fortran for efficiency reasons, a connection between these two languages is provided by the tool f2py, the Fortran to Python interface generator. Finally, the output of the simulations is stored by means of the h5py package, which represents a convenient interface to the HDF5 binary data format.

Data Analytics to be applied in the ISI Demonstrator Scenarios in ICARUS

It is important to note that the epidemic simulation engine runs outside the ICARUS platform. This demonstrator, due to its peculiar approach focused on scientific research and the fact that we are working with health data (that are not to be included in the ICARUS common aviation model), cannot exploit the platform analytics powers.

The introduction of age stratification into the model requires a major redesign of two fundamental parts of the numerical engine, namely the way air travels are implemented and the computation of transition probabilities for the epidemic dynamics. For the flights, in order to obtain the correct travel probabilities, ISI needs to take into account the demographic structure of the various subpopulations, with the further constraint represented by a no-drift effect for the number of individuals of each age group over time. Concerning the disease epidemics, the procedure yielding the probabilities corresponding to transitions induced by effective contacts with infective compartments becomes more involved, since now we must consider the different interactions among the various age groups. This is accomplished by using suitable contact matrices obtained from synthetic populations based on real-world data about the demographic structure of different regions. Furthermore, simple multiplication operations are replaced by multidimensional dot products between contact matrices and vectorial representations of the age-stratified individuals belonging to each compartment, which implies that special care is required in order to implement such operations in a computational efficient way.

We will assess the accuracy of the improved modeling framework exploiting age-stratified data by comparing the results provided by the simulations with historical time series of several US and European countries seasonal flu epidemics, using performance indicators such as the logarithmic score (due to its operational use by major national and international health agencies) and the Maximum absolute percentage error (MAPE).

Table 3-8: ICARUS Data Analytics for the ISI Demonstrator Scenario

Algorithms Types / Families	Purpose & expected output	Algorithms	Considered Alternatives	Input	Limitations/ Caveats/ Considerations
Multivariate random generators	Provides means for performing stochastic simulations leveraging on real-world data	Multinomial distribution	N/A	Worldwide air travel origin-destination bookings	-
Performance indicators	Represent measures to assess the accuracy of a forecasting method	Logarithmic score; Maximum Absolute Percentage Error	N/A	Output results from simulations and historical time series of seasonal flu epidemics	-

3.1.6 Perspectives for Demonstrator 4: CELLOCK

Current Data Analytics Practices in CELLOCK

CELLOCK provides a complete inventory and warehouse management system named **BoB** (**Buy-on-Board),** used by both caterers and airlines. Caterers use it for the warehouse management, monitoring all their product stock and handling all plane loadings, including Food and Beverages (FnB) trays as well as Duty-Free/Sales-on-Board trays. Airlines use BoB for inventory and sales monitoring.

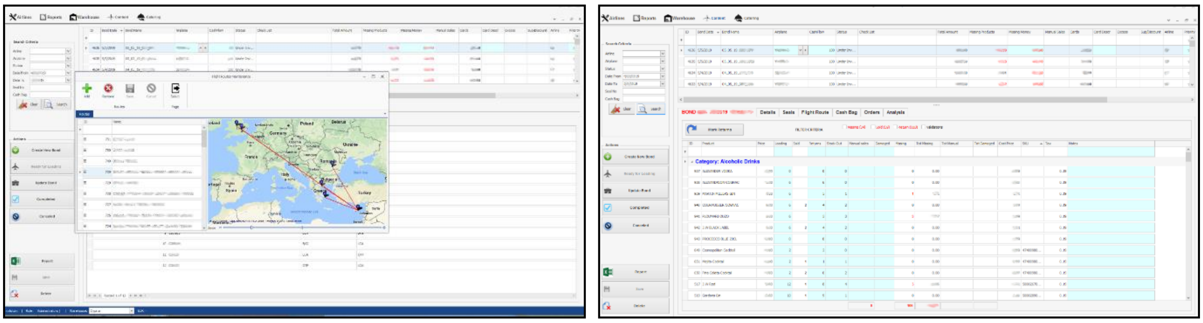


Figure 3-2: CELLOCK's BoB: bond loading page & sales page

BoB supports a wide number of pre-built reports, while the end-users are also able to create customized reports to accommodate their specific needs. The users of BoB can also get extensive visualized results, based on the data collected. This provides a quick overview on the predefined reports, as well as the customized ones, such as sales achieved by flight or destination.

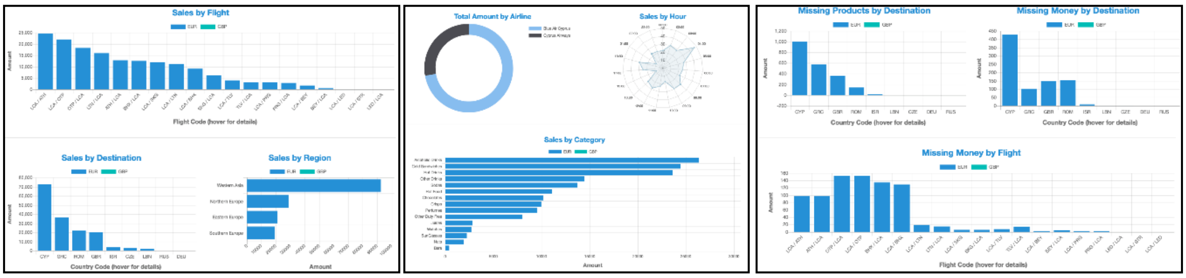


Figure 3-3: CELLOCK's BoB: sales visualization pages & discrepancies visualization graphs

Despite the comprehensive dashboard with visualized results, and the option to use or create customized reports, BoB lacks any advanced predictive analytics.

Data Analytics to be applied in the CELLOCK Demonstrator Scenarios in ICARUS

The scenarios to be executed by CELLOCK focus on daily complicated operations of both airlines and in-flight product providers/caterers, and more specifically on the Food and Beverage (FnB) and Duty-free trays loading (1st scenario) and on automated product discount and offer suggestions (2nd scenario). The overall target is to increase ancillary revenue for airlines, reducing cabin waste and tray loadings, while increasing sales and automate product discount suggestions.

The two planned scenarios are examined in the tables below, accompanied by the proposed analytic methods and procedures.

Table 3-9: ICARUS Data Analytics for the CELLOCK Demonstrator Scenario 1

Scenario 1: Predict in-flight product sales					
Description					
The main objectives of this task are: a) extracting useful knowledge and insights from historical data regarding various different aspects for in-flight sales and b) predicting the exact number of sales (regression problem) per product and product category for each flight. The first part of this task (statistical analysis) will utilize historical data about retail and FnB in-flight sales, number of passengers, airplane loading for FnB, flights discrepancies, as well as other related external data, such as weather data, flight status data, economic data of countries. The second part of this task will utilize the extracted knowledge and insights from the statistical analysis, as well as the aforementioned datasets in order to train a machine learning algorithm for predicting the number of in-flight product sales, with the target of optimizing tray loadings and minimizing in-cabin waste.					
Algorithm Types/ Families	**Purpose & expected output**	**Algorithms**	**Considered Alternatives**	**Input**	**Limitations/Caveats/ Considerations**
Statistical Analysis	Statistical view	Summary statistics	-	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Airplane loading for F & B (CELLOCK_06), Flights discrepancies (CELLOCK_07)	
Visual Analytics	To provide descriptive analysis and initial insights. More specifically, to examine relationships among input features, as well	-	-	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Airplane loading for F & B (CELLOCK_06),	Domain knowledge is required to evaluate the visualisations.

	as between different data sources.			Flights discrepancies (CELLOCK_07), Weather data, Flight status data, Economic data of countries	
Feature Correlation	To provide descriptive analysis and initial insights of pattern relationships. Specifically, in order to identify how different flight characteristics (e.g. seasonality, departure country, etc.) affect the sales of products or the overall sales of each product category.	Pearson's and Spearman's Correlation	Linear Regression	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Airplane loading for F & B (CELLOCK_06), Flights discrepancies (CELLOCK_07), Weather data, Flight status data, Economic data of countries	Categorical data need to be converted to numerical.
Clustering	To provide descriptive analysis and initial insights of pattern relationships. Specifically, in order to identify groups of flights that show similar sales of products or similar sales for each product category.	K-means	DBSCAN	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Airplane loading for F & B (CELLOCK_06), Flights discrepancies (CELLOCK_07), Weather data, Flight status data, Economic data of countries	Categorical data need to be converted to numerical.
Regression	Predict the sales of each product per flight based on historical in-flight sales, historical flight data (e.g. departure, destination, etc.), weather	Gradient Boosting Tree	Random Forest, Deep Feedforward Networks	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Airplane loading for F & B (CELLOCK_06),	Categorical data need to be converted to numerical.

	conditions (temperature), seasonality and economic data.			Flights discrepancies (CELLOCK_07), Weather data, Flight status data, Economic data of countries	
Regression	Predict the total sales for each product category per flight based on historical in-flight sales, historical flight data (e.g. departure, destination, etc.), weather conditions (temperature), seasonality and economic data.	Gradient Boosting Tree	Random Forest, Deep Feedforward Networks	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Airplane loading for F & B (CELLOCK_06), Flights discrepancies (CELLOCK_07), Weather data, Flight status data, Economic data of countries	Categorical data need to be converted to numerical.

Table 3-10: ICARUS Data Analytics for the CELLOCK Demonstrator Scenario 2

Scenario 2: Predict profitable product discounts and offers to increase inflight sales.					
Description					
The main objectives of this task are: a) extracting useful knowledge and insights from historical data regarding various different aspects for in-flight sales of discounted products, b) predict the exact number of sales (regression problem) per discounted product for each flight, c) identify products that could be included together in a bundle (association rule learning) per flight and d) recommend product bundles offers for each flight. The first part of this task (statistical analysis) will utilize historical data about retail and FnB in-flight sales, discounted products, number of passengers, flights discrepancies, as well as other related external data, such as flight status data, economic data of countries. The second part of this task will utilize the extracted knowledge and insights from the statistical analysis, as well as the aforementioned datasets in order to predict and suggest discounts and offers for each flight.					
Algorithm Types/ Families	**Purpose & expected output**	**Algorithms**	**Considered Alternatives**	**Input**	**Limitations/Caveats/ Considerations**
Statistical Analysis	Statistical view	Summary statistics	-	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Flights discrepancies (CELLOCK_07)	
Visual Analytics	To provide descriptive analysis and	-	-	Retail and F&B in-flight sales (CELLOCK_01),	Domain knowledge is required to evaluate the visualisations.

	initial insights. More specifically, to examine relationships among input features, as well as between different data sources.			Number of passengers (CELLOCK_02), Flights discrepancies (CELLOCK_07), Flight status data, Economic data of countries	
Feature Correlation	To provide descriptive analysis and initial insights of pattern relationships. Specifically, in order to identify how different flight characteristics (e.g. seasonality, departure country, etc.) affect the sales of discounted products or bundle of products.	Pearson's and Spearman's Correlation	Linear Regression	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Flights discrepancies (CELLOCK_07), Flight status data, Economic data of countries	Categorical data need to be converted to numerical.
Clustering	To provide descriptive analysis and initial insights of pattern relationships. Specifically, in order to identify groups of flights that show similar sales of discounted products or bundle of products.	K-means	DBSCAN	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Flights discrepancies (CELLOCK_07), Flight status data, Economic data of countries	Categorical data need to be converted to numerical.
Regression	Predict the sales of the discounted products/bundles per flight based on historical in-flight sales, historical flight data (e.g.	Gradient Boosting Tree	Random Forest, Deep Feedforward Networks	Retail and F&B in-flight sales (CELLOCK_01), Number of passengers (CELLOCK_02), Flights discrepancies (CELLOCK_07), Flight status	Categorical data need to be converted to numerical.

	departure, destination, etc.), seasonality and economic data.			data, Economic data of countries	
Association Rules	Identify products (or product categories) associations in order to provide bundles that could be offered per flight based on in-flight sales and flight routes.	Apriori	-	Retail and F&B in-flight sales (CELLOCK_01), Flights discrepancies (CELLOCK_07)	-
Recommendation Systems	Recommend product bundle offers per flight based on in-flight sales and flight routes.	Collaborative Filtering	-	Retail and F&B in-flight sales (CELLOCK_01), Flights discrepancies (CELLOCK_07)	Categorical data need to be converted to numerical.

The plan for the remaining project implementation period is to demonstrate the aforementioned scenarios and optimize the prediction algorithms for tray loadings and discounts. The feedback from the Cellock experts as well as the consortium will provide a guidance for improving the results and the scope of the main data analytics scenario.

3.2 ICARUS Data Policy and Assets Brokerage Framework

The first version of the ICARUS Data Policy and Assets Brokerage Framework was presented in D2.2, where its dual scope was defined as follows:

(I) The framework will formalise all attributes and qualities that affect, or are in any way relevant to, the ways in which data and data-based assets can be shared / traded and handled subsequently to their acquisition, including licenses, IPR, data sensitivity, privacy risks, even data content and structure where relevant.

(II) The framework will enable the creation of structured, machine-processable data and data-based asset contracts for the aviation industry. It will define how contractual terms pertaining to asset trading agreements should be expressed into an appropriate machine-processable format and it will describe how stakeholders shall interact in the context of the foreseen aviation data sharing scenarios and what is the system's expected behaviour.

The above remain valid and the second version of the framework, which will be presented in the next sections, primarily aims to accomplish these targets. The ICARUS Data Policy and Assets Brokerage Framework aspires to go further than a set of definitions and conceptually plausible guidelines and manage to provide a strong theoretical foundation for the construction of a robust aviation data and data-based assets' marketplace. Therefore, it has been further refined to assimilate insights gained through (a) the initial application of the framework, which has been inevitably limited by the current preliminary implementation phase, (b) discussions with aviation stakeholders, some of which have also been documented in the WP1 deliverables, which have been helpful in revealing considerations and scepticism that if not addressed properly could significantly hinder the platform's adoption and, to a lesser extent, (c) identified technical limitations that could make the framework inapplicable in real life scenarios and (d) updates on the marketplaces and data sharing landscape analysis.

It should be mentioned that the way the framework was described in D2.2, the term "Assets" in its title was considered (either explicitly stated or implied) to be limited to the data assets. However, the ICARUS brokerage functionalities extend to the outcomes of data analysis and visualisation and even the actual analysis steps and the underlying algorithms are allowed in this context to be considered as an asset. These assets are not limited to data, but they clearly hold strong links to the underlying aviation data, and this is why the original term "data assets" has been replaced by "data and data-based assets" in the framework's scope statement above. Similar to data, the aforementioned different types of assets are also protected by IPR and fall under the ICARUS Data Policy and Brokerage Framework.

For the remainder of the current section, the term "assets" will thus include all the above and any differentiation and/or limitation enforced by their distinct characteristics, will be explicitly outlined and discussed in subsequent sections.

3.2.1 Aviation Marketplace Insights and ICARUS positioning

D2.2 provided an extended landscape analysis of data marketplaces and data sharing incentives and barriers in general (irrespectively of the industry scope), as well as a state-of-the-art review on data sharing initiatives in aviation. The insights gained were leveraged to guide the positioning of the ICARUS data policy and assets brokerage framework towards the right direction for the aviation industry and were combined with collected requirements from domain stakeholders and discussions among the project's partners to shape the first definition of the framework. The current section will report on the updates of the previous analysis, focusing now more on the marketplace aspects that will help refine the ICARUS brokerage framework, but will also largely determine its success.

The marketplaces that are of interest for ICARUS follow the many-to-many paradigm (many providers and many consumers), which is the most common category. As already explained, the traded commodities are not to be limited to datasets, a decision that causes the number of potentially relevant platforms to explode. However, it is not possible, or even desirable, to identify and exhaustively explore all existing marketplaces. The aim here is to understand which are the dimensions that affect (or should affect) the design of an aviation data-enabled marketplace and therefore a typology is needed to help classify the marketplaces and identify their differentiating features.

(Stahl et al, 2016) identify 12 dimensions along which data marketplaces can be examined, the majority of which are considered as relevant for ICARUS and have been adopted and extended through other classification schemes, serving both academic and market purposes, to construct the list of marketplace dimensions below. It should be noted that, as (Schomm et al, 2013) find, adopting complex marketplace classification schemes with multiple features hinders, perhaps counterintuitively, the extraction of meaningful conclusions, as it is impossible to populate all dimensions for all marketplaces, thus reducing the conciseness of an attempted report. However, the scope of the section is not to provide a market analysis, which has been performed in the context of D7.2, and many of the dimensions are only included to ensure that all important characteristics of the ICARUS marketplace are documented and that any limitations and decisions they impose will be properly reflected in the framework's definition.

Dimension 1: Type of core offering

Platforms that fall under the digital marketplace definition provide commodities of four types:

1. Raw data in any form, for which the platforms only hold dataset listings (catalogues) with links pointing to other sites, i.e. they do not host the data at all, but act as directories to find the datasets in other sites/ platforms. These platforms are out of scope for ICARUS and will not be examined further.
2. Data assets and the underlying mechanisms required to trade them. In most cases the platforms enforce some minimum common underlying schema/model/format/structure on the data being traded in order to enable the brokerage functionalities.
3. Data enrichment and analysis tools that operate on top of the underlying data. In these cases, the corresponding platforms often do not encourage members to download datasets as raw data, but instead provide the tools and services needed to explore and process the data.
4. Data-enabled intelligence services as ready to be consumed reports and/or custom consultant services based on the underlying data. In many of these cases, the marketplace aspect is usually very limited and the platforms primarily sell business intelligence services. They often already have a collection of pre-processed and curated domain-relevant data and aim to attract companies and organisations that seek consulting services and business intelligence reports to be created on the data they will bring combined with the existing.

ICARUS is positioned on the intersection of types 2,3 and 4, as it enables its users to trade data assets (type 2), to combine datasets and perform analysis, create visualisations and select (almost) any part of the process to be considered a tradeable asset (superset of type 4) and it also offers the tools needed to enrich, process and analyse data (type 3).

The ICARUS positioning in this dimension is by itself quite unusual and even bold (with the support of OAG that is excluded from the analysis in the following lines as part of the ICARUS consortium), as it aims to equally support all three types, whereas most aviation-related data marketplaces focus on one type. As an example, RDCaviation[1], which holds data feeds of core aviation information, mostly promote their exploitation through their flagship products: APEX[2], an airline performance analytics platform, and AirportCharges[3] that focuses on airport charges and en route data analysis. The SKYtelligence platform, provided by SmartSky[4] offers an aviation data marketplace, along with tools to leverage them towards creating applications and services, but it is not clear from the publicly available information whether it would be possible to only acquire the data or if consumers are tied to the platform in order to use them. Pure data marketplaces are not common in aviation and platforms that do provide aviation datasets may differ significantly in the way they address the end user's need for data. Cirium[5], the platform through which FlightGlobal provides aviation data feeds (as well as aviation data analytics services), follows a more ad hoc request-for-data process instead of offering dataset search functionalities. AirlineData[6] on the other hand, a US-based air traffic database, follows a different approach and focuses on the data, offering full access to the complete database to all the platform members, who can freely perform queries and access the raw data they need at all times.

This last part, regarding the way platform members can access the provided assets, cannot be seen independently from the core offering, as a marketplace should by definition enable users to discover and explore the assets being traded. The DX Network, one of the largest blockchain-based business data marketplaces, which focuses on data asset trading, offers granular access to the data through its custom query language, but most platforms offer more traditional asset search functionalities. (syncsort, 2019) find that allowing users to pick and choose the information they need depending on what they want to accomplish is very important in data marketplaces, yet the way in which they are structured does not usually allow such fine-grained control.

ICARUS strives to provide a high level of flexibility through its advanced asset query mechanism, designed to facilitate users identify not only the assets that are of interest to them, but also their subsets and/or combinations. The technical details of the mechanism put in place are documented in technical deliverables, whereas the query creation process which is an important part of the brokerage workflow is described in more detail in the corresponding section (Section 3.2.3).

Dimension 2: Marketplace and asset ownership

According to (Stahl et al, 2016), marketplace ownership is a key differentiating factor that shapes the way a platform operates. Specifically, it is important to know whether the platform is owned by one or more of the asset providers, which would make it inevitably seller-biased, by one or more asset consumers, which would similarly make it buyer-biased, or if it is independent, i.e. it can be examined

[1] https://rdcaviation.com/

[2] https://www.rdcapex.com/

[3] https://www.airportcharges.com/

[4] https://www.smartskynetworks.com/skytelligence/

[5] https://www.cirium.com/

[6] https://www.airlinedata.com/

and designed with respect to the actual platform as a brokerage enabler and not based on how to promote the needs of a specific role (provider or consumer). It should be mentioned that it is very common in aviation data-related marketplaces to have one of the data providers acting as a platform provider as well. Skywise[7], by Airbus, is a notable example, since Airbus is a significant data provider contributing data to the platform.

In the case of ICARUS, the platform can be considered independent, i.e. not biased in favour of either providers or consumers, and therefore the adopted brokerage framework should ensure that both interacting parties are equally facilitated throughout the asset brokerage workflow.

Having said that, the proclaimed platform independence should not be mistakenly assumed to ensure that asset providers and asset consumers will trust it. In fact, as explained in (TowardsDataScience, 2018) it is extremely important **to allow participants to not trust the platform**, i.e. to foresee technical measures that enforce clear asset provenance, to guarantee contracts cannot be tampered with, to ensure the asset remains available for the complete contract duration etc, so that the platform trustworthiness is based on concrete and unarguable facts and not on the members' willingness to acknowledge it. These qualities hint towards DLT solutions and are therefore by design ensured in ICARUS, which as stated in D2.2 and in the technical deliverables, builds upon DLT and specifically the Ethereum platform. Another important aspect in allowing participants to not trust the platform is the access it has to the assets. Data ownership in the cloud is a controversial topic with many implications that continues to raise concerns, however the right of providers and consumers to not trust the platform implies that it should not only be illegal for the platform to access the assets, but it should be instead technically impossible. In the case of data assets, this is also ensured in ICARUS through the encryption mechanism put in place.

Dimension 3: Participation & target audience

Participation schemes in many-to-many marketplaces range from open to all to invitation/members-only, depending on the type of assets being traded, the supported transaction mechanisms, the domain and the type of stakeholders. As discussed in D2.2 in the context of key data sharing consideration II regarding trust, the aviation domain has very strong KYC (Know-Your-Customer) requirements, so for industry stakeholders to trust this marketplace, participants should be limited to well-known organisations and businesses operating in the core aviation industry or have explicit and clear connection to it, as defined by the three aviation tiers (Primary Aviation, Extra-Aviation, Aviation-derived) described in D1.1. In order to become a member of the ICARUS ecosystem, an organisation will have to go through a thorough identity validation process. Controlled network membership is also applied in the ICARUS Blockchain to ensure that only verified participants are allowed to publish blocks.

It should be stressed that controlled membership as a domain requirement can be easily verified through a quick visit on the websites of the aviation-targeted data and intelligence platforms. It is also

[7] https://www.airbus.com/aircraft/support-services/skywise.html

often the case that in order for an organisation to become a member, a prior membership to another aviation organisation/ group/ initiative / collaboration is required, e.g. as is the case with the CAPA Data Centre[8] which is only available to CAPA Members.

Dimension 4: Domain and domain models

The domain aspect, as opposed to the participation dimension, does not reflect who is expected to use the platform, but rather what the user can expect to find in it. It shows whether the marketplace is targeted to a specific domain, hence the assets expected to be traded through it are also relevant to (and/or optimised for) the specific domain or generic. In ICARUS there is a clear link to the aviation industry. However, when considering all three identified aviation data tiers, it becomes obvious that the scope of relevant assets is extremely broad and challenging.

Although the literature identifies the domain as an individual dimension, the current analysis considers the domain models relevant and they will be discussed in this context to account for the fact that being domain-dependent can be treated superficially, i.e. only enforcing a topic uniformity, but can also have deep consequences in the ways the marketplace is structured, which is the case in ICARUS. One of the main data sharing considerations identified in D2.2 was the lack of common data models which are necessary to enable data fusion and through it more advanced data analytics. Interestingly, aviation-based data and service marketplaces, at least in the information that is publicly available in their websites, do not often discuss the provision of a common underlying model – a counter-example here being GrayMatter's Airport Analytics (AA+)[9] solution. ICARUS places the need to conform to a common model at its core (as described in section 2.2.3) and binds many of its advanced services to the common data and metadata models that are created.

Data assets are mapped to the common data model (which is based on other domain models and standards and is described in Section 2.2.3) through a process described in technical deliverables. The non-data assets are also expected to adhere to certain common patterns and underlying models in order to be useful and directly exploitable by aviation stakeholders in general, but especially to members of the ICARUS ecosystem. The implementation of an algorithm adjusted to specific aviation data (e.g. trained for cargo airlines or optimised for level 1 airports), the implementation of an algorithm "packaged" with the required data to be applied on, the steps of an advanced data curation pipeline expressed in a ready-to-be-executed way, a visual static report of a performed analysis, an interactive dashboard invoking predefined functions in the background calibrated by the applied user controls, all constitute valid non-data asset examples to be considered in ICARUS. Inevitably, in order to be directly usable by the asset consumer, they are all bound to the ICARUS analytics and visualisation tools, in the sense that they need to be directly consumed and executed by them, hence they need to be expressed in a common language and format. Furthermore, the domain aspects that

[8] https://centreforaviation.com/data

[9] https://www.airportanalytics.aero/

reveal not only how the assets were created but also their value from a business perspective, are also captured in appropriately designed metadata (embedded in the ICARUS metadata schema).

The technical details of how this is achieved are out of scope for the current deliverable, however ensuring that this is true is inherently related to the assets' brokerage framework. As discussed in the marketplace landscape analysis presented in D2.2, relying only on descriptions which could be proven outdated, misleading or in any other way inconsistent with the actual offering is a common pitfall that can quickly discourage potential consumers. Instead, consistency should be enforced by design in order to offer advanced discoverability functionalities and facilitate intuitive asset exploitation.

The data policy and assets brokerage framework, as explained in detail in the next sections, leverages and is by design dependent on the common models and structures.

Dimension 5: Time frame

Marketplaces may offer support for static and/or real-time data, a factor that significantly affects the technical choices that need to be made, but also the business scenarios, hence also the brokerage workflow that can and should be foreseen. As documented in D2.2, especially in the DLT-enabled spectrum, numerous real-time data marketplaces have emerged leveraging the abundance of IoT data. A notable example of such marketplaces is Streamr[10], also briefly presented in D2.2, as it is blockchain-backed and offers smart contracts leveraging the Ethereum platform, like ICARUS. ICARUS however is not planned to support real-time data streams, a decision that has been already documented in previous deliverables.

Dimension 6: Type of data

According to (TowardsDataScience, 2018), an intuitive way to categorise data and data-based asset marketplaces is in regard to the type of data they allow their stakeholders to trade:

- Personal data (potentially sensitive) being traded through platforms that aim to help users monetize their data (indicatively: Datum[11], Datawallet[12]).
- Business data, which are traded through platforms that support business-to-business transactions (indicatively the DX Network[13]).
- Sensor data, traded through platforms that support IoT data streams and allow sensor owners to monetise the data produced by their devices (indicatively Streamr[10]).

This distinction is not helpful when considering leading data marketplaces which cover potentially all the above categories (as is the case with DAWEX[14]), nevertheless it achieves a very effective first level of marketplace grouping – even if in reality the sensor data type is not mutually exclusive with the other two. ICARUS is clearly positioned in the business data type and sensor data are not considered

[10] https://www.streamr.com/

[11] https://datum.org/

[12] https://datawallet.com/

[13]https://dx.network/

[14] https://www.dawex.com/en/

as relevant, as there is no support for real-time data streams. This does not preclude a data provider making available a dataset produced by one or more sensors, but it would be misleading to claim that ICARUS supports sensor data, as this would imply inherent support for IoT data streams which is not the case.

Dimension 7: Pricing model

Pricing models are of paramount importance in a marketplace setting and include free schemes, freemium, pay-per-use, flat rates etc. A distinction should be made between the pricing models enabled for the asset per se and the ways in which the marketplace platform will make profits (e.g. through subscription fees), which is an exploitation aspect and is not discussed here. The first part, the asset pricing model was one of the discussed brokerage challenges in D2.2 (Asset Brokerage Challenge III). The updated ICARUS sharing model, which will be presented in the next section, defines the pricing models selected based on the stakeholder requirements and shall be concluded and enforced in the context of the exploitation activities in WP7.

Dimension 8: Payment methods

This dimension becomes particularly relevant when exploring the DLT-enabled payment solutions that are available in marketplaces that embrace these technologies. Contract payments in ICARUS, as already documented in the technical deliverables and as revealed by the definition of the sharing model and the brokerage workflow that will be presented in the next sections, are considered to be performed externally to ICARUS with monetary transactions and not, as it is often the case when smart contracts are used, through smart coins. This is not a technically oriented choice and is relevant to, if not stemming from, the way the ICARUS asset sharing framework is designed. The decision is based on the following two points:

- Stakeholders in aviation are not accustomed to such payment methods and their limited familiarity with this technology could create scepticism and ultimately discourage them from participating in such sharing agreements.
- The amount paid per transaction, in the case payments are performed through smart coins, is inevitably revealed to all ledger participants. However, in the aviation industry, this type of information may need to be kept confidential among the asset provider and the asset consumer.

It should be noted however that this is a dimension that could change in the future and the ICARUS approach could be extended to leverage smart coins as well, especially as targeted aviation initiatives are emerging, e.g. the IATA coin (IATA, 2018).

Dimension 9: Data access

The technical means offered to access the data and data-based assets are more than an implementation decision, as a balance needs to be achieved among the needs of the target user in this respect, the nature of the provided assets and the effort required to support them. Possible alternatives here are simple download, APIs, specialised software and access through web interface. Selecting the appropriate data access means was identified and discussed as one of the asset brokerage challenges in D2.2 (Asset Brokerage Challenge IV: Provision Means of Data). Since then, the relevant needs and requirements have been considered and it has been decided that the ICARUS marketplace will support downloading and access through web interfaces and APIs.

Dimension 10: Trustworthiness

This is one of the subjective categories defined by the classification framework presented by (Stahl et al, 2016). The word subjective denotes that classification within this dimension cannot be easily verified but is based on the researcher's judgement. Trustworthiness here refers to the asset provider's trustworthiness and the ability of the prospective consumer to track and verify the original source of the traded asset. As already explained, ICARUS leverages the Blockchain technologies to ensure that data contracts' provenance can be verified. Furthermore, due to the controlled membership in the platform and the enforced KYC principles, it is expected that any scepticism regarding a member's trustworthiness, given that the validity of the identification information is guaranteed, will be really limited.

The analysis performed along these 10 marketplace dimensions was particularly helpful in outlining some design principles that the ICARUS brokerage framework should follow in order to fulfil its mission, as expressed in the dual set scope in the section's introduction. The next sections will present in detail the final version of the ICARUS data policy and assets brokerage framework.

3.2.2 Data Sharing Model

D2.2 reported on some preliminary assumptions and decisions on which the framework was built, which touch upon more technical dimensions of the complete ICARUS offering. The necessity of identifying, creating and maintaining links between the framework and the implementation, although quite intuitive, has been also explained in the introductory section and is documented in the description of its first version. As stated there, the framework drives the design and development of the project's data sharing mechanisms and cannot be examined independently of the overall architecture and envisioned usage workflows. Therefore, prior to describing the updated data sharing model, the list of technical decisions affecting it (as presented in D2.2) will be revisited and accordingly commented and extended.

Decision 1: Each data asset corresponds to a single dataset which can be easily represented in a tabular format. This means that irrespectively of the way a dataset is internally stored in the ICARUS platform, as far as the end users are concerned, a tabular format can be assumed to facilitate the analytics tasks.

Decision 2: All data assets available in ICARUS conform to the ICARUS common data model. This was not explicitly presented as a decision related to the framework in D2.2, however it constitutes one of the central ICARUS requirements and an enabler of almost all its advanced functionalities and as such, it should also be stressed in this context. The important role of the common data model was also described in the context of the fourth dimension of the marketplace landscape review in the previous section, whereas the details of how it is enforced and how datasets are mapped to the model is discussed in section 2.2.3.

Decision 3: Datasets provided by the ICARUS stakeholders will not leave their premises unencrypted, unless they comprise only public information. As previously explained, this requirement emerged in the context of the WP1 MVP validation activities and since then its significance in inspiring trust from the stakeholders to the ICARUS marketplace and overall platform has been continuously confirmed.

Prior to uploading a dataset, the data provider will select which columns should be encrypted, leaving only specific columns unencrypted. These columns, which hold spatiotemporal information, will be used to enable efficient data browsing and selection without revealing any sensitive/proprietary information prior to data acquisition. Since the ICARUS platform cannot access the encrypted columns, the initial steps of its data brokerage workflow should ensure that discovering datasets remains not only possible (which is a prerequisite for any marketplace), but also an easy and pleasant process. To further facilitate this process, the ICARUS data model has been extended to include "encrypt-ability" aspects of each field.

Decision 4: ICARUS will adopt a DLT-based solution for the data brokerage. This decision is at the core of the framework and its advantages have been documented in D2.2 but will be also discussed in section 3.2.3 of the current deliverable along with the description of the Blockchain-enabled data sharing workflows that are foreseen. The adopted DLT solution in ICARUS is the Ethereum platform, mainly due to implementation-related reasons. As technologies evolve, it is not in the scope of the framework to enforce a specific solution, but rather to establish the necessity of a DLT-based approach, which is also supported by the Blockchain solution assessment diagram by IATA (IATA, 2018), and build on top of the advantages that it provides.

Decision 5: Data entry to the ICARUS platform will require the provision of certain metadata by the data provider. ICARUS promotes the adoption of a powerful metadata schema, which has been specifically designed to ensure that all important asset features have been foreseen not only as "nice to have" accompanying information, but as facilitators and/or regulators of the ways an asset should be perceived, acquired and consumed. Although not always relevant to license policies and brokerage dimensions, the framework relies on and leverages various fields of the metadata schema which either play a role in understanding relevant limitations and potential risks and benefits of an asset being shared through the ICARUS system or can be seen as guarantees of certain asset attributes under a sharing agreement.

Decision 6: ICARUS does not provide any form of support (collection, ingestion, curation, acquisition, provision) for real-time data streams. This has been documented and justified in various deliverables and constitutes an important choice that greatly affects the ICARUS vision and the means to achieve it.

Having outlined the context in which the framework will operate, the next step towards its definition is to first define the ICARUS data sharing model, as explained in the section's introduction. The way the attributes that constitute the sharing model are organised is shown in the diagram below.

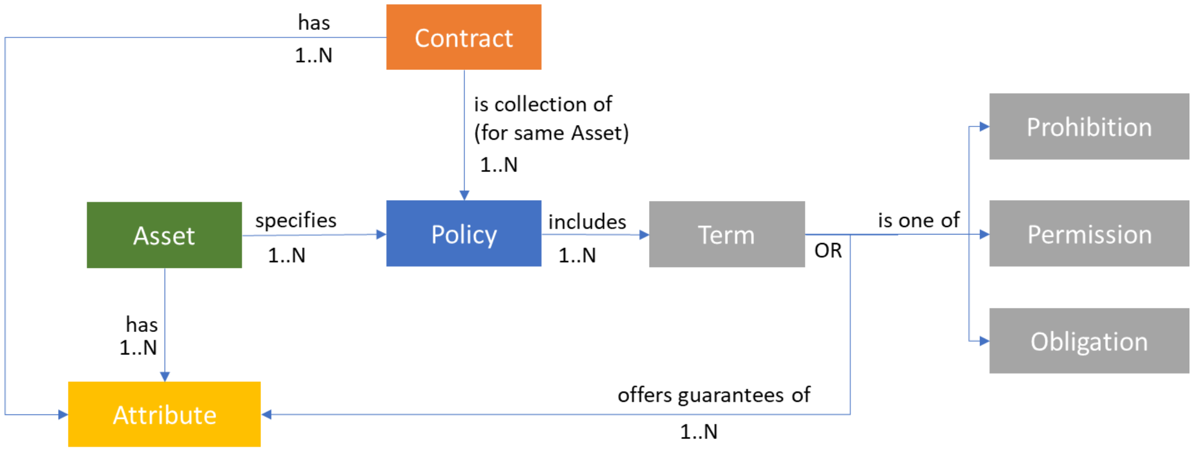


Figure 3-4: ICARUS Asset Sharing Model - High-Level View

The model comprises three core entities, namely the Asset, the Policy and the Contract and two supporting entities, namely Attributes and Terms, with the latter being specified as one of Prohibition, Permission and Obligation or an attribute guarantee. The model remains almost unchanged since its first definition in D2.2 with one important differentiation point: The Asset entity presented here is the evolution of the "Data Asset" entity and accounts for the fact that the complete ICARUS Data Policy and Assets Brokerage Framework foresees trading assets beyond data, as also explained in the introduction.

As algorithms and reports (analytics, visualisations) are forms of intellectual property, their inclusion in the model can be in most cases performed seamlessly. This does not imply that licensing and trading in their case is a straightforward task, but rather that they share the implications of data as tradeable assets and the complexities emerging when derivative work of a specific asset can become a new tradeable asset.

An Asset in ICARUS is defined as one of the following:

1. A specific dataset from a data provider or an open data source
2. An application that can be used as-is to perform a specific analytics and/or visualisation task inside ICARUS. Such an application may take the form of instructions that need to be expressed in the language used by the ICARUS analytics and visualisation tools and they may or may not be linked to specific data input sources.
3. The result of applying specific data analytics and/or visualisations on specific data, which can be perceived as a report.

A Policy is the way all legal, IPR, license, quality etc. terms are expressed. Each Asset specifies a number of Policies which control how it can be shared and accessed. A Policy comprises a group of terms. Terms, as explained, are either specific prohibitions, permissions or obligations or expressions of certain facts and/or qualities that represent attribute guarantees. There is an important distinction between the way the term Policy is used here compared to the Data Access Policies defined in Section 2.3.1. The aim of the data sharing model is to enable the definition, expression and enforcement of what constitutes the terms (in their common meaning) of a data sharing contract. These terms are of two kinds in ICARUS:

1. Attributes of the asset being traded which act as guarantees for both the provider and the consumer of what the first needs to provide and what the latter is expecting to acquire.
2. Prohibitions, permissions and obligations of both interacting parties in respect to the asset being traded.

Because of the digitalisation of the data sharing process, some of the above Terms (the term now being used in the data sharing context), may be context-aware and could also under specific circumstances transition from attributes to one of {obligation, prohibition, permission}. As an example, quality related attributes of an asset act as guarantees of the expected quality metrics of the asset, but can also be expressed as obligation of the data provider to maintain a certain quality level when the provision of data updates is foreseen in the contract. Access-related metadata of an asset are a special type of policies, which in the scope of an online marketplace are by definition context-aware. The way the respective policies are formulated has been examined in Section 2.3.1 and the usage of XACML to express them has been explained. However, it should be noted that the access related policies of the data sharing model are part of the framework's theoretical foundations, whereas Section 2.3.1 presents the way they have been materialised.

The last entity shown in the diagram, the Contract, represents the official data sharing agreement between a data provider and a data consumer in regard to one single Asset under specific Policies which are either snapshots of the asset's qualities to be used as guarantees or framings of the expected/allowed actions of the interacting parties in regard to the asset.

It should be mentioned here that contracts in ICARUS are hybrid, in the sense that although a smart contract mechanism is in place, a decision has been made to maintain a textual part of the underlying agreement. To avoid confusion, a definition of smart contracts is required: "smart contracts are self-executing contracts with the terms of the agreement between buyer and seller being directly written into lines of code. The code and the agreements contained therein exist across a distributed, decentralized blockchain network". The need to be self-executable precludes any text written in natural language. However, based on the requirements elicited from the aviation stakeholders (both participating in ICARUS and external) and on the landscape analysis on smart contracts in aviation and automated data sharing contracts in the scope of intellectual property, this was not considered a viable option in ICARUS. The limited familiarity of stakeholders in the aviation industry with this technology and the gravity of data sharing agreements in aviation both in terms of monetary value and legal implications push towards a hybrid model, at least in the current maturity level of the smart contract solutions. Therefore, attributes and terms that can be unambiguously defined and evaluated are included in the self-executable contract, i.e. the smart contract, which is always bound to a textual agreement where the remaining terms that cannot yet be expressed in the smart contract's language (e.g. indemnification clauses) are written. This accelerates the drafting process of the contract and automates the execution of the asset trading process in the majority of cases, whilst allowing legal departments and procedures to operate in the same way as before in the case of a misbehaving party or any dispute that may arise.

The following two tables present details on the way the data sharing model is instantiated. The information shown in them is an update of what was presented in D2.2 and reflects the decision to follow the hybrid contract approach, the extension of assets to include more than datasets and the feedback received regarding the features that are both needed and can be made available. The indicative information shown here is embedded in the complete ICARUS metadata model presented in Annex II, therefore only asset attributes that belong to sharing and the distribution metadata categories are included.

Table 3-11: ICARUS Smart Contract relevant to the data sharing model

Attribute	Description & exemplary values
asset id hash	Unique identification of the asset in ICARUS – hashed to avoid being in any way exposed
asset filters	Any evaluatable filter on the asset. e.g. In the case of data assets, this could include spatiotemporal coverage based on specific asset columns/fields. All metadata model fields can be used as valid filters combined with the desired value(s) and/or value range.
asset fields	Applicable in data assets only, includes the fields of the ICARUS common data model that should be present in the dataset
validation date	Timestamp when the contract was validated
duration	Contract duration exprssed as dates range
provider	The id of the asset provider (ethereum address)
consumer	The id of the asset consumer (ethereum address)
free terms hash	Hash of the contract part written in natural language

Table 3-12: ICARUS Policies included in a valid contract

Policy Category	Terms Scope	Exemplary values
Pricing	cost calculation scheme	fixed per row | fixed per asset | request dependent
	amount	amount in euro | available upon request
	payment method	credit/debit card| bank transfer| online payment services | other
Responsibility	copyright ownership	owner of asset
	addressed to	individual | group | legal entity
	liability & indemnification	custom clauses (included in the natural language textual part of the contract if needed)
Rights & Usage	license	custom | CC | CDLA | Open Data Commons | ...
	derivation	modify (Y|N) | excerpt (Y|N) | annotate (Y|N) | aggregate (Y|N)
	attribution	required | not required
	reproduction	allowed | prohibited

	distribution	allowed | prohibited
	target purpose	business | academic | scientific | personal | non-profit
	target industry	limited to Aviation | excluding Aviation | all
	offline retention	allowed | prohibited
	re-context	allowed | prohibited
Privacy & Protection	privacy & sensitivity compliance	custom clauses (included in the natural language textual part of the contract if needed)
	liability	custom clauses (included in the natural language textual part of the contract if needed)
	applicable law	custom clauses (included in the natural language textual part of the contract if needed)

Table 3-11 only shows the smart contract fields that are relevant to the data sharing model presented here (i.e. the attributes) and not fields that stem from the technical details of the underlying implementation. One of the fields that are not shown, the stage, is of particular importance for the workflows enabled by the framework and will be examined in the next section.

Table 3-12 does not refer to the access policies, as the way they are instantiated has been described in detail in section 2.3.1. It should be noted that section 2.3.1 refers to access policies in general in the ICARUS platform which are also applied to restrict asset discoverability and exploration also prior to acquisition. Policies included in a valid contract override general asset policies, e.g. a data provider cannot unilaterally invalidate an existing contract by simply altering the asset’s policies, the latter being completely within his/her rights.

As explained before, the ICARUS Data Policy and Assets Brokerage Framework seeks to achieve a balance between expressivity and applicability / efficiency; therefore, the proposed sharing model may still be simplified when compared to the complete set of considerations and options of asset trading in aviation. It is believed that the underlying assumptions do not harm the applicability of the model and that the design decisions that were made and presented here, will ensure the smooth operation of the sharing mechanism and will inspire trust in the stakeholders to use it.

3.2.3 Blockchain-enabled Asset Sharing Workflows

The second part of the ICARUS Data Policy and Assets Brokerage Framework, as explained in the section’s introduction, is related to the definition of the workflows that capture the basic provider-consumer interactions. D2.2 presented the simplest version of a data trading workflow, assumed to be completed smoothly and based on it discussed some anticipated deviations and variations. This section will adopt the same structure, although the updated version of the simple (core) workflow introduces by itself a higher level of complexity.

As in the first version, the core asset brokerage workflow comprises three phases, with phase I now being split into two alternatives (Phase I-a and Phase I-b, depending on the type of the asset).

Prior to entering phase I, the following prerequisites are considered as satisfied:

- Data assets have been successfully mapped to the ICARUS common data model.

- Other asset types have been successfully registered in the ICARUS application catalogue, which also ensures that a minimum level of uniformity in their structure is guaranteed.
- Asset providers have gone through the metadata provision and data access policy definition steps.
- ICARUS administrator has gone through the metadata provision and policy definition steps for the open datasets.

Phase I-a: Data Assets Exploration

The process corresponds to Phase I of the workflow presented in D2.2 and remains almost unchanged. It is initiated by an ICARUS user performing a query to search for data. The query construction process is three-fold:

- The user selects which fields from the common data model should be available in the returned results.
- The user defines filters to be applied on specific fields:
  - For the datasets' unencrypted fields, the user may define filters that are used to specify and/or refine the (possibly) multi-dimensional spatiotemporal bounds of the query. The multi-dimensionality emerges from the fact that numerous fields containing spatial and temporal information may be at the same time present in a dataset and unencrypted.
  - For the encrypted fields, ICARUS allows data providers to mark an encrypted field as searchable / indexed. This is only available for specific fields and aims to facilitate the asset's discoverability without compromising its content. When a field is marked both as encrypted and indexed, then prior to the dataset's encryption and depending on the field's type, its unique values or its value bounds are extracted. Hence, during the query creation, the user may also define filters for certain encrypted fields, although the two filter types are not handled in the same way.
- The user defines filters to be applied on the metadata of the data assets.

The system then combines the three parts into one query which is performed to identify the matching data assets from the ICARUS database. These are either individual datasets that match the whole query or appropriate combinations of datasets.

Phase I-b: Non-data Assets Exploration

With the ICARUS marketplace providing assets that are not limited to data, the user should be facilitated in the exploration of such assets as well, in order to identify potentially interesting cases that could lead to sharing agreements. Aviation-related data lie at the core of the marketplace and ICARUS aspires to create an ecosystem of data-enabled insights sharing.

Similar to Phase I-a, this phase is also initiated by an ICARUS user performing a search query but now for non-data assets. It is reminded that non-data assets also share a common underlying metadata model, as well as a common definition template which ensures they can be processed by the ICARUS tools. Apart from the metadata foreseen by the ICARUS metadata schema (Annex II) and the ones defined by the Sharing Model, non-data assets are accompanied by numerous, more targeted, and more technically oriented, metadata, the majority of which are calculated and assigned directly by the system as a result of the steps the provider went through in order to create them. A detailed list of these metadata attributes is out of scope here due to their technical nature. The key take-away is that queries created to search for non-data assets should be flexible and customisable, yet the complexity level may increase significantly, especially when looking for combos, i.e. assets that comprise other assets (algorithms, data, visualisations), each of which may adhere to different policies and IPR.

Both alternatives of Phase I end with the query results being returned to the user. In all cases, the results that are presented to the user adhere to the limitations imposed by the defined terms and policies. Browsing and reviewing the presented result list, the user may choose to issue a request to obtain a specific asset.

Phase II: Smart Contract Drafting

In order to enter this phase, phase I should have ended with a request being issued by the user who performed the query to the provider of the selected asset in order to purchase it. In the first definition of the workflows presented in D2.2, an assumption was made in this phase that the asset provider would accept the request and that the contract terms would be drafted based on the available metadata and would be directly accepted by both parties. Refusing the request and negotiating over the contract terms were treated as deviations from the normal flow (Deviation 1 and Deviation 3 respectively in D2.2). Although this was acceptable in that early version, both deviations have been now included in the core workflow, as the received feedback showed that they are very common processes. Therefore, the phase is re-designed as follows:

Step II.1: The data owner is notified of the request. In case the request is not considered interesting, the provider may refuse to proceed and the process ends without any contract being drafted. If the provider chooses to follow through, the next step is the definition of the contact terms (as foreseen by the sharing model). It should be noted that many of the terms will be pre-calculated as they stem from the asset attributes and policies, as well as from the request that initiated the process. The provider should then be guided on providing the remaining required terms and on updating the pre-calculated ones if needed. Apart from the structured terms, the provider in this phase should also provide the natural language part of the agreement (unless not desired). Once this process is completed, a smart contract is created and uploaded to the blockchain. The contract is now in draft stage, as denoted by the field “Stage” which is used as a flag for the contract’s status and is in this case set to “Draft”.

Step II.2: Being notified for the draft contract, the prospective consumer should proceed to review the terms and may act in the following three ways:

a. Reject the contract and end the process. In this case the smart contract's "Stage" field is set to "Rejected".
b. Accept the contract as-is. The smart contract's "Stage" field is set to "Accepted". This concludes Phase II.
c. Edit the contract terms. Editing is possible both for the smart contract part and for the natural language document. Once editing is completed, the smart contract is updated accordingly and its "Stage" field is set to "Negotiating".

Step II.3: This step can be reached only from the sub-step II.2c. The asset provider is notified for the updates in the contract and should proceed to review them. There are three possible actions:

a. Reject the contract and end the process. In this case the smart contract's "Stage" field is set to "Rejected".
b. Accept the updated contract as-is. The smart contract's "Stage" field is set to "Accepted". This concludes Phase II.
c. Edit the contract terms. Editing is possible both for the smart contract part and for the natural language document. Once editing is completed, the smart contract is updated accordingly and its "Stage" field is set to "Draft". It should be noted that editing the contract results in a different state depending on who performed the edit, the provider or the consumer. The process from here goes back to step II.2.

The negotiation over the contract terms corresponds to a loop between steps II.2 and II.3. This loop ends once the contract stage is set to either "Rejected" or "Accepted". In the first case, the workflow is terminated in Phase II, otherwise it moves to Phase III.

Phase III: Smart Contract Validation

To enter this phase, the smart contract has been accepted by both parties (and its "Stage" field is set to "Accepted"). When this is done, the data consumer should proceed with the payment. Once the data provider validates that the payment was successfully made, the smart contract stage changes to "Paid", which is the stage that denotes that the smart contract is considered valid and the ICARUS platform will allow the consumer to obtain the asset. Access to the asset will be automatically ensured for the time interval defined by the duration field of the smart contract.

The rationale for keeping the payment process detached from the platform and not implementing a smart coin solution has been explained in Section 3.2.1. However, the non-automated way in which it is acknowledged can be seen as a potential weakness of the workflow. It should be stressed that ICARUS is based on strong KYC principles and therefore trust is to a great extent guaranteed among its members. However, to add an extra level of security, technical measures are also foreseen to be put in place. Specifically, to avoid issues caused by potentially malicious data providers, the implemented system should have a security mechanism to bypass the data provider's validation process and set the stage to "Paid", hence validate the contract independently, to address cases where the consumer has proof of a completed payment, but the provider does not honour the agreement.

From an implementation perspective, additional technical measures are becoming available to help address this issue (e.g. blockchain oracles[15]) and could be considered and adopted in the future. As already explained, the technical details of the way the workflow foreseen by the framework is instantiated are not important. Nevertheless, as with all legally binding agreements, specific legal procedures are also foreseen when disputes arise, which are also out of scope to be described here.

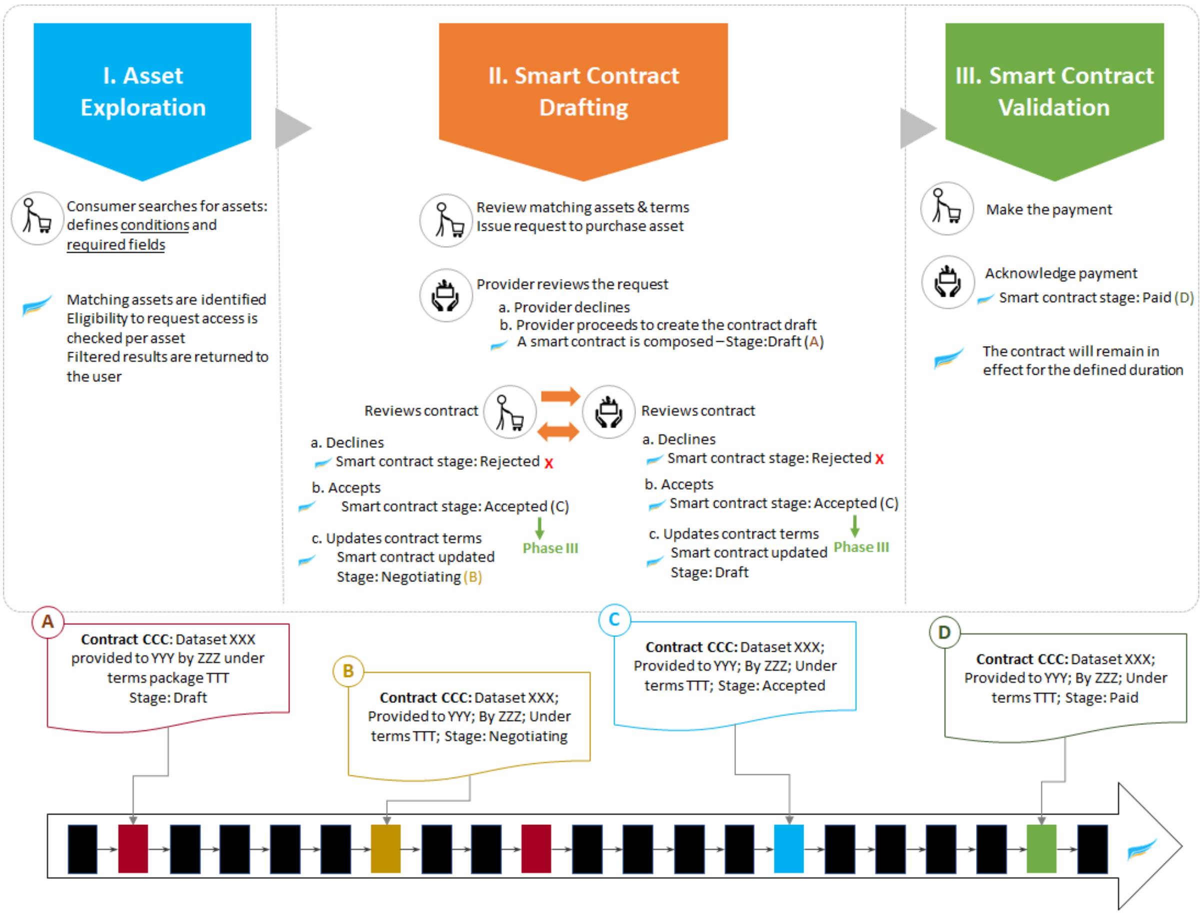

I. Asset Exploration
II. Smart Contract Drafting
III. Smart Contract Validation
Consumer searches for assets: defines conditions and required fields
Matching assets are identified
Eligibility to request access is checked per asset
Filtered results are returned to the user
Review matching assets & terms
Issue request to purchase asset
Provider reviews the request
a. Provider declines
b. Provider proceeds to create the contract draft
A smart contract is composed – Stage:Draft (A)
Reviews contract
a. Declines
Smart contract stage: Rejected x
b. Accepts
Smart contract stage: Accepted (C)
c. Updates contract terms
Smart contract updated
Stage: Negotiating (B)
Phase III
Reviews contract
a. Declines
Smart contract stage: Rejected x
b. Accepts
Smart contract stage: Accepted (C)
c. Updates contract terms
Smart contract updated
Stage: Draft
Phase III
Make the payment
Acknowledge payment
Smart contract stage: Paid (D)
The contract will remain in effect for the defined duration
A
Contract CCC: Dataset XXX provided to YYY by ZZZ under terms package TTT
Stage: Draft
B
Contract CCC: Dataset XXX; Provided to YYY; By ZZZ; Under terms TTT; Stage: Negotiating
C
Contract CCC: Dataset XXX; Provided to YYY; By ZZZ; Under terms TTT; Stage: Accepted
D
Contract CCC: Dataset XXX; Provided to YYY; By ZZZ; Under terms TTT; Stage: Paid

Figure 3-5: Core Asset Trading Workflow

Having defined the core Blockchain-enabled asset sharing workflow, possible deviations are examined. As already mentioned, deviations 1 and 3 documented in D2.2 have now been assimilated to the core flow.

Deviation 2 was about cases that a prospective asset consumer wants to issue requests to purchase more than one assets that were returned as a single result of the performed query. As explained then, this is expected to be a very common case, since data asset results may be combinations of datasets and non-data assets may also be composed by more than one underlying assets. The solution foreseen in such cases remains the same: For each of the assets included in the result, a separate request is issued and the process spawns parallel processes, equal in number with the number of different assets. Each process is a complete workflow comprising the three phases described above. No

[15] https://blockchainhub.net/blockchain-oracles/

additional possible deviation is foreseen. Other possible implications and limitations will be discussed in section 4.2.

4 ICARUS Considerations and Positioning

4.1 Data Management Considerations and Open Questions

In Table 4-1, a number of data management considerations is effectively discussed in order to externalize the project positioning and current status. Since many of the considerations are practically inherited from D2.1, the revised ICARUS perspectives are exposed, highlighting whether such considerations are resolved in this deliverable or maintain their "open challenge" status for the forthcoming implementation activities.

Table 4-1: Data Management Considerations

Area	Key Consideration / Open Challenge	Description	Comments / Status in ICARUS
Data Collection	Performance aspects when checking-in very large data assets in the ICARUS Platform	Since the data that the aviation stakeholders have to handle includes potentially very large, batch files containing historical aviation data, performance issues may naturally arise when a stakeholder attempts to check in and transmit such files.	In order for data providers to experience low, predictable latency in uploading the data assets (and eventually in executing queries) in ICARUS, an early experimentation was performed with high-performance, secure, reliable data transfer protocols (such as GridFTP) to achieve the required throughput without losing any data in the process. However, it remains open as it such a protocol is planned to be supported in a future, stable ICARUS platform release.
	Audit trail for the check-in process	In order to address the situations when a data asset is in a limbo state prior to its check-in, ICARUS shall provide a complete audit trail of the check-in process, recording the status and providing the related feedback to the data provider.	In its beta platform release, ICARUS already supports a full audit trail of the check-in process which allows a data provider to leave the data check-in job definition as open in any step (i.e. Mapping, Cleaning, Anonymization, Encryption) and resume at a later stage.
	Check-in process for Data Collection Level 3 (APIs)	The check-in process for data collection level 3 (APIs) poses many difficulties that require human intervention.	In preparation of its beta release, ICARUS has focused entirely on the data upload modality, thus the APIs check-in process has inevitably lagged behind and remains an open issue for the implementation how it will be handled (although certain similarities to the currently supported check-in process are expected).
	Data check-out process	The decommissioning or deletion of data assets that may happen, for example, when a data asset is no longer valid or its provider wishes to withdraw it. However, such a decision has further repercussions in case a data asset	The data check-out process is prohibited for data assets that are already part of active data contracts till their expiration. As the data provider may block any further purchases of the specific data asset, it should be eventually possible to "delete" a data

Area	Key Consideration / Open Challenge	Description	Comments / Status in ICARUS
		that a provider wishes to check-out is already part of active data contracts. To this end, the data check-out process shall be examined in detail in collaboration	asset in terms of: (a) the metadata of such a data asset need to contain information that the data were deleted, and if they were archived, how and where an archival copy can be requested, (b) ensuring the trusted destruction of the actual data.
	Data collection roadmap in ICARUS	ICARUS needs to sketch a thorough plan for data population and data maintenance, starting from the data providers within the consortium and expanding to different data providers with which ICARUS interacted during its external MVP activities.	Although D2.1 anticipated the data collection roadmap in this deliverable, it was decided to be reported on the management report due to its confidential nature in contrast to the public type of the present deliverable.
	Data update strategy dictating replacement	In cases of encrypted data, the overall data update strategy presents certain challenges that need to be overcome.	The data update strategy has been defined based on the experience from the data check-in process. If the data are encrypted, the non-encrypted columns may not be adequate in order to uniquely identify which rows to replace, thus the data replacement strategy cannot be effectively applied.
Data Cleaning	Dynamic data validation rules definition	The data validation rules that applied on the data validation process is shall be more dynamic, flexible, complete, coherent and efficient and should be designed in collaboration with the data providers and the demonstrators of the project taking into consideration the various aspects of each data source, such as the different context, schema and the type of information included. Additionally, the re-evaluation of these rules shall be allowed in order to ensure that new requirements are addressed.	In the design process of the data validation rules the consortium elaborated with the data providers and the demonstrators of the project in order to extract their needs in terms of data validation that should be performed in the context of the data cleaning process. After analyzing the collected feedback, as well as the information for the nature of the data that they will provide in the ICARUS platform, the extended list of validation rules that were presented in section 2.2.2 was compiled. Additionally, this list can be easily extended as the project evolves with new validations rules if the need for this arises.
Data Cleaning	Data cleansing process performance aspects	The data cleaning process should be designed carefully taking into consideration that most of the data cleansing tasks can be computationally intensive and time-consuming tasks which may introduce delays in the check-in process	The data cleaning process was carefully designed in order to adhere the principles of the effective and efficient data cleaning. The effectiveness of the process in terms of performance is addressed with the design specifications of the relative component of the ICARUS architecture in which the best practices for big data processing were adopted.

Area	Key Consideration / Open Challenge	Description	Comments / Status in ICARUS
Data Cleaning	Bias-free data completion process	The data completion or missing values handling can be addressed with the methods and processes that are offered within the data cleaning process. However, applying any of these methods and processes without a valid analysis of the root cause of the missing values and the impact on the statistical analysis may inevitably introduce a significant effect on the conclusions that can be drawn from the analysis.	The data cleaning process incorporates a variety of data completion or missing vales handling methods as described in section 2.2.2. However, the data cleaning process itself cannot provide any assertions regarding the suitability of each method on the dataset that will be used in the process as the content and the nature of the included information is agnostic to the process. Thus, the responsibility of ensuring that no biased information will be introduced in the outcome of the process lies with the data provider who should have a deep understanding of the underlying context of the information of the dataset and the implications that might be introduced with each method offered and he/she should select the most appropriate ones.
Data Mapping & Linking	Evolution and lifecycle management of the ICARUS common aviation schema	As the project progresses and new datasets are checked in in the ICARUS platform, no matter how thorough and well-designed the ICARUS common aviation schema is, it is inevitable that it will always need to evolve to address new needs, and refinements and updates will have to be introduced in a consistent manner, that is centrally controlled and managed by an ICARUS administrator.	Despite the increased complexity and the effort that will be required as the size of the model increases, abiding to the defined process in section 2.2.3 is a prerequisite in order to maintain the integrity and consistency of the ICARUS common aviation model. It is already planned to expand the data model in the next platform releases in the following ways: (a) support for more aviation data standards (e.g. Airline Industry Data Model (AIDM), Flight Information Exchange Model (FIXM)) and full support for SSIM, (b) impose compliance with specific code lists (to facilitate the upcoming data linking at query time), (c) make clearer in the model which related term comes from which standard (in order to provision for exporting data in specific supported industry standards the user selects in the future platform releases).
	GraphQL Schema definition; Dependency with the ICARUS common aviation schema;	As defined in D2.1, a set of challenges were stemming from the use of GraphQL to ensure alignment with the ICARUS common aviation data model and successful and effective operation of the GraphQL engine.	Although in D2.1, the data linking approach was tightly interconnected with the use of GraphQL, this approach proved ineffective in the ICARUS beta platform release due to the end-to-end data encryption that is enforced. It was thus revised to the data linking approach that was framed in section

Area	Key Consideration / Open Challenge	Description	Comments / Status in ICARUS
	Implementation of Resolvers		2.2.3 that does not pre-link the results. The use of GraphQL may be eventually reconsidered for the permanent secure spaces a data consumer has in ICARUS, but the experimentation stage of such permanent spaces forces keeping the specific challenge as open.
	Mapping conflict management in specific datasets	Multiple properties of the data that are checked in (referring to flights code sharing in the OAG data) are mapped in the same property in the ICARUS common aviation data model (*Mapping Case 5 (N:1)*).	Although the ICARUS common aviation data model does not impose any cardinality restrictions (i.e. that a specific property must appear once), it was decided to postpone handling such an issue until stakeholders upload their data to realize whether it is an exception or a common phenomenon. If it is an exception, it shall be handled with different properties in the data model, but if it is a common phenomenon, it shall be directly supported in the future releases of the ICARUS platform.
Data Provenance	Efficient metadata lifecycle management	In order to ensure provenance at dataset level as imposed in ICARUS, the role of metadata for recording each and every data-related activity becomes crucial.	Separation of concerns between the storage of the accompanying metadata (at the different categories) and the (encrypted) storage and indexing of the corresponding data asset is ensured. The metadata utilized are aligned to the ICARUS metadata schema.
	Minimal intrusive provenance	Capturing provenance is not free of cost and certain provenance decisions (e.g. fine-grained, at data level) may severely impact performance due to intrusiveness.	Since the ICARUS data provenance method is designed to be minimally intrusive and lightweight at dataset level, this issue is not considered as critical at the moment this report was written. However, in case any performance issue is detected in the implementation, the provenance efficiency will be revisited.
	Storage overhead	Although ICARUS shall be handling big data and scalable storage is to be ensured by design, the metadata schema and the provenance approaches to be adopted pose significant storage overhead.	In the early experimentation with the ICARUS beta platform, the storage overhead does not appear to be unmanageable since the audit trail remains at dataset level. In case any issue appears though, ICARUS shall revisit its provenance approach and decide how to "compress" the provenance data.
Data Encryption	Trust in uploading data in the ICARUS Platform	The aviation stakeholders in order to upload their data and trust the platform require end-to-end encryption for all the data operations performed in the platform.	The consortium incorporated the end-to-end data encryption approach that has been described in section 2.3.2 across the ICARUS platform. This approach is successfully addressing the concerns of the aviation

Area	Key Consideration / Open Challenge	Description	Comments / Status in ICARUS
			stakeholders and increases their trust in the platform.
Data Encryption	Computation needs and repercussions of data encryption to real-time data sharing	The ICARUS platform needs to effectively address the computation needs and overhead that is introduced the end-to-end encryption that is incorporated in the platform operations in order to avoid the inevitably delay that will be introduced in the real-time data sharing within the aviation data value chain.	The consortium decided to adhere the principle of the data security at all costs in order to support the scenarios of the ICARUS demonstrators and increase the trust of the ICARUS stakeholders. The trade-off for this decision is the delay in the real-time data sharing. As the project evolves, the consortium will investigate further this issue and explore any possible technological solution that might resolve it. However, at the time of writing this issue remains open.
Data Access Control	Change management in data access policies	The management of the data access policy lifecycle is a crucial, yet challenging task. As new needs emerge, it is expected that policies may be changed and revoked, yet ICARUS needs to able to handle effectively the management of the data access policies and ensure that undesired behavior is avoided during runtime at all costs. In case where multiple policies are changed and deployed simultaneously (even if editing the policies for a specific data asset is always locked and limited to one user per data provider), ICARUS should be able to perform conflict detection and resolution in an effective way.	The Data Access Control method is designed in a way that management of the access control policies is performed efficiently and effectively. Moreover, the design and specifications of the component of the ICARUS platform that provides the implementation of this method is enhanced with advanced management capabilities for the whole access policy lifecycle. Thus, the issue has been effectively resolved.
Data Access Control	Know-your-customer identity management	ICARUS shall apply state-of-the art KYC (Know-your-customer) guidelines with very rigorous customer acceptance policies at organization level and detailed customer identification procedures to prevent its platform from being misused, intentionally or unintentionally, by unauthorized stakeholders. However, if this policy poses more barriers to entry for aviation stakeholders than appeasing their security concerns it might be reconsidered.	ICARUS adopted the KYC guidelines as part of the decisions undertaken to increase the trust of the stakeholders to the platform and to ensure that all the security aspects of the platform are effectively addressed. For this reason, the platform incorporated a strict registration process that adheres these guidelines. This decision has not changed till the moment of writing and it will be re-evaluated during the platform evaluation phase depending on the collected feedback.
Data Access Control	3rd party vs own Certificate Authority	In order for the data safeguarding methods as a whole to seamlessly be put into action, a Certificate	In ICARUS, the option of setting an ICARUS Certificate authority in order to enable the effective management of

Area	Key Consideration / Open Challenge	Description	Comments / Status in ICARUS
		Authority (CA) needs to issue certificates for all involved stakeholders that will be used to establish secure communication between them and the platform. The options for this is the usage of 3rd party Certificate Authority or setting its own certificate authority.	the issued certificates has been selected at the moment.
Data Encryption	Key revocation	Revocation of the produced symmetric keys that are utilised in the data encryption method described in 2.4.2 might be needed for various reasons. The ICARUS standing is that if a data consumer tries to reuse a decryption key in the future without an active data contract, he/she will have to face the legal repercussions since the data contract terms will have been violated, in the same way that data providers traditionally react at the moment. Furthermore, the feasibility of technically supporting a solution will be examined.	Providing a technical solution to support an effective and efficient revocation process for the produced decryption keys is a nontrivial problem. The consortium will investigate the possibility of providing a technical solution that will not pose any issues in the described data encryption method and that it will ensure the undisrupted operation of the secure data sharing functionalities. However, at the time of writing this deliverable, the issue remains open.

4.2 Data Value Enrichment Considerations and Open Questions

The current section serves a dual role:

I. To collect the challenges that were discussed in D2.2 for each of the two main aspects of the data value enrichment methods, i.e. data analysis and data brokerage, and report on the current ICARUS standing in regard to their status, i.e. whether they remain relevant and whether they have been addressed and to what extent.

II. To report on new identified challenges related to the data value enrichment methods, some of which may stem from previously identified challenges and the approach ICARUS has taken to address them. This is why the comments and status column in Table 4-2 often contains discussion and remarks for other/ new challenges.

Any considerations that arise from unresolved challenges will be leveraged for better decision-making during the design processes of the ICARUS offering and will serve as guidance during the implementation phases in order to proactively address potential issues. It should be noted that in D2.2 a distinction was made between data sharing and data brokerage, the first addressing more the IPR and incentivisation issues, and the latter the actual brokerage aspects such as contract drafting and stakeholder interactions. The two areas are now integrated into one, shown in the table below as "Asset Sharing", to denote that they are both addressed by the ICARUS data policy and assets brokerage framework and the processes and workflows foreseen by it.

Table 4-2: Data Value Enrichment Considerations

Area	Key Consideration / Open Challenge	Previous Reference/ Discussion	Comments / Status in ICARUS
Data Analytics	Data Availability and Quality	D2.2 Section 2.3	The need to obtain appropriate data in order for data analytics processes to produce meaningful results, although intuitive, has been explained also in D2.2. Data appropriateness here refers to both quantity and quality so as to avoid "Garbage In-Garbage Out" situations. A recent IATA white paper (IATA, 2019) also acknowledges this as a key aspect for the proliferation of data and data science in aviation. This is obviously an intrinsic problem in data analytics and ICARUS cannot by itself address it. However, through actively enabling and promoting data sharing, through concrete methods that have been described here as well as in other ICARUS deliverables, ICARUS aims to significantly increase data availability and incentivise provision of high-quality data.
Data Analytics	Type of Problem	D2.2 Section 2.3	Identifying the right approach for a given problem is an important part of every data analysis workflow. Hence, the "type of problem" challenge stresses the importance of exploring the distinctive features of a given problem prior to commencing the analysis effort, which is an inherent data analysis difficulty that reduces the automation level of certain steps in the data value enrichment methods. As the project progresses and the targeted use cases become more concrete, the proposed data analytics approaches are being refined, as can be seen in section 3.1 in the current deliverable and as will be further documented in the WP5 deliverables. It should be stressed that, as also explained in section 3.1, domain knowledge is extremely important for data analysis in general and for addressing this issue in particular.
Data Analytics	Stakeholder Perspective	D2.2 Section 2.3	The complexity of the aviation value chain and the interconnections among use cases and stakeholders require assumptions and simplifications to be performed. As with the previous Data Analytics challenges, this is also not addressed in a generic manner, but per case in the scope of ICARUS. The way the different perspectives are considered is, up to an extent, outlined in Section 3.1 in the current deliverable and will be further examined, where necessary, in the WP5 deliverables.
Data Analytics	Mentality	D2.2 Section 2.3 IATA integration of analytics Collaboration and domain knowledge	Successfully integrating data analytics processes in the operations of aviation organisations and companies, also recognised by IATA (IATA, 2019) as one of the key aspects for the proliferation of data and data science in aviation, requires a change in mentality. ICARUS will contribute towards this goal by providing the framework, the tools and the services required to build data-enabled solutions in the aviation industry. The change in mentality is closely related to the previously discussed need for domain knowledge, as it implies

Area	Key Consideration / Open Challenge	Previous Reference/ Discussion	Comments / Status in ICARUS
			the need for collaboration among data analysts and aviation domain experts that have a deep understanding of the underlying business and its operations. Only when built through such collaborative processes the data analysis models are able to capture the real-life peculiarities of the aviation domain and become applicable - hence truly valuable.
Asset Sharing	Data Privacy and Sensitivity	Deliverable D2.2 Section 4.5	Potential data privacy and sensitivity issues were discussed in D2.2 as part of the Key Data Sharing Consideration I. ICARUS places the data provider at the centre of the decision making regarding the data assets that will be uploaded in the ICARUS platform, either for own usage or for sharing purposes. Relevant legal aspects and safeguarding methods have been examined in all WP2 deliverables and also appropriate tools will be provided to the users, e.g. for data anonymisation and for defining asset features related to regulatory compliance. Nevertheless, the data provider remains the main responsible for exposing sensitive data and for ensuring regulatory compliance where applicable. Any legal implications will therefore be handled externally to ICARUS following the foreseen legal procedures.
Asset Sharing	Trust	Deliverable D2.2 Section 4.5	Trust towards the platform, as well as among its members has been extensively discussed (also in section 3.2.1 of the current deliverable) and several ICARUS design decisions stem from the need to ensure trustworthiness across the complete asset brokerage workflow. Data encryption, Blockchain and smart contracts, controlled membership and enforcement of KYC principles in general, are among the measures taken towards this goal.
Asset Sharing	Security	Deliverable D2.2 Section 4.5	Fine grained access control to data and non-data assets, private secure execution spaces and data encryption are among the foreseen ways in which ICARUS addresses concerns around the platform's security levels.
Asset Sharing	Data IPR, licensing and ownership	Deliverable D2.2 Section 4.5	As already explained, these issues pertain mainly to legal aspects of intellectual property sharing. Data IPR, as discussed in D2.2, is by itself a contentious issue. Considering that ICARUS extends brokerage to derivative data and also non-data assets (e.g. algorithms, visualisations built on multiple underlying datasets possibly from different owners etc.), the number of implications and contradictions that may arise becomes impossible to handle. As stated before, contract management needs to account for two contradicting forces: the requirement to homogenise the brokerage process under a common framework to increase efficiency and the demand for customised rights, terms and conditions and negotiation mechanisms in the complex landscape of data-enabled assets' IPR. The ICARUS data policy and assets brokerage framework, along with the defined metadata schema and the underlying sharing model, provide guidance and support throughout the complete workflow of aviation data-enabled assets' sharing. Nevertheless, it is the ICARUS positioning that addressing these challenges cannot be achieved through automated

Area	Key Consideration / Open Challenge	Previous Reference/ Discussion	Comments / Status in ICARUS
			processes but eventually requires legal procedures to be put in place.
Asset Sharing	Lack of common data and metadata model	Deliverable D2.2 Sections 4.5 & 5.4	The design and usage of the ICARUS common data model has been extensively discussed and its benefits for the proliferation of data sharing and data analytics in aviation have been justified. The lack of a common data model is a long-lasting obstacle in the aviation data value chain, which affects all its steps, from data integration to sharing and analysis. Nevertheless, addressing these issues through the ICARUS data model introduces new challenges related to the sustainability of the approach. Data models, especially in an evolving landscape such as that of the aviation industry, cannot be considered static and therefore updating mechanisms should be foreseen and backwards compatibility ensured. ICARUS foresees a controlled process for updating the data model, however this is a challenging task that cannot be fully automated, as contradictions may arise and a balance should always be kept between the model's expressivity and the complexity it enforces on the processes that rely on it as described in section 2.2.3. As for the common metadata model, the current deliverable to a large extent constitutes a report on how the ICARUS metadata model is designed and used, especially as a brokerage enabler.
Asset Sharing	Data Policy Language	Deliverable D2.2 Section 5.4	Section 2.3.1 in the current deliverable explains how XACML was selected as the data policy language and how it is used for the definition and enforcement of the access policies. With regard to the broader policies embedded into the data contracts, the respective policy language remains open.
Asset Sharing	Data Pricing (and asset pricing in general)	Deliverable D2.2 Section 5.4	Data Pricing was identified as one of the main asset brokerage challenges in D2.2 and was also discussed in the current deliverable as one of the important marketplace dimensions. The ICARUS approach regarding data and non-data asset pricing is reflected in the metadata model. However, as the ICARUS marketplace grows and as the aviation data analysis landscape evolves, new pricing models may be considered for adoption. The current ICARUS approach puts the data provider at the centre of the pricing decision making. Yet recent studies in the field envision aviation smart coins and dynamic pricing schemes that will be based on the financial value created based on a given data asset (MITSloan, 2018), therefore the current pricing model may need to be eventually reconsidered. On a final remark, it should be noted that there is an additional challenge stemming from the adopted pricing model. Specifically, due to the fact that pricing is agreed upon and all foreseen charges paid prior to the contract validation step, per row pricing of a given dataset cannot be combined with the provision of data updates, as this would require to either (a) continuously charge the consumer for the updates and invalidate and re-validate the contract upon each payment or (b) in advance charge the consumer for the

Area	Key Consideration / Open Challenge	Previous Reference/ Discussion	Comments / Status in ICARUS
			complete foreseen dataset (i.e. for rows that are currently not available but the provider states will become available). Neither mechanism is foreseen in ICARUS and therefore this constitutes a limitation, yet not considered severe, of the brokerage framework.
Asset Sharing	Provision means of data	Deliverable D2.2 Section 5.4	The way data are accessed was one of the data brokerage considerations discussed in D2.2 and also constitutes the 9th dimension of the asset marketplaces review presented in this deliverable. ICARUS has decided to provide data assets as downloadable datasets and as resources to be further explored and processed in the secure private spaces.
Asset Sharing	License compatibility analysis	Deliverable D2.2 Section 5.4	ICARUS foresees both the creation and brokerage of assets that constitute combinations of other underlying assets and/or derivative work of previously existing assets. This places license compatibility analysis at the core of the challenges that need to be addressed in regard to asset sharing. Nevertheless, it has been concluded that this cannot be provided as an automated process at this stage, as it would require natural language processing on legal documents, which is not in the scope of ICARUS. This decision has been also documented in D1.3 as part of the decision to exclude the relevant MVP feature. This decision does not imply that license compatibility is not of interest, only that any controlling mechanisms will heavily rely on manual checking and, subsequently, legal procedures when disputes need to be resolved.
Asset Sharing	Liability and accountability	Deliverable D2.2 Section 5.4	As with IPR, licensing and ownership, ICARUS foresees in the metadata model and the contract terms ways to define liability and accountability clauses, but ultimately legal procedures need to be put in place to handle these issues and ICARUS cannot automate this process. Nevertheless, the usage of Blockchain facilitates accountability mechanisms and as the technology becomes more mature and widespread and legal standards start being adopted (blockchainhub, 2019), additional measures could be put in place in ICARUS to facilitate relevant processes.
Asset Sharing	Smart contracts' enforceability, GDPR compliance and security	Deliverable D2.2 Section 5.4	Enforceability and GDPR compliance of smart contracts remain challenging issues in the evolving DLT-enabled data marketplace landscape and addressing them is beyond the scope of ICARUS. However, certain methodology and design decisions have been made to partially address some relevant aspects. The most notable among them is that a legal document written in natural language accompanies each smart contract to allow the definition of terms that cannot be expressed as part of the self-executing smart contracts and which could, if needed, be used in legal procedures. Regarding DLT, and specifically Ethereum, security, the transactions foreseen by the brokerage workflow have been found to be insusceptible to common Blockchain manipulation schemes, therefore adopting additional security mechanisms has not been considered at this point.

Area	Key Consideration / Open Challenge	Previous Reference/ Discussion	Comments / Status in ICARUS
Asset Sharing	Alignment of smart contract and natural language contract terms	Deliverable D2.2 Section 5.4	The reasons for providing hybrid asset sharing contracts which comprise a smart contract and a natural language legal document have been explained. The resulting implication of ensuring the two parts are not contradicting each other or causing ambiguity is extremely challenging. Addressing it would require, at least, natural language processing over legal documents and thus goes beyond the ICARUS scope. However, since both interacting parties (provider and consumer) are certain to have legal departments that handle the asset sharing contracts, ensuring alignment is considered primarily their responsibility.

As a final note, it is worth mentioning that IATA identifies data sharing in general by itself as a challenge and a key aspect for the proliferation of data science in aviation and specifically states that "an open culture towards data pays dividends when being part of a trusted and governed sharing ecosystem". This last sentence is part of the ICARUS mission statement and data sharing is at the core of the ICARUS assets brokerage framework.

5 Conclusions & Next Steps

The present deliverable documents the produced results of the activities performed in the final iteration of all WP2 tasks, namely T2.1 “Data Collection, Provenance and Safeguarding Methods”, T2.2 “Data Curation, Harmonisation and Linking Frameworks”, T2.3 “Deep Learning and Prescriptive Analytics Algorithms” and Task T2.4 “Data Policy and Assets Brokerage Frameworks”. In this regard, the deliverable extends the work performed in D2.1 and D2.2 and is structured on two main axes, linked to the data management methods (which correspond to the T2.1 and T2.2 activities), and the data value enrichment activities, which comprise data analysis (within T2.3) and data and data-enabled assets sharing (in T2.4).

In detail, the “data management methods” axis involves the work presented in section 2 of the current report and its main outcome is the definition of the proposed ICARUS approach regarding: (i) Data Collection that showcases how data populate the ICARUS for the first time (through the data check-in process) or evolve (through the data update process); (ii) Data Curation which includes the definition of a data cleansing workflow, the data provenance model and the mapping and linking methods. Overall, the defined data curation approach addresses all data pre-processing needs and documents how ICARUS shall follow the trails of data assets and actions performed on them; (iii) Data Safeguarding methods that comprise detailed data access control mechanisms based on rules expressed in XACML, data encryption functionalities and data anonymisation processes that ensure data protection from unauthorised access on multiple levels.

The “data value enrichment methods” axis, involves the work presented in section 3 and its main outcome is two-fold:

- The analytics methods to be supported in ICARUS are refined and examined in more detail to provide more actionable information that can be leveraged during the implementation phase and can also help to timely identify potential issues. Furthermore, data analysis perspectives are examined in relation to each demonstrator based on the initially available insights.
- The data marketplaces landscape is revisited, 10 differentiating dimensions are identified and the ICARUS positioning along each one of them is defined. The previously defined data sharing model is extended to include also non-data, but data-enabled, assets and its features are refined. The final version of the ICARUS Data Policy and Assets’ Brokerage Framework is presented in detail and its main workflow, which foresees all stakeholders’ interactions during an asset sharing process, is described.

D2.3 concludes with a broad discussion on several challenges that have been identified in D2.1 and D2.2, as well as new considerations that have emerged, and the ways in which ICARUS foresees to address them. This discussion, along with the aforementioned outcomes, shapes the ICARUS positioning and will continue to serve as a guideline for the activities in other work packages, mainly WP3 and WP4 that are responsible for the design and implementation of the ICARUS platform, through which the ICARUS framework and its workflows will be instantiated.

Annex I: References

ACI, EUROCONTROL & IATA (2017) A-CDM Implementation Manual v5, March 2017, Available online at: https://www.eurocontrol.int/sites/default/files/publication/files/airport-cdm-manual-2017.PDF

ACI (2019) Airport Community Recommended Information Services (ACRIS), Available online at: https://aci.aero/about-aci/priorities/airport-it/acris/

Aeronautical Information Exchange Model (AIXM), Available online at: http://aixm.aero/

Blockchainhub (2019), Smart Contracts, Available online at https://blockchainhub.net/smart-contracts/

EC Joinup (2019) DCAT-AP v1.2.1, May 2019, Available online at https://joinup.ec.europa.eu/release/dcat-ap/121

IATA White Paper (2018), Blockchain in Aviation- Exploring the Fundamentals, Use Cases, and Industry Initiatives, Available online at https://www.iata.org/publications/Documents/blockchain-in-aviation-white-paper.pdf

IATA White Paper (2019), Data Science Hype or Ripe for Aviation?, Available online at https://www.iata.org/events/Documents/Datathon-2019/Data-Science-Hype-or-Ripe-for-Aviation-White-Paper.pdf#__prclt=iSofPoel

IATA (2019) SSIM - Standard Schedules Information, Available online at: https://www.iata.org/publications/store/Pages/standard-schedules-information.aspx

Maletic, J., & Marcus, A. (2000). Data Cleansing: Beyond Integrity Analysis. Iq, 1–10.

MITSloan (2018), Why the Data Marketplaces of the Future Will Sell Insights, Not Data, Available online at https://sloanreview.mit.edu/article/why-the-data-marketplaces-of-the-future-will-sell-insights-not-data/

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

Schomm, F., Stahl, F., & Vossen, G. (2013). Marketplaces for data: an initial survey.ACM SIGMOD Record,42(1), 15-26.

Stahl, F., Schomm, F., Vossen, G., & Vomfell, L. (2016). A classification framework for data marketplaces.Vietnam Journal of Computer Science,3(3), 137-143.

Syncsort (2019), Enterprise Data Marketplaces – Enabling a Data-driven Culture, Available online at https://www.em360tech.com/wp-content/uploads/2019/01/EB_Data-in-the-Marketplace_181207_Letter.pdf

TowardsDataScience (2018), Data Marketplaces: The Holy Grail of our Information Age, Available online at https://towardsdatascience.com/data-marketplaces-the-holy-grail-of-our-information-age-403ef569fffb

UN/CEFACT (2019) Core Component Technical Specification, Version 3.0, Available online at: https://www.unece.org/cefact/codesfortrade/ccts_index.html

Wu, S. (2013) A review on coarse warranty data and analysis. *Reliability Engineering and System, 114*: 1–11.

Annex II: Updated ICARUS Metadata Schema

As defined in D2.1 and taking into consideration the updated DCAT Application Profile for data portals in Europe (DCAT-AP, v1.2.1) that was released in May 2019, as well as the initial experience from the ICARUS beta platform release, an updated ICARUS metadata schema (or the ICARUS Application Profile in the DCMI terminology) has been defined along the following categories:

- **Core Metadata** encapsulating the basic information accompanying a data asset, e.g. a unique identifier for the data asset following specific naming conventions, the title by which the data asset is formally known and a brief (free-text) description of the data asset.
- **Semantic Metadata** referring to semantic annotations for the data asset, as well as its mapping to the ICARUS data model and its linking to other data assets.
- **Distribution Metadata** that provide a better understanding for the availability of a data asset, define its accessible forms and allow for retrieving certain data asset extract (as defined by the data asset provider).
- **Sharing Metadata** shedding light on the rights and the policies associated to a data asset.
- **Trading Metadata** keeping track of the data contracts that have been made and registered in ICARUS.
- **Preservation Metadata** presenting the quality assessment of a data asset, as well as information related to its provenance.

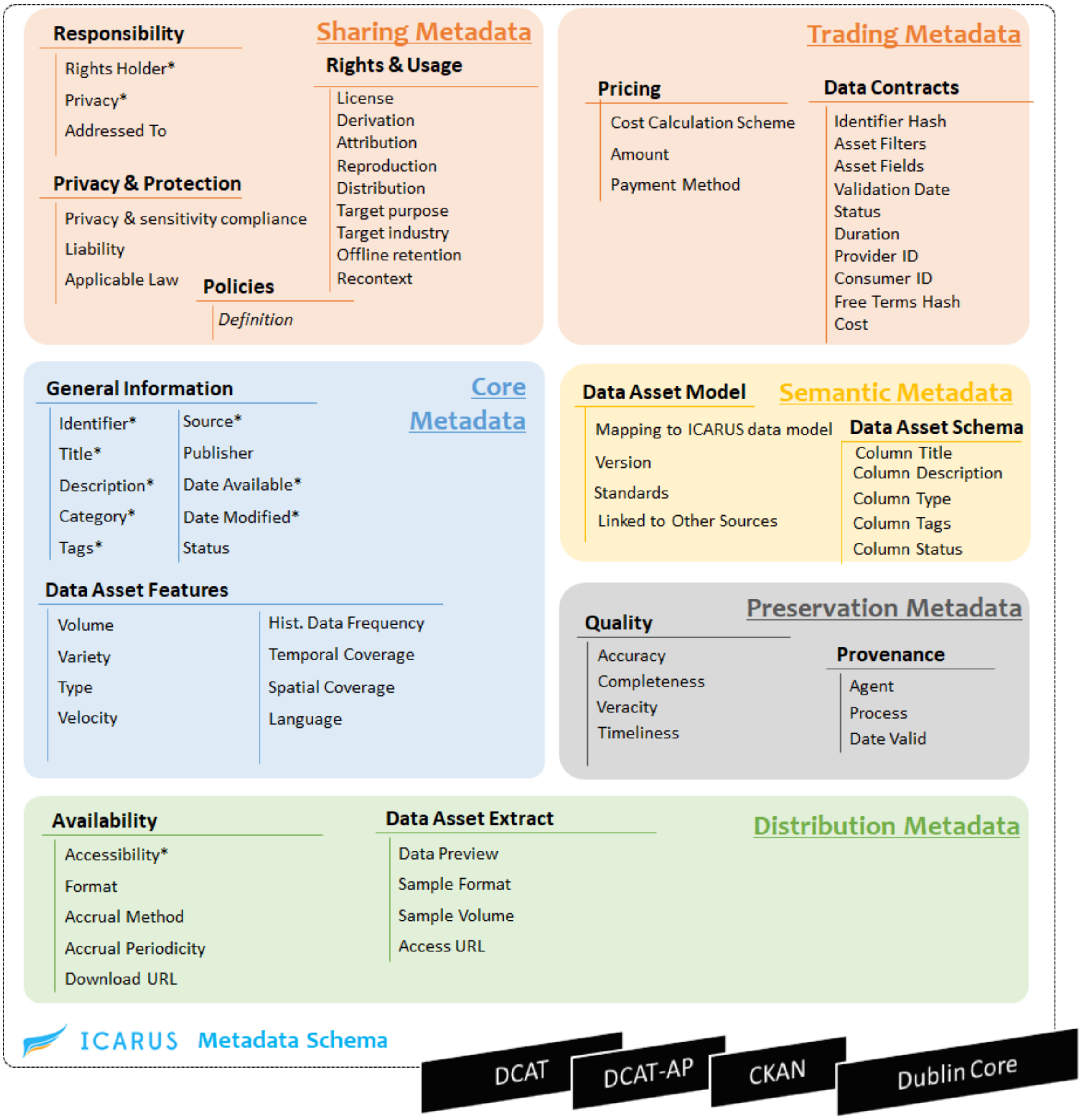


Figure II-1: ICARUS Metadata Schema Overview

Table 0-1: ICARUS Metadata Schema Details

Category		Metadata	URI	Description	Type	Cardinality	Origin
Core Metadata	General Information	**Identifier**	dct:Identifier	An unambiguous reference to a data asset in the context of ICARUS.	Identifier	1..1	DCMI
		Title	dct:Type	The name of the data asset by which it can be easily identified.	Text	1..1	DCMI
		Description	dct:Description	A brief overview that acts as an account of a data asset’s contents.	Text	1..1	DCMI
		Category	skos:Concept	A classification of the data asset to 1st Tier (Primary Aviation), 2nd Tier (Extra-Aviation & Linked Open Data) and 3rd Tier (Aviation-derived Data).	Code	1..1	DCAT-AP
		Tags	icarus:Tags	A list of pre-defined keywords, concepts and / or arbitrary textual tags associated with a data asset.	Text	1..N	CKAN
		Source	dct:Source / foaf:Agent	An entity (e.g. organization, individual or service) from which the data asset originates.	Name	1..1	DCMI
		Publisher	dct:Publisher / foaf:Agent	An entity responsible for making a data asset available in ICARUS.	Name	0...1	DCMI
		Date Available	dcterms:issued	The date when a data asset became or will become available in ICARUS, using an encoding scheme, such as the W3CDTF profile of ISO 8601.	DateTime	1...1	DCMI

Category		Metadata	URI	Description	Type	Cardinality	Origin
		Date Modified	dcterms:modified	The date when a data asset was last changed, using an encoding scheme, such as the W3CDTF profile of ISO 8601.	DateTime	0…1	DCMI
	Features	**Volume**	icarus:Volume	The scale / amount of data within a data asset, e.g. X GBs / records / transactions in total or per hour / day / month.	Measure	1..1	-
		Variety	icarus:Variety	The different forms of the data in terms of being considered as Structured, Unstructured or Semi-structured.	Code	1..1	-
		Type	dct:Type	The nature or genre of the data asset using a controlled vocabulary, such as the IANA Media Types (referring to Text, Image, Video, Audio).	Code	1..1	DCMI
		Velocity	icarus:Velocity	The speed with which the data asset becomes available in ICARUS, i.e. Streaming, Real-time, Near Real-time, Micro-Batch, Batch.	Code	1..1	-
		Historical Data Frequency	icarus:Frequency	The rate at which the historical data have been collected according to a controlled list, i.e. Hourly, Daily, Weekly, Monthly, Yearly, other.	Code	0..1	DCMI / DCAT-AP
		Temporal Coverage	dct:temporal	A named period, date, or date range that the data asset covers.	DateTime / Duration	1..1	DCMI
		Spatial Coverage	dct:spatial	Named places or locations to which the data asset refers, using a controlled vocabulary such as the Thesaurus of Geographic Names [TGN].	Code	1..1	DCMI

Category		Metadata	URI	Description	Type	Cardinality	Origin
		Language	dct:language	The language of the data asset, use a controlled vocabulary such as RFC 4646 / Language of the metadata composed of an ISO639-2/T three-letter language code and an ISO3166-1 three-letter country code.	Code	1..1	DCMI
Distribution Metadata	**Availability**	**Accessibility Method**	icarus:access	The method by which a data asset is accessible to a data consumer, e.g. through API, as a downloadable file, as database extract, other.	Code	1..1	-
		Format	dct:format	The file format of a data asset, using a controlled list, e.g. csv, xml, json, other.	Code	1..N	DCMI
		Accrual Method	dct:AccrualMethod	The method by which up-to-date data are added to the data asset, if applicable.	Text	0..1	DCMI, DCAT
		Accrual Periodicity	dct:AccrualPeriodicity	The frequency with which up-to-date data are added to the data asset, if applicable.	Measure	0..1	DCMI, DCAT
		Download URL	dcat:downloadURL	A URL that is a direct link to a downloadable file in a given format.	Text	1..N	DCAT
	Data Asset Extract	**Data Preview**	icarus:Preview	A description of the sample (even fabricated) extract provided for a data asset.	Text	1..1	VoID
		Sample Format	icarus:sampleFormat	The file format of a data asset sample extract, using a controlled list, e.g. csv, xml, json, other.	Code	1..1	-
		Sample Volume	icarus:sampleVolume	The scale / amount of data within a data asset sample.	Measure	1..1	-

Category		Metadata	URI	Description	Type	Cardinality	Origin
		Access URL	dcat:accessURL	A URL that gives direct access to the distribution of a data asset sample extract.	Text	1..1	DCAT
Sharing Metadata	**Responsibility**	**Rights Holder**	dct:rightsHolder	A person or organization owning or managing rights over the data asset and acting as the data provider.	Name	1..1	DCMI
		Privacy	icarus:privacy	The desired visibility of a data asset, i.e. Confidential (not to be shared), Proprietary/Private (to be shared with appropriate licensing), Public (available to all).	Code	1..1	-
		Addressed To	icarus:audience	The intended audience for responsibility (i.e. individual, group, legal entity)	Text	0..1	DCMI *New*
	Rights & Usage	**License**	dct:license	The legal statement / terms giving official permission to a data asset in each case or on a case-by-case basis, e.g. CC Attribution-NonCommercial-ShareAlike (CC BY-NC-SA), or Bilateral Agreement.	Text	1..1	DCMI
		Derivation	icarus:derivation	An indication whether the creation and distribution of any update, adaptation, or any other alteration of a data asset or of a substantial part of the data asset that constitutes a derivative data asset is allowed, with permissions to modify, excerpt, annotate, aggregate the original data asset.	Text	0..1	- *New*

Category		Metadata	URI	Description	Type	Cardinality	Origin
		Attribution	icarus:attribution	An indication whether it is required to give credit to copyright holder and/or provider	Text	0..1	- *New*
		Reproduction	icarus:reproduction	An indication whether from a given data asset, temporary or permanent reproductions can be created by any means and in any form, in whole or in part.	Text	0..1	- *New*
		Distribution	icarus:distribution	An indication whether restricted or unrestricted publication and distribution of a data asset is allowed.	Text	0..1	- *New*
		Target purpose	icarus:targetPurpose	The intended use that the data provider allows, i.e. for business purposes, for academic purposes, for scientific purposes, for personal purposes, for non-profit purposes.	Text	0..1	- *New*
		Target industry	icarus:targetIndustry	The target industry within and beyond the aviation data value chain stakeholders.	Text	0..1	- *New*
		Offline retention	icarus:offlineRetention	An indication whether storage beyond the ICARUS platform (i.e. local downloading) is permitted.	Text	0..1	- *New*
		Re-context	Icarus:recontext	An indication whether restricted or unrestricted use of a data asset in a different context is allowed.	Text	0..1	- *New*
	Policies	**Definition**	icarus:policies	A set of policies associated with a data asset, according to section 2.3.1.	N.A.	0..N	-

Category		Metadata	URI	Description	Type	Cardinality	Origin
	Privacy & Protection	**Privacy & sensitivity compliance**	icarus:sensitivityCompliance	A set of custom clauses (included in the natural language textual part of the contract if needed), referring to obligations for privacy and sensitivity compliance.	Text	0..1	- *New*
		Liability	icarus:liability	A set of custom clauses (included in the natural language textual part of the contract if needed) referring to the data liability disclaimer and conditions.	Text	0..1	- *New*
		Applicable Law	icarus:applicableLaw	A set of custom clauses (included in the natural language textual part of the contract if needed), including the regulatory framework of the country that is responsible for settlement of any disputes.	Text	0..1	- *New*
Trading Metadata	**Contracts**	**Identifier Hash**	icarus:assetIDHash	A unique identification of the asset in ICARUS that is hashed to avoid being in any way exposed in the blockchain.	Text	1..1	- *New*
		Asset Filters	icarus:assetFilters	Any evaluatable filter on the asset, e.g. in the case of data assets, this could include spatiotemporal coverage based on specific asset columns/fields. Most fields in the core metadata model can be used as valid filters combined with the desired value(s) and/or value range.	Text	0..1	- *New*
		Asset Fields	icarus:assetFields	Applicable in data assets only – the list of the fields of the ICARUS common aviation data model that should be present in the dataset.	Text	0..1	- *New*

Category		Metadata	URI	Description	Type	Cardinality	Origin
		Validation Date	Icarus:validationDate	A timestamp for the different status conditions a contract comes across, e.g. when the contract was drafted and signed by the data provider, when it was signed by the data consumer, when it became effective (i.e. when the payment was performed and confirmed by the data provider).	Text	1..1	- *New*
		Status	icarus:contrStatus	An indication of the status of the contract, e.g. draft, signed, paid, rejected, etc.	Text	1..1	- *New*
		Duration	icarus:contrDuration	The contract duration expressed as dates range.	Text	1..1	- *New*
		Provider ID	icarus:providerID	The ID of the asset provider in blockchain (ethereum address).	Text	1..1	- *New*
		Consumer ID	icarus:consumerID	The ID of the asset consumer in blockchain (ethereum address).	Text	1..1	- *New*
		Free Terms Hash	icarus:termsHash	A hash of the contract part written in natural language.	Text	1..1	- *New*
		Cost	icarus:cost	The price for the acquisition of a data asset as foreseen in the contract including its currency, but not stored in the blockchain.	Text	1..1	- *New*
	Pricing	**Cost Calculation Scheme**	icarus:costScheme	A selection of the applicable cost calculation scheme for a data asset that may range from fixed per row and fixed per asset to request dependent.	Text	0..1	- *New*

Category		Metadata	URI	Description	Type	Cardinality	Origin
		Amount	icarus:amount	The price to purchase a dataset along with its currency.	Text	0..1	- *New*
		Payment Method	icarus:paymentMethod	The applicable payment method that the data provider has defined in order for the payment to be conducted "offline" (outside the platform), e.g. credit/debit card, bank transfer, online payment services.	Text	0..1	- *New*
Semantic Metadata	**Data Asset Model**	**Mapping to ICARUS data model**	icarus:mappingConfiguration	The mapping of the data asset to the ICARUS data model stored offline and used for the transformation of the data.	Text	1..1	- *Updated*
		Version	dct:version	The version of the mapping of the data asset to the ICARUS data model	Text	1…N	DCMI
		Standards	dct:conformsTo	A standard or any other specification to which a data asset conforms.	Text	0..N	DCMI
		Linked to Other Sources	dct:relation	The external data assets to which a data asset is linked (at model / schema level). Note: the related data assets to which a data asset may be linked (at model / schema level) in ICARUS are dynamically provided.	Text	0..N	DCMI
	Data Asset Schema	**Column Title**	icarus:ColTitle	The title of each column included in a data asset's schema as mapped in the ICARUS common aviation data model.	Text	1..N	- *Updated*

Category		Metadata	URI	Description	Type	Cardinality	Origin
		Column Description	icarus:ColDescription	A brief overview that acts as an account of a data asset's column.	Text	1..N	-
		Column Type	icarus:ColType	The nature of a column using a controlled vocabulary.	Code	1..N	-
		Column Tags	icarus:ColTags	A list of pre-defined keywords, concepts and / or arbitrary textual tags associated with a data asset's column.	Text	1..N	-
		Column Status	Icarus:ColStatus	A status indicator to denote whether a column is anonymized and / or encrypted.	Text	1..N	- *New*
Preservation Metadata	**Quality**	**Accuracy**	icarus:Accuracy	A collective assessment/measurement by data consumers within ICARUS of a data asset's correctness and precision, e.g. whether the dataset is error-free.	Value	1..1	-
		Completeness	icarus:Completeness	The degree to which a data asset is sufficient in scope and depth.	Value	1..1	-
		Veracity	icarus:Veracity	The degree to which a data asset is free of bias, using a controlled list, i.e. Raw Data asset, Pre-processed Data asset, Anonymized Data asset, Processed Data asset, Synthetic Data asset.	Code	1..1	-
		Timeliness	icarus:Timeliness	A date or a period range during which a data asset is considered as valid and up-to-date.	DateTime / Duration	0..1	-

Category		Metadata	URI	Description	Type	Cardinality	Origin
	Provenance	**Agent**	icarus:Agent	The agent who is responsible for a specific provenance process to be recorded.	Text	1..N	-
		Process	icarus:ChangeActivity	A log of all changes in the data asset since its initial publication that are significant for its authenticity, integrity, and interpretation.	Text	1..N	-
		Date Valid	icarus:DateChanged	A date or a period range during which a provenance change is valid / happens.	DateTime / Duration	0..1	-

It needs to be noted that, as implied in the previous sections, the data providers may be the sources or the publishers of data assets, yet they are always considered as the IPR holders of the respective data assets, with the exception of the open data.

As noticed in the table above, specific controlled vocabularies (in terms of thesauri, taxonomies and standardised lists of terms) can be extensively used for assigning values in a standardized, homogeneous manner to certain metadata properties. In alignment with the DCAT requirements for Application Profiles, such controlled vocabularies should be published under an open licence; be operated and/or maintained by an institution of the European Union, by a recognised standards organisation or another trusted organisation; be properly documented; have labels in multiple languages, ideally in all official languages of the European Union; contain a relatively small number of terms (e.g. 10-25) that are general enough to enable a wide range of resources to be classified; have terms that are identified by URIs with each URI resolving to documentation about the term; have associated persistence and versioning policies.