Ref. Ares(2018)2305100 - 30/04/2018

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"

D1.1 – Domain Landscape Review and Data Value Chain Definition

Workpackage:	WP1 – ICARUS Data Value Chain Elaboration		
Authors:	University of Cyprus (UCY), Suite5, UBITECH, CINECA, AIA, PACE, ISI, CELLOCK, TXT, OAG		
Status:	Final	**Classification:**	Public
Date:	30/04/2018	**Version:**	1.00

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

ICARUS Project Profile

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

Partners

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

Document History

Version	Date	Author (Partner)	Remarks
0.10	01/02/2018	Loukas Pouis, Dimosthenis Stefanidis, George Pallis, Marios D. Dikaiakos (UCY)	Table of Content (ToC) and partner contribution assignment
0.20	10/03/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Contributions for Section 2
0.21	12/03/2018	Giorgio Pedrazzi (CINECA)	Contributions for Section 2
0.22	13/03/2018	Dimitrios Miltiadou, Konstantinos Perakis (UBITECH)	Contributions for Section 2
0.30	14/03/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Contributions for Section 1, Section 3 and Section 5
0.40	15/03/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Integrated Draft
0.50	16/03/2018	George Pallis, Marios D. Dikaiakos (UCY)	Updated Integrated Draft for Internal Review
0.51	26/03/2018	Fenareti Lampathaki (SUITE5), Haris Zacharatos (CELLOCK)	Internal Review Feedback
0.60	28/03/2018	Dimitrios Miltiadou, Konstantinos Perakis (UBITECH)	Updated Section 2.2 upon Internal Review
0.61	29/03/2018	Giorgio Pedrazzi (CINECA)	Updated Section 2.1 upon Internal Review
0.62	31/03/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Updated Section 1, 2, 3 and 5 upon Internal Review
0.70	27/04/2018	Fenareti Lampathaki (SUITE5)	Complete contribution for Section 4 including all inputs from the demonstrators' partners (AIA in 4.1.1-4.2-4.4, PACE&TXT in 4.1.2-4.2-4.4, ISI in 4.1.3-4.2-4.4, CELLOCK in 4.1.4-4.2-4.4) and OAG (in 4.1.5 and 4.2)
0.71	29/04/2018	George Pallis (UCY)	Review Feedback for Section 4
0.80	30/04/2018	Fenareti Lampathaki (SUITE5)	Updated Section 4
0.90	30/04/2018	Loukas Pouis, Dimosthenis Stefanidis (UCY)	Integrated Section 4 and Final Draft for Submission
1.00	30/04/2018	Dimitris Alexandrou (UBITECH)	Final version submitted to the EC

Executive Summary

The ICARUS project aims to capitalize on the latest Big Data technological breakthroughs in order to provide an infrastructure able to transform, link and analyze interrelated data coming from diverse sectors, delivered in different formats and types.

The work in this deliverable begins by presenting a state-of-play analysis of the big, linked and open data landscape, suggesting in each category the most promising frameworks and tools to be considered. More precisely, the categories considered are Data Collection (data anonymization, data quality, semantic enrichment and annotation), Data Processing and Management (data curation, data linking, data storage, query processing) and Data Analytics (machine learning, deep learning and data visualization). Additionally, EU projects that are relevant to ICARUS are presented. However, this deliverable does not provide an in-depth analysis of the state-of-the-art research methods and methodologies, which will be described in WP2.

Furthermore, it describes the process followed to identify stakeholders that will be potentially interested in the ICARUS outcome and can also benefit from the ICARUS data value chain. After understanding the current state in the field of Big Data by reviewing the key findings of various related industry studies, an early indication of the needs of the prospective ICARUS users is extracted through the conduction of questionnaires with the demonstrators and other stakeholders, as well as with the analysis of the data and other information sources.

The analysis of the questionnaire responses contributed in highlighting the main obstacles and difficulties that stakeholders are currently facing. More specifically, the most difficult processes for organizations are the data anonymization and data linking, while their main concerns for data sharing are privacy/confidentiality and security. Moreover, organizations that use data marketplaces and APIs find it easier to collect data, while organizations that use custom in-house mechanisms find it harder. Additionally, almost 50% of the respondents do not have in place mechanisms for big data, due to budget/cost constraints and lack of experience. What is more important though, is that the majority of the respondents are interested in a data marketplace platform that offers functionalities such as secure experimentation playground for experimenting with datasets before purchasing them, a service that recommends similar datasets with the ones currently explored and a dashboard with interactive visualization capabilities.

Moreover, the deliverable presents a large set of information/data from different sources that would feed the ICARUS data value chain and that will be used in the platform. Finally, a regulatory data sharing framework for data protection, IPR and data sharing is defined in order to be used by the brokerage engine of the platform.

Mostly, this deliverable contributes to Deliverable D1.2 (“The ICARUS Methodology and MVP”) in order to define the ICARUS methodology and value chain definition and formulate the platform’s MVP. The tasks of this Deliverable will be constantly monitored and will be reported in Deliverable D1.3 (“Updated ICARUS Methodology and MVP”), as they remain active until the 15th month of the project. The results of this deliverable will be used not only for the WP1, but also for other WPs: in WP2, to define the main data management, transformation, intelligence extraction and sharing methods that will be supported by the ICARUS platform; in WP3, to help the design of the architecture and of the core features of the ICARUS platform; in WP7, as input to the market analysis to be conducted.

Table of Contents

1 INTRODUCTION 9

1.1 DOCUMENT PURPOSE AND SCOPE 9
1.2 DOCUMENT RELATIONSHIP WITH OTHER PROJECT WORK PACKAGES 10
1.3 DOCUMENT STRUCTURE 11

2 STATE-OF-PLAY IN BIG, LINKED, OPEN DATA 12

2.1 DATA COLLECTION SERVICES 13
2.1.1 Data Anonymization 13
2.1.2 Data Quality 24
2.1.3 Semantic Enrichment and Annotation 30
2.2 DATA PROCESSING AND MANAGEMENT SERVICES 35
2.2.1 Data Curation 35
2.2.2 Data Linking 38
2.2.3 Data Storage 45
2.2.4 Query Processing 52
2.3 DATA ANALYTICS SERVICES 57
2.3.1 Machine Learning 57
2.3.2 Deep Learning 61
2.3.3 Data Visualization 65
2.4 RELEVANT EU PROJECTS 72

3 AVIATION DATA VALUE CHAIN REQUIREMENTS ANALYSIS 74

3.1 ICARUS STAKEHOLDERS AND TARGET AUDIENCE 74
3.2 KEY FINDINGS FROM INDUSTRY STUDIES 77
3.3 ICARUS SURVEY KEY FINDINGS 81
3.3.1 Respondents Profile 81
3.3.2 Data Collection 83
3.3.3 Data Analytics 91
3.3.4 ICARUS Platform 96
3.3.5 ICARUS Survey Key Findings 98

4 AVIATION DATASETS COLLECTION, PROTECTION, IPR AND BROKERAGE 100

4.1 AVIATION DATASETS COLLECTION FROM THE ICARUS CONSORTIUM 100
4.1.1 Demonstrator 1 (AIA): Data Profiling 101
4.1.2 Demonstrator 2 (PACE/TXT): Data Profiling 104
4.1.3 Demonstrator 3 (ISI): Data Profiling 106
4.1.4 Demonstrator 4 (CELLOCK): Data Profiling 108
4.1.5 OAG Aviation Data 111
4.2 INITIAL DATA NEEDS OF THE ICARUS DEMONSTRATORS 115
4.3 AVIATION DATASETS COLLECTION BEYOND THE ICARUS CONSORTIUM 124
4.3.1 Aviation Data Sources 124
4.3.2 Open Data Repositories & Linked Open Data Cloud 129
4.3.3 Other Relevant Data Sources 131
4.4 ICARUS DATA PRELIMINARY ASSESSMENT 132
4.4.1 Data Positioning to ICARUS Data Tiers 132
4.4.2 ICARUS Demonstrators Data High-level Evaluation 134
4.4.3 Initial Data Needs Prioritization in ICARUS 134

4.5 AVIATION DATA PROTECTION AND SHARING 136
4.5.1 Data IPR and Licensing State-of-Play 136
4.5.2 Legal and Regulatory Legislation applicable to the ICARUS Data Tiers 141
4.5.3 Key Considerations for ICARUS Data Sharing 143

5 CONCLUSION 145

6 REFERENCES 147

7 ANNEX 156

7.1 DISSEMINATED QUESTIONNAIRE 156

List of Figures

Figure 1-1: D1.1 Relationship with other Deliverables and Work Packages 11
Figure 2-1: ICARUS Concept 12
Figure 2-2: Mechanisms for Tagging Content [35] 30
Figure 3-1: High-level process to derive ICARUS Stakeholders Requirements 74
Figure 3-2: ICARUS Data Value Chain 75
Figure 3-3: Type of Organization 82
Figure 3-4: Organization Operating Business Domains 82
Figure 3-5: Data Challenges for the Organizations 83
Figure 3-6: Domains of Data Collected by the Organizations 84
Figure 3-7: Data Velocity 85
Figure 3-8: Data Volume 85
Figure 3-9: Data Sources 86
Figure 3-10: Mechanisms for Collecting Data 87
Figure 3-11: Reasons for Collecting Data 87
Figure 3-12: Availability of Data Linking Mechanism 88
Figure 3-13: Main Issues of Collecting Data 88
Figure 3-14: Domains of Data Provided by the Organizations 89
Figure 3-15: Reasons for not Sharing Data 89
Figure 3-16: License of Provided Data 90
Figure 3-17: Format of Provided Data 90
Figure 3-18: Mechanisms for Providing Data 91
Figure 3-19: Security Mechanisms for Data 91
Figure 3-20: Usage of Big Data Analysis Platform 92
Figure 3-21: Issues Preventing Processing/Analyzing Big Data 92
Figure 3-22: Purpose of Data Analytics 93
Figure 3-23: Type of Big Data Processing Frameworks 93
Figure 3-24: Popular Big Data Platforms 94
Figure 3-25: Popular Programming Language for Data Analytics 94
Figure 3-26: Popular Libraries/Frameworks for Data Analytics 95
Figure 3-27: Usage of Collaborative Tools for Data Analytics and Visualization 95
Figure 3-28: Popular Data Visualization Tools 96
Figure 3-29: Interest in ICARUS Functionalities 97
Figure 3-30: Interest in Specific Data Domains 98
Figure 3-31: Concerns of Data Sharing Through an Intermediary 98
Figure 4-1: ICARUS Data Assets Available classified in the tiers of the aviation data value chain 133
Figure 4-2: ICARUS Data Assets needed as classified in the tiers of the aviation data value chain 133
Figure 4-3: Initial assessment of the ICARUS Data Assets needed 135
Figure 4-4: Interrelations among the ICARUS demonstrators regarding the data assets needed (note: the numbers on the arrow signify relevance for other demonstrators, the boxes on data assets depict similarity) 136

List of Tables

Table 2-1: Data Anonymization Software and their main aspects 23
Table 2-2: Data Quality Software and their main aspects 29
Table 2-3: Semantic Enrichment and Annotation Software and their main aspects 34

TABLE 2-4: DATA CURATION SOFTWARE AND THEIR MAIN ASPECTS 38
TABLE 2-5: DATA LINKING SOFTWARE AND THEIR MAIN ASPECTS 44
TABLE 2-6: DATA STORAGE SOFTWARE AND THEIR MAIN ASPECTS 52
TABLE 2-7: QUERY PROCESSING SOFTWARE AND THEIR MAIN ASPECTS 57
TABLE 2-8: TYPICAL MACHINE LEARNING SOFTWARE AND THEIR MAIN ASPECTS 60
TABLE 2-9: DEEP LEARNING SOFTWARE AND THEIR MAIN ASPECTS 64
TABLE 2-10: NOTEBOOK-BASED VISUALIZATION TOOLS AND THEIR MAIN ASPECTS 65
TABLE 2-11: DATA VISUALIZATION SOFTWARE AND THEIR MAIN ASPECTS 71
TABLE 3-1: ICARUS TARGET AUDIENCE 77
TABLE 3-2: RELATED INDUSTRY STUDIES KEY FINDINGS 81
TABLE 3-3: ICARUS SURVEY KEY FINDINGS 99
TABLE 4-1: AIA DATA PROFILING – GENERAL INFO 101
TABLE 4-2: AIA DATA PROFILING – DATA ASSETS FEATURES 102
TABLE 4-3: AIA DATA PROFILING – DATA ASSETS AVAILABILITY & RIGHTS 103
TABLE 4-4: PACE/TXT DATA PROFILING – GENERAL INFO 104
TABLE 4-5: PACE/TXT DATA PROFILING – DATA ASSETS FEATURES 105
TABLE 4-6: PACE/TXT DATA PROFILING – DATA ASSETS AVAILABILITY & RIGHTS 105
TABLE 4-7: ISI DATA PROFILING – GENERAL INFO 106
TABLE 4-8: ISI DATA PROFILING – DATA ASSETS FEATURES 107
TABLE 4-9: ISI DATA PROFILING – DATA ASSETS AVAILABILITY & RIGHTS 107
TABLE 4-10: ISI DATA PROFILING – INDICATIVE DATA ASSET EXTRACT 108
TABLE 4-11: CELLOCK DATA PROFILING – GENERAL INFO 108
TABLE 4-12: CELLOCK DATA PROFILING – DATA ASSETS FEATURES 109
TABLE 4-13: CELLOCK DATA PROFILING – DATA ASSETS AVAILABILITY & RIGHTS 110
TABLE 4-14: OAG DATA PROFILING – GENERAL INFO 112
TABLE 4-15: OAG DATA PROFILING – DATA ASSETS FEATURES 113
TABLE 4-16: OAG DATA PROFILING – DATA ASSETS AVAILABILITY & RIGHTS 114
TABLE 4-17: ICARUS DEMONSTRATORS DATA NEEDS PROFILING – GENERAL INFO 117
TABLE 4-18: ICARUS DEMONSTRATORS DATA NEEDS PROFILING – DATA ASSETS DESIRED FEATURES 124
TABLE 4-19: AVIATION DATA SOURCES PROFILING 126
TABLE 4-20: AVIATION DATA SOURCES PROFILING – DATA ASSETS FEATURES 127
TABLE 4-21: AVIATION DATA SOURCES PROFILING – DATA ASSETS AVAILABILITY & RIGHTS 128
TABLE 4-22: OPEN DATA REPOSITORIES AND SOURCES 131
TABLE 4-23: OTHER RELEVANT DATA SOURCES 132
TABLE 4-24: ICARUS DEMONSTRATORS DATA ASSETS AVAILABLE ASSESSMENT 134
TABLE 4-25: HIGH-LEVEL COMPARISON OF STANDARD DATA LICENSES 139

1 Introduction

The main objective of the ICARUS project is to deliver a novel platform that leverages data, primary or secondarily related to the aviation domain, to help companies and organizations whose operations are directly or indirectly linked to aviation to simultaneously enhance their data reach, as well as share / trade their existing data sources and intelligence, in order to gain better insights, improve their operations and increase passengers' safety and satisfaction. Using methods such as big data analytics, semantic data enrichment and blockchain powered data sharing, ICARUS will address critical barriers for the adoption of Big Data in the aviation industry and will enable aviation-related big data scenarios through a multi-sided platform that will allow exploration, curation, integration and deep analysis of original, synthesized and derivative data, characterized by different volume, velocity and variety, in a trusted and fair manner.

During the first phase of the ICARUS methodology (WP1), the needs and requirements of the market and, in particular, of the aviation industry, are elicited. First of all, a state-of-play analysis is conducted to provide insights on the big data landscape and the latest technological advancements, describing existing products, frameworks and platforms related to the ICARUS concept, regarding data collection, data processing and management, data analytics and visualization. However, an in-depth analysis of the state-of-the-art research methods and methodologies is not presented in this deliverable but will be described in WP2. Furthermore, the needs and requirements of the potential stakeholders of ICARUS, as well as the current barriers and limitations of the aviation industry are extracted through an online survey and the analysis of related industry studies. Finally, an in-depth investigation for representative aviation data has been undertaken with the contribution of the ICARUS demonstrators and the initial requirements and repercussions for data protection, IPR and brokerage are elicited.

1.1 Document Purpose and Scope

Deliverable D1.1 aims to investigate the current landscape of Big, Linked and Open Data, and identify tools and frameworks that can be integrated into the ICARUS services and the platform backbone infrastructure. Additionally, it identifies the stakeholders of the ICARUS project and derives the initial needs of the demonstrators and stakeholders through an online survey and an analysis of selected industry reports. D1.1 also introduces a pool of data sources that practically initiate the ICARUS data value chain, and investigates the emerging data sharing and protection aspects. The work in D1.1 aims to report the initial activities of ICARUS WP1, regarding the context of the tasks T1.1 "State-of-Play in Big, Linked, Open Data", T1.2 "Aviation Data Value Chain Requirements Analysis" and T1.3 "Aviation Datasets Collection, Protection, IPR and Brokerage".

T1.1 includes a state-of-play analysis on existing frameworks, platforms and technologies that can be integrated into the ICARUS services and platform backbone infrastructure. A comparative analysis for the reported as candidate components will be conducted in order to provide the current technological landscape for Big Data that will be considered during the implementation phases, as they would be the ones that better match the demonstrators' and project's needs.

The main aim of T1.2 is to identify the requirements for constructing the ICARUS data value chain and to map the whole set of stakeholders that are potentially interested in and can also benefit from the ICARUS data value chain. More precisely, the objective is to identify the directly and indirectly to ICARUS linked sectors and extract an early indication of the needs of the prospective ICARUS users. This will lead to the definition of preliminary user requirements that will be used as a high-level description of the features that need to be developed to serve these sectors and allow the cross-sector exchange of data, alongside with value added services that will renovate the data management activities of these sectors.

The main objective of T1.3 is the generation of a pool of data coming from the different sources identified that would feed the ICARUS data value chain and that will be used in the platform, identifying how they are related and how they can benefit the whole sector. As such, it is expected to identify and document the major data sets of these sub-sectors and their metadata, alongside with existing standards and semantic models and vocabularies used in these domains, thus creating a database of information sources that will be considered during the data handling activities of the platform. Furthermore, T1.3 aims to identify, monitor and analyze relevant legal and regulatory legislation relevant to the data to be used by ICARUS. This relates both to the business value and IPR handling of data, as well as to any personal or private data that might be used during the implementation of the project. Related to the first case, ICARUS will conduct an investigation on data IPR policies that the integrated to the platform systems support, in order to design and implement a holistic data sharing framework.

1.2 Document Relationship with other Project Work Packages

This deliverable (D1.1 - "Domain Landscape Review and Data Value Chain Definition") is the outcome of the initial activities undertaken in the context of Task 1.1 "State-of-Play in Big, Linked, Open Data", T1.2 "Aviation Data Value Chain Requirements Analysis" and T1.3 "Aviation Datasets Collection, Protection, IPR and Brokerage" which remain active until the 15th month of the project. Figure 1-1 depicts the relationship of Deliverable D1.1 to other Deliverables and Work Packages (WPs) in ICARUS.

Deliverable D1.2 ("The ICARUS Methodology and MVP"), as well as Deliverable D1.3 ("Updated ICARUS Methodology and MVP") will directly use the D1.1 results to define the ICARUS methodology and value chain definition and formulate the platform's MVP.

With the identification of the targeted stakeholders, their initial needs and the key technology findings (described in Section 3), this deliverable (D1.1) will feed the ICARUS deliverables D2.1 ("Data Management and Value Enrichment Methods"), D3.1 ("ICARUS Architecture, APIs Specifications and Technical and User Requirements") and D7.1 ("Initial Project Exploitation Plan–v1").

Furthermore, the identified data sources and the defined regulatory data sharing framework (described in Section 4) will feed the deliverables D2.1 and D2.2 ("Intuitive Analytics Algorithms and Data Policy Framework") for describing the data that will be on the ICARUS platform and define the data exchange policy, as well as D7.1 in order to identify the IPRs of the project's outputs for the market analysis that will be conducted.

Finally, the state-of-play analysis of the big, linked and open data landscape (described in Section 2) will provide valuable inputs to the deliverable D3.1.

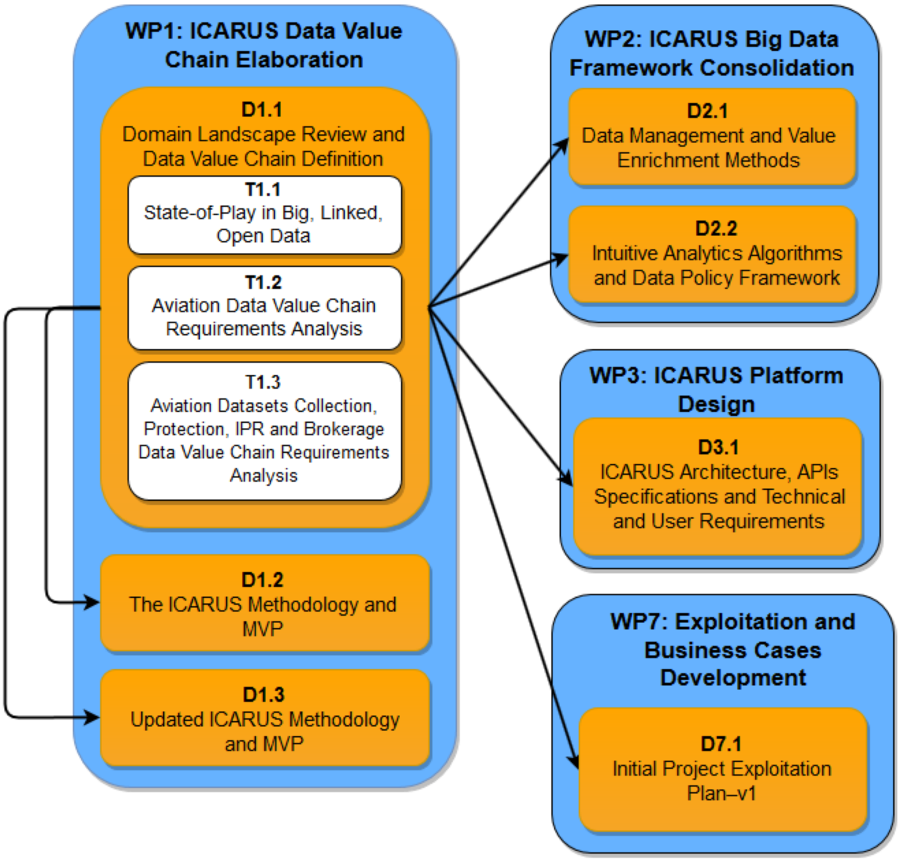

WP1: ICARUS Data Value Chain Elaboration
D1.1
Domain Landscape Review and Data Value Chain Definition
T1.1
State-of-Play in Big, Linked, Open Data
T1.2
Aviation Data Value Chain Requirements Analysis
T1.3
Aviation Datasets Collection, Protection, IPR and Brokerage Data Value Chain Requirements Analysis
D1.2
The ICARUS Methodology and MVP
D1.3
Updated ICARUS Methodology and MVP
WP2: ICARUS Big Data Framework Consolidation
D2.1
Data Management and Value Enrichment Methods
D2.2
Intuitive Analytics Algorithms and Data Policy Framework
WP3: ICARUS Platform Design
D3.1
ICARUS Architecture, APIs Specifications and Technical and User Requirements
WP7: Exploitation and Business Cases Development
D7.1
Initial Project Exploitation Plan–v1

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

1.3 Document Structure

The remainder of this deliverable is structured as follows: Section 2 presents a state-of-play analysis on existing methods, component and technologies related to the ICARUS concept and the three core Data Services Bundles. Particularly, various tools were investigated regarding the Data Collection Services Bundle (data anonymization, data quality, semantic enrichment and annotation), Data Processing and Management Services Bundle (data curation, data linking, data storage, query processing) and the Data Analytics Services Bundle (data analytics, deep learning and data visualization). Furthermore, various EU projects relevant to ICARUS are presented. Section 3 describes the process followed to derive the initial needs of the ICARUS stakeholders. More precisely, it presents the identified project stakeholders and target audience, and documents the key findings derived from the online survey and other relevant industry studies. Section 4 presents a set of data sources that would feed the ICARUS data value chain, while it also presents preliminary perspectives on data protection, IPR and data sharing. Finally, Section 5 concludes this deliverable.

2 State-of-Play in Big, Linked, Open Data

Big Data is revolutionizing all aspects of our lives ranging from enterprises to consumers, from science to government, and of course, the aviation sector could not be excluded. Creating value from Big Data is a multi-step process: data gathering, information extraction and cleaning, data integration, modeling and analysis, interpretation and deployment. Many discussions of Big Data focus only on one or two steps, ignoring the rest.

In each step of the process, there are many open challenges that need to be addressed. A main concern is the privacy and data ownership as there can be inappropriate use of personal data, exposing personal information through linking of data from multiple sources. Furthermore, most data sources are notoriously unreliable and may contain errors, missing values and inconsistencies. For instance, sensors can be faulty, humans may provide biased opinions, remote websites might be stale and so on. Hence, data sources need to be assessed and mechanisms for information extraction and cleaning need to be considered. Another major challenge is the data integration, aggregation, and representation. Frequently, the information collected will not be in a format ready for analysis, as effective large-scale analysis often requires the collection of heterogeneous data from multiple sources and therefore, data need to be transformed and stored in specific data structures. And even then, typical machine learning algorithms cannot handle the high volume of data, hence, there is a need of advanced algorithms that can scale efficiently to address this problem. For big data to fully reach its potential though, there is a need to scale not just for the system, but also from the perspective of humans, as humans need to visualize meaningful information in order to understand the meaningful insights from the data. ICARUS aims to tackle these problems by providing three core data services bundles (Figure 2-1): Data Collection; Data Processing and Management; Data Analytics and Visualization.

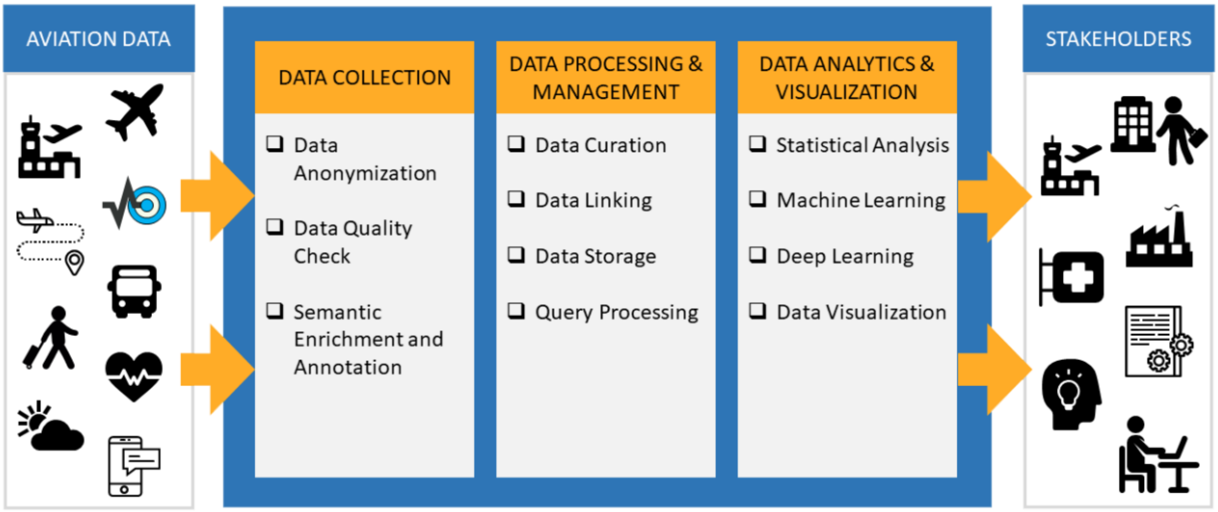


Figure 2-1: ICARUS Concept

Data Collection involves Data Anonymization techniques that tackle the privacy issue, as it provides mechanisms for the hiding the identity of the data origin. Additionally, it includes Data Quality that aims to identify data inconsistencies and assess the reliability of the data source. Semantic Enrichment and Annotation are also considered in this bundle, as it can enhance the content with information about its meaning, enabling users to move quickly to more intelligence-rich information activities.

The second core services bundle, Data Processing and Management, tackles the challenges of filtering and cleaning noisy data (Data Curation) and effectively combining multiple data from different heterogeneous sources (Data Linking). Data Storage and Query Processing are also included in this bundle in order to store data, based on standardization schemes and efficiently retrieve any information required.

The bundle of Data Analytics involves Statistical Analysis and Machine Learning techniques, as well as Deep Learning, aiming to deal with the large-scale data with high dimensionality and extract meaningful insights from the data. Moreover, Data Visualization is also included, as it is a critical asset from the human perspective. The data analysis is useless if the end users (humans) cannot “absorb” the results of the analysis and get lost in a sea of data.

2.1 Data Collection Services

During the Data Collection step, data is accessed, retrieved, collected, transformed, checked, semantically annotated and inserted in the ICARUS data lake verifying the respect of regulations on privacy. Data quality checks are instrumental to ensure the integrity and veracity of the data while data registration needs to be accompanied by the data policy definition (type, terms, frequency of updates, etc.) to ensure compliance with the owner’s IPR.

For the data collection services the Big Data Value Association (BDVA) in its SRIA [1] defines challenges in these topics:

- Data protection and privacy: as part of the data lifecycle, data protection and management must be aligned. Control, auditability and lifecycle management are key for governance, cross-sector applications and the GDPR.
- Data quality: methods for improving and assessing data quality have to be created, together with curation frameworks and workflows.
- Semantic annotation of unstructured and semi-structured data: Data needs to be semantically annotated in digital formats, without imposing extra effort on data producers.

2.1.1 Data Anonymization

Managing data privacy is becoming an increasingly difficult challenge in big data and cloud projects littered with data silos. New data regulations (like GDPR in the European Union, rules for protect broadband consumers in the USA, Cybersecurity law in China, etc.) illustrate that this challenge is just beginning. This trend underscores the importance of anonymization – one of the most important tools in a data scientist’s “privacy toolbox” [2].

To address the problem of data privacy protection, one of the possible solutions is the use of data anonymization techniques [3]. Anonymization can be viewed as a technique to remove an individual’s identifying information from a dataset so that the remaining data cannot be linked to that individual.

In respect to privacy issues in a dataset there are four types of possible specifications for variables (attributes):

- Identifying variables that must be removed from the data set.
- Quasi-identifying (QID) variables are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity so that they can be combined with other quasi-identifiers to create a unique identifier. These variables must be transformed.

- Sensitive variables (or sensitive attribute (SA)) can be kept as-is but they can be protected using privacy models, such as t-closeness or l-diversity.
- Insensitive attributes can be kept unmodified.

The transformed data needs to actually be useful as well. This is called the “privacy vs. utility tradeoff.” If a dataset is perfectly anonymized, there is no risk in identifying an individual from that data, but that data also might (and probably will) be useless. In the cyber security world, there’s a saying that the safest computer is one that won’t function. And here the same point applies: ensuring anonymity usually requires sacrificing utility [2].

With Data anonymization the information that discloses the identity is removed from datasets, so that the people who are defined by the information can remain unknown [3] i.e. sensitive date is de-identified though its format and data type is preserved. Internet is making data more reachable, however most of the data has been limited and the personal identifiable information has been removed.

Data anonymization allows proclamation of entire information that is useful for queries and analysis while maintaining the privacy of sensitive data against diversified types of attacks. Prevailing techniques for data anonymization can be categorized as follows [4] [5]:

1) Data Nature

- Relational data
- Transactional data
- Graphical data
- Unstructured textual data
- Metadata
- Images

2) Objectives of Anonymization

- K-anonymity: the goal in this is to make every record distinct from definite ‘k’ number of records when trying to identifying the record.
- L-diversity: it guarantees L-different values for each group’s sensitive attributes. Thus, an attack can recognize a user’s sensitive information with maximum probability of 1/L.
- T-closeness: the dissemination of sensitive data is accounted for and the dissemination distinction between sensitive data and its values within groups does not exceed T
- Additional objective: it aims for preventing some inferences that are based on presumption that an attacker can have some knowledge.

3) Anonymization approaches

- Generalization: the attribute like age is stiffened into datasets. For example, age is generalized into age ranges.
- Suppression: the attribute like gender is detached from complete dataset.
- Perturbation: addition of noise to the attributes like salary in dataset.
- Permutation: the sensitive linkages between instances are swapped like purchasing medication by an individual.

- Synthetic data approach: it is also used to protect the privacy and confidentiality of a set of data. Synthetic data holds no personal information and cannot be traced back to any individual; therefore, the use of synthetic data reduces confidentiality and privacy issues.
- Pseudonymization: it is a procedure by which the most identifying fields within a data record are replaced by one or more artificial identifiers, or pseudonyms.
- Additional data specific approaches.

In the next paragraphs, a short description of the main methods for different types of data are presented and in various available software are listed.

2.1.1.1 Relational Data anonymization

Data anonymization is the technique for altering the data before being used or published so as to avoid the identification of sensitive attributes. The original format i.e. position, size and type of data, is not lost during anonymization process, and thus, the data appears realistic even for test data environments and can still be processed to get useful information. The anonymization techniques can be divided in two macro categories:

- Syntactic anonymization: it includes K- anonymity, L-diversity anonymous, and T-closeness anonymous.
- Differential privacy: it is a privacy approach with probability of output of two different data sets will nearly be same.

The tools for transforming sensitive personal relational data use selected methods from the broad area of statistical disclosure control. The basic idea is to transform datasets in ways that make sure that they adhere to well-known syntactic privacy models (K-anonymity, L-diversity and T-closeness) that mitigate attacks that may lead to privacy breaches:

1) **K-anonymity**: in current years, K-anonymity, promoting a new definition of privacy, becomes popular. The goal in this is to make every record distinct from definite 'k' number of records when trying to identifying the record.

This method warranties that in series of k groups, the sensitive attributes are unknown which indicates that the probability of identifying a person is less than 1/k and the privacy level is directly depending on size of k. K-anonymity is not only suitable for sensitive attributes since the statistical features of data are suppressed as much as possible. An attacker can escalade a consistency attack or background-knowledge attack for establishing linkage between sensitive and identifiable personal data that leads to privacy breach. Sweeney highlights some limitations of k-anonymity model as follows [3]:

• It doesn't guarantee a "privacy" kind of attack in which the attacker is having background knowledge of the targeted victim to eliminate possible values in a sensitive attribute. Also, it is prone to homogeneity attack, wherein the little diversity in the sensitive attributes can be discovered by an attacker. Thus, L- diversity is generated adopting sturdier definitions of privacy.

• Although identity disclosure is protected by the existing k- anonymity property, attribute disclosure is not protected.

- It protects identifiable attributes, yet it does not safeguard sensitive relationships in datasets.

Personal anonymity requirements are not accounted, and a k-anonymize table can have loss of significant information from the dataset that may be a valued source of information for many purposes (e.g. trend analysis, public fund allocation, medical research etc.). It is mostly apt for the definite sensitive data only, but it may give undesirable information leakage in case of numerical sensitive data values (like salary).

2) **L-diversity**: The L-diversity model is an extension of the K-anonymity model which reduces the granularity of data representation using techniques like generalization and suppression, so that any given record maps onto at least k-1 other records in the data. The L-diversity model handles some of the weaknesses in the k-anonymity model where protected identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group exhibit homogeneity. The L-diversity model adds the promotion of intra-group diversity for sensitive values in the anonymization mechanism.

3) **T-closeness**: T-closeness is a further refinement of L-diversity group-based anonymization that is used to preserve privacy in data sets by reducing the granularity of data representation. This reduction is a tradeoff that results in some loss of effectiveness of data management or mining algorithms in order to gain some privacy. The t-closeness model extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

Differential privacy is a way to protect data in a database maximizing the accuracy of queries from statistical databases while minimizing the chances of identifying its records. It is impossible to publish information from a private statistical database without revealing some amount of private information, while the entire database can be revealed by publishing the results of a surprisingly small number of queries. This problem can be solved introducing randomness in the query response [6]. Other privacy models are also available in specific software (Incognito, δ-disclosure privacy, etc.).

The methods to make data adhere to these theoretical models include, for example, [7].

- **Generalization**: this technique replaces (or records) quasi-identifiers values for less specific values, that are semantically consistent. In this technique, a value is replaced by another more generic, that is faithful to the original. For example, the date of birth could be generalized to a range such as year of birth, in order to reduce the risk of identification;
- **Suppression**: in this technique, the key identifier or the quasi-identifier is deleted to form the anonymized table. It is used in the context of statistical databases, which provides only summaries of the table data instead of individual data;
- **Randomization**: this technique consists of the replacement of the actual data values in order to remove the strong link between the data and the individuals. There are many ways to implement randomization. Some of them are: (i) Noise Addition, which consists of adding random noise to original data; (ii) Permutation, which consists of shuffling the values of attributes in a table so that some of them are artificially linked to different data subjects; (iii) Differential privacy, which adds appropriate noise to the query response;

- **Pseudoanonymization** consists of replacing one attribute (typically a unique attribute) in a record by another. Encryption with secret keys, Hash Functions and Tokenization are the main ways to implement Pseudoanonymization.

- **Packetization** produces non-overlapping groups (or buckets) and then, for each group, releases its projection on the QIDs and also its projection on SAs.

- **Data perturbation** is a procedure that changes data (e.g. by adding noises) to protect privacy. This method is popularly used to protect summary data in statistical analysis, but it is not commonly used in protecting relational data.
- **Synthetic data approach** is also used to protect the privacy and confidentiality of a set of data. Synthetic data holds no personal information and cannot be traced back to any individual; therefore, the use of synthetic data reduces confidentiality and privacy issues.

2.1.1.2 Transactional data

Unlike relational data, transaction data has some unique properties that make its anonymization more difficult. One important property is that transaction data is often high dimensional compared to relational data. That is, considering each item as an attribute, a set of transactions will have a high number of attributes compared to a typical relational dataset. It has been shown that k-anonymization is not useful in high dimensional data because it can significantly destroy data utility; that is because, with high dimensional data, there is a low chance for records to share attribute values, hence more generalization (or distortion) needs to be applied to the data.

Anonymizing transaction data is quite different from K-anonymization of relational data because the data has no well-defined set of quasi-identifiers and sensitive values. Any subset of items in a transaction could play the role of quasi-identifiers for the remaining (sensitive) ones. Another fundamental difference is that transactions have variable lengths and high dimensionality. To protect the privacy of transaction data in such conditions, the K^m-anonymity privacy model has been proposed [8]: given a set of transactions T, no adversary who has background knowledge of up to m items of a transaction can use these items to identify less than k transactions from T.

However, K^m-anonymity have two main limitations [7]:

- Approaches do not support detailed privacy requirements enforcement. For example, in K^m-anonymity, all possible combinations of m items are required to be protected. In real applications, not all items need to be protected. Overprotection could lead to unnecessary loss of data utility. It is desirable that a data publisher can specify, in detail, how data is to be protected.

- Generalization is dependent on a hierarchy, which is not flexible enough as a generalized item has to be a parent node of items that need to be protected. Loukides et al. [9] proposed a constraint-based anonymization method (COAT) to reduce information loss that may have occurred in K^m-anonymization. To protect a set of transactions, COAT allows a data publisher to specify privacy constraints and utility constraints.

2.1.1.3 Graphical Data Anonymization

Anonymizing social network data is much more challenging than anonymizing relational data [10]. First, it is much more difficult to model background knowledge of adversaries and attacks about social network data than that about relational data. On relational data, it is often assumed that a set of attributes serving as a quasi-identifier is used to associate data from multiple tables, and attacks mainly come from identifying individuals from the quasi-identifier. However, in a social network, many pieces of information can be used to identify individuals, such as labels of vertices and edges, neighborhood graphs, induced subgraphs, and their combinations. Second, it is much more difficult to measure the information loss in anonymizing social network data than that in anonymizing relational data. Typically, the information loss in an anonymized table can be measured using the sum of information loss in individual tuples. Given one tuple in the original table and the corresponding anonymized tuple in the released table, it is possible to calculate the distance between the two tuples to measure the information loss at the tuple level. However, a social network consists of a set of vertices and a set of edges. It is hard to compare two social networks by comparing the vertices and edges individually. Two social networks having the same number of vertices and the same number of edges may have very different network-wise properties such as connectivity, betweenness, and diameter. Thus, there can be many different ways to assess information loss and anonymization quality.

There are basically three types of sensitive information that one may want to keep private and may be under attack in a social network environment: node information, link information and edge weight information. The **node information** is the information attached to a vertex. For example, the emails sent by an individual, the personal information such as age, sex, zip code, and transaction data such as purchased items. The link information is about the relationships among the individuals which may be considered sensitive. Links can be used to represent financial exchanges, friend relationships, conflict likelihood, sexual relations, and disease transmission. Depending on the application, the edge weight information can semantically represent “degree of friendship”, “trustworthiness”, and “behavior” etc. If considering routing problem, (for information spread and marketing), edge weights may correspond to the cost of information propagation. To protect edge weight privacy, perturbation-based approaches to preserve linear property, such as modifying all edge weights so that the shortest path remained to be the shortest path, have been proposed recently.

2.1.1.4 Metadata Anonymization

Metadata, also known as data about data, is information that characterizes or gives details on digital media like music, images, movies, Office files, etc. [11]. Metadata provides information about the internal structure of the file. Such information, which is required to extract the content from the binary representation, does not change for a given type of file. For instance, digital cameras insert metadata into each picture they produce: the date, the camera model, the post-processing software used, and even, for some high-end models, the GPS coordinates of the place where the images have been taken. Office documents like PDF or Libre/Microsoft Office generally contains authors, operating system, company information, and even the history of revisions into each document.

These data can also compromise the anonymity and privacy of users in a network context. More disturbingly, some metadata is added during the data acquisition stage of the file creation process, possibly without the user’s knowledge nor agreement.

2.1.1.5 Unstructured Text Anonymization

Like other types of data, text data may also contain identifiers and sensitive information. Therefore, releasing text data (like a patient discharge report or a legal document) could infringe privacy. Protecting privacy in text is challenging because there are semantic relationships among items which an adversary can use to infer additional information [12].

The process of anonymization is important for sharing data without exposing to outsiders any sensitive information contained in documents. An anonymization process that aims at the definitive deletion of sensitive information in text is usually called redacion (when performed by humans), declassification or sanitization. If this information is replaced by a specific label or entry, in order to be later included in the text, the process is called de-identification and the reverse process of inclusion is called re-identification.

One common assumption about relational and transaction data is that the released data does not contain identifier information. This assumption is reasonable for such types of data because they are structured or semi-structured, consequently, it is easy to automatically pinpoint identifier information and remove it before releasing the data. However, this assumption is not reasonable in text data as text is unstructured data and identifier information may exist among items in different forms, which are difficult to identify and remove.

Scrubbing is used to locate and then replace identifier or sensitive information in text. One possible way of detecting sensitive information from the content of these documents is to identify text structures that constitute names or unique identifiers, known as named entities (NE), which represent real entities in the extra-linguistic universe. However, protecting text data by scrubbing is not sufficient because an adversary can still identify an individual by unique combinations of the information left in the text. Like in the case of relational and transaction data, both identity and sensitive information may be disclosed as a result.

With relational and transaction data, generalization has been shown to be a better approach than removing information, in terms of preserving utility while protecting privacy. On the other hand, scrubbing often uses dictionaries or statistical learning techniques that may miss the detection of some identifiers. Thus, it is worth considering if text data can be transformed so that generalization may be applied. Gardner et al. [13] proposed HIDE, a framework to anonymize text in three steps:

1. attributes are extracted from text using a named entity recognizer;
2. a person-centric identifier is used to classify extracted attributes into QID and SA, and as a result, the text data is transformed into relational data;
3. k-anonymity is adopted to anonymize the relational data.

Another alternative to protect privacy in text is to make the overall content more general (text generalization). A typical method of doing this is to identify terms (e.g. nouns and noun phrases) and then generalize them in text by using an ontology.

Cumby et al. [14] treat the protection problem as a multi-class classification problem by proposing the k-confusability privacy model. The model is very similar to k-anonymity in relational data, in which each document is required to be classified into at least k different topics, assuming that both sensitive and non-sensitive topics are known, and each topic is defined as a set of related terms. A document matches a topic when it contains all terms of the topic.

Scrubbing, transformation-then-anonymization and text generalization are different, in terms of how a data publisher wants an output to be based on the purpose of publishing data. In scrubbing, terms are replaced by a code which may not have any meaning or be completely removed from the text. In transformation-then-anonymization, the output is not text itself, but the algorithm generates an anonymous relational or transaction data. In text generalization, it generates a more general text than the original text. However, the same issue remains in these approaches in that they do not consider semantic relationships between the terms in the text. Therefore, an adversary can still infer sensitive information, using the non-sensitive information.

In addition, semantic relationships among the terms in text create a context. Protecting privacy in text without considering its context may not guarantee privacy because an adversary may use the context to narrow down their "guess". This type of attack is called semantic attack. In this case, semantic enrichment software can be used for anonymization, deleting annotated words in the text.

2.1.1.6 Image Anonymization

Data can also come in the form of images. It is important in a video or image to detect faces, car number plates and other image information in various scales and orientations and applies blurring filters to make the information unreadable [15].

Anonymization of personal data can be reached with 6 different approaches

- Blurring
- Pixelation of picture
- Bar mask over eyes
- Negative of photo
- Mask identity – avatar face with same expressions
- Masked face - characteristics of another face is used to transform the face

The process of automatic face de-identification in videos combines face detection, face tracking and face masking. The first step in face de-identification for video though is face detection. Most of the software in this field is commercial but it is possible to use specifically trained deep neural networks using open source platforms.

2.1.1.7 Related software

ARX [16] is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data. The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes. ARX is able to handle large datasets on commodity hardware.

UTD Anonymization Toolbox [17] currently contains 6 different anonymization methods over 3 different privacy definitions:

- Mondrian Multidimensional k-Anonymity
- Datafly
- Incognito
- Incognito with l-diversity

- Incognito with t-closeness
- Anatomy

sdcMicro [18] can be used for the generation of anonymized (micro)data, i.e. for the creation of public- and scientific-use files. In addition, various risk estimation methods are included. Note that the package includes a graphical user interface that allows to use various methods of this package.

Partition_for_Transaction [19] is a top-down anonymization algorithm for set-valued data (or transaction), based on local generalization. It works under k-anonymity constrain.

Apriori_based_Anonymization [20] is a counting tree based data anonymization algorithm for set-valued dataset.

GraphAnon [21] transforming a graph into supergraphs that are resistance to identity and attribute disclosure attacks.

Mat [22] is a toolbox composed of a GUI application, a CLI application and a library, to anonymize/remove metadata.

NLM-scrubber [23] is a new, freely available, HIPAA compliant, clinical text de-identification tool.

MIST [24] The MITRE Identification Scrubber Toolkit (MIST) is a suite of tools for identifying and redacting personally identifiable information (PII) in free-text medical records.

Anonymizer [25] anonymizes images using detection and blurring technology. Software detects faces and car number plates in various scales and orientations and applies blurring filters to make the faces unidentifiable and the number plates unreadable.

Facepixelizer [26] is a privacy image editor with face detection capabilities allowing you to easily blur faces or other sensitive parts of picture.

Name	Last release	Platform	License	API	GUI	Privacy Models	Anonymization approach	Remarks
Relational Data Anonymization								
ARX [16]	28 Jul 2017	Windows, Linux, OSX	Apache License 2.0	Java	YES	K-anonymity L-diversity T-closeness k-map D-disclosure	Generalization Suppression Perturbation Permutation	• It is the state of the art software for relational data. • High scalability and ease of use. • Frequently updated.
UTD Anonymization Toolbox [17]	3 Jan 2012	Windows, Linux	GPL v3	NO	NO	K-anonymity L-diversity T-closeness	Generalization Suppression Incognito Mondrian Anatomy	• Inactive development.
sdcMicro [18]	26 Jan 2018	R Supported platforms	GPL v2	R	Web Interface	K-anonymity L-diversity	Local recoding k-anonymity Numerical rank swapping Noise addition MDAV PRAM Sampling	• Specialized on microdata. • Frequently updated.
Transactional Anonymization								
Partition_for_Transaction [19]	27 Aug 2015	Python Supported Platforms	The MIT License (MIT)	NO	NO	K-anonymity	Local generalization	• Inactive development.
Apriori_based_Anonymization [20]	3 Sep 2015	Python Supported Platforms	The MIT License (MIT)	NO	NO		Apriori-based Anonymization is a counting tree-based data anonymization algorithm for set-valued dataset	• Inactive development.
Graphical Data Anonymization								
GraphAnon [21]	11 Nov 2017	Linux	The MIT License (MIT)	NO	NO	k-degree-anonymous		• One of the few available to treat graph data. • Active development.
Meta Data Anonymization								

Name	Last release	Platform	License	API	GUI	Privacy Models	Anonymization approach	Remarks
Mat [22]	3 Jan 2016	Linux	GPL v3	NO	YES		Metadata Fields anonymization	• Development is currently on hold.
Unstructured Text Anonymization								
NLM-scrubber [23]	9 Aug 2016	Linux, Windows	Free no open	NO	NO		Scrubbing	• It is free but not open source. • It is inactive. • Can be replaced by NER software.
MIST [24]	25 Aug 2014	Linux, Windows, MacOS	BSD license	NO	NO		Replacement	• It is inactive. • Can be replaced by NER software
Image Anonymization								
Anonymizer [25]	30 Mar 2017	Windows, Linux	Comm.	YES	NO		Image detection and blurring	• Only commercial
Facepixelizer [26]	2015	All platforms	Comm.	NO	WebApp browser computation		Face detection and substitution	• Only commercial

Table 2-1: Data Anonymization Software and their main aspects

2.1.2 Data Quality

2.1.2.1 Data Quality Assessment

The purpose of data quality assessment is to identify data errors and erroneous data elements and to measure the impact of various data-driven business processes [27]. Both aspects are critical. Data quality assessment can be accomplished in different ways, from simple qualitative assessment to detailed quantitative measurement. Assessments can be made based on general knowledge, guiding principles or specific standards. Data can be assessed at the macro level of general content or at the micro level of specific fields or values. The purpose of data quality assessment is to understand the condition of data in relation to expectations or particular purposes or both and to draw a conclusion about whether it meets expectations or satisfies the requirements of particular purposes. This process always implies the need also to understand how effectively data represents the objects, events and concepts it is designed to represent.

The word dimension is used to identify aspects of data that can be measured and through which data's quality can be described and quantified. As high-level categories, data quality dimensions are relatively abstract. The dimensions explored in the DQAF include completeness, validity, timeliness, consistency and integrity. Data quality dimensions are important because they enable people to understand why data is being measured.

Aspects of data quality include:

- Accuracy
- Completeness
- Update status
- Relevance
- Consistency across data sources
- Reliability
- Appropriate presentation
- Accessibility

The Data Quality Metric (DQM) formula usually needs to build data quality indexes or weights, which depend on specific business scenarios It is essential to define their own weights for data quality metrics [28].

2.1.2.2 Approaches to Data Quality

Manual checking of some properties and constraints of the data is of course possible but not affordable in the long run, an automatic approach is needed. A possible solution could be to implement project specific controls and checks of the data. This has the obvious advantage to integrate quality checks in different parts of the application. There are however several drawbacks:

- Development time consumption: developers have spent time on the design, development and test the controls.
- Code readability: both the application specific code and data quality one will reside in the same code base.
- Hard and costly maintainability: additional checks will require a new release of the software.

- Application-specific checks: every application will have its own controls, maybe redundant on some data input. Data quality checks can't be executed in isolation without the related application (or it can be done with additional overhead in the development phase).

A better solution is to have a generic Data Quality Framework, able to read many data sources, easily configurable and flexible to accommodate different needs.

2.1.2.3 Data Quality in Big Data Scenario

In today's data intensive society, Big Data applications are becoming more and more common. Their success stems from the ability to analyze huge collections of data opening up new business perspectives. Devising a novel, clever and non-trivial use case for a given collection of data is not enough to guarantee success. Data is the main actor in any big data application, therefore it's of paramount importance that the right data is available and the quality of such data must meet certain requirements.

One of the main targets of a big data application is to extract valuable business knowledge from the input data**.** Such a process does not involve a trivial computation of summary statistics from the raw data. Furthermore, traditional tools and techniques cannot be applied efficiently to the huge collections of data that are becoming commonplace.

The most common data quality issues observed when dealing with Big Data can be best understood in terms of the key characteristics of Big Data – Volume, Velocity, Variety, Veracity, and Value as reported by Anmol Rajpurohit in KDnuggets [29].

- **Volume**: in the traditional data warehouse environment, comprehensive data quality assessment and reporting was at least possible (if not, ideal). However, in the Big Data projects the scale of data makes it impossible. Thus, the data quality measurements can at best be approximations. It is important to re-define most of the data quality metrics based on the specific characteristics of the Big Data project so that those metrics can have a clear meaning, be measured (good approximation) and be used for evaluating the alternative strategies for data quality improvement. Despite the great volume of underlying data, it is not uncommon to find out that some desired data was not captured or is not available for other reasons.
- **Velocity**: the pace of data generation and collection makes it hard to monitor data quality within a reasonable overhead on time and resources (storage, compute, human effort, etc.). So, by the time data quality assessment completes, the output might be outdated and of little use, particularly if the Big Data project is to serve any real-time or near real-time business needs. In such scenarios, it is needed to re-define data quality metrics so that they are relevant as well as feasible in the real-time context. Sampling can help to gain speed for the data quality efforts, but this comes at the cost of a bias (which eventually makes the end result less useful) because of the fact that samples are rarely an accurate representation of the entire data. Another impact of velocity is that it is needed to do data quality assessments on-the-fly.
- **Variety:** one of the biggest data quality issues in Big Data is that the data includes several data types (structured, semi-structured, and unstructured). Thus, often a single data quality metric will not be applicable for the entire data and you would need to separately define data quality metrics for each data type. Moreover, assessing and improving the data quality of unstructured or semi-structured data is way more tricky and complex than that of structured data. Data from different sources often has

serious semantic differences. This problem is made worse by the lack of adequate and consistent meta-data from each data source.

- **Veracity:** the data might have some inherent impreciseness and uncertainty. Besides data inaccuracies, Veracity also includes data consistency (defined by the statistical reliability of data) and data trustworthiness (based on data origin, data collection and processing methods, security infrastructure, etc.). These data quality issues in turn impact data integrity and data accountability. While the other V's are relatively well-defined and can be easily measured, Veracity is a complex theoretical construct with no standard approach for measurement. In a way this reflects how complex the topic of "data quality" is within the Big Data context. Data users and data providers are often different organizations with very different goals and operational procedures. Thus, it is no surprise that their notions of data quality are very different. In many cases, the data providers have no clue about the business use cases of data users. This disconnect between data source and data use is one of the prime reasons behind the data quality issues symbolized by Veracity.
- **Value:** organizations are harnessing Big Data for many diverse business pursuits, and those pursuits are the real drivers of how data quality is defined, measured, and improved. A common and old definition of data quality is that it is the "fitness of use" for the data consumer. This means that data quality is dependent on what you plan to do with the data. Thus, for a given data two different organizations with different business goals will most likely have widely different measurements of data quality. This nuance is often not well understood – data quality is a "relative" term. A Big Data project might involve incomplete and inconsistent data; however, it is possible that those data quality issues do not impact the utility of data towards the business goal. In such a case, the business would say that the data quality is great (and will not be interested in investing in data quality improvements). The Value aspect also brings in the "cost-benefit" perspective to data quality – whether it would be worth to resolve a given data quality issue, which issues should be resolved on priority, etc.

Data quality in Big Data projects is a very complex topic, where the theory and practice often differ. In practice, data quality does play an important role in the design of Big Data architecture. All the data quality efforts must start from a solid understanding of high-priority business use cases, and use that insight to navigate various trade-offs to optimize the quality of the final output.

A continuous Data Quality check on input, intermediate and output data is therefore strongly advisable. Some open source solutions are described in Table 2-2.

2.1.2.4 Related software

Griffin [30] is a Data Quality Service platform built on Apache Hadoop and Apache Spark. It provides a framework process for defining data quality model, executing data quality measurement, automating data profiling and validation, as well as a unified data quality visualization across multiple data systems. It tries to address the data quality challenges in big data and streaming context.

Apache Griffin is model driven solution, user can choose various data quality dimension to execute his/her data quality validation based on selected target data-set or source data-set (as the golden reference data). It has corresponding library supporting it in back-end for the following measurement:

- Accuracy - Does data reflect the real-world objects or a verifiable source
- Completeness - Is all necessary data present

- Validity - Are all data values within the data domains specified by the business
- Timeliness - Is the data available at the time needed
- Anomaly detection - Pre-built algorithm functions for the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset
- Data Profiling - Apply statistical analysis and assessment of data values within a dataset for consistency, uniqueness and logic.

OpenRefine [31] (formerly Google Refine) is a standalone open source tool suitable for data curation, cleaning and transformation. While it is similar to spreadsheets applications, it behaves more like a database. More specifically, it operates on rows of data which have cells under columns, which is very similar to relational database tables. OpenRefine goes beyond basic transformations by offering a set of advanced cell transformations, as well as the ability to create transformation scripts that can be applied to multiple datasets. It also supports dataset linking and extension via external web services, for example geocoding addresses to geographic coordinates.

Data Quality [32] is a framework developed by Agile Lab. Compared to typical data quality products, this framework performs quality checks at raw level. It doesn't leverage any kind of SQL abstraction like Hive or Impala because they perform type checks at runtime hiding bad formatted data. Hadoop is mainly unstructured data (files), so quality checks are performed at row level without typed abstractions. With DQ you are allowed to:

- Load heterogeneous data from different sources (HDFS, DB etc.) and various formats (Avro, Parquet, CSV, etc.)
- Apply SQL queries on Sources (powered with spark Dataframe API)
- Select, define and perform metrics on DataFrames
- Compose and perform checks
- Evaluate quality and consistency on data, determined by constraints.
- Perform trend analysis, based on previous results.
- Transform results in order to make reports that you like.
- Save results on HDFS in multiple formats (csv, avro, parquet) or/and datastore etc.

Talend Open Studio for Data Quality [33] is an open source data profiling tool that's ready to download and free to use. With Talend Open Studio for Data Quality, it is possible to evaluate current data quality and identify strengths and shortcomings. With a user interface, this data profiler enables users to:

- Easily connect to and drill down into a wide range data sources including databases, packaged applications, and varied file formats.
- Generate a rich variety of data profile statistics, from simple record counts, to analyses of text fields and numeric fields, to building of frequency tables that show how often different values occur.
- Test data for conformance to internal business rules and external standards such as correct syntax for email addresses, international postal codes, and credit card numbers.

Name	Last release	Platform	License	API	GUI	Features	Remarks
Griffin [30]	7 Nov 2017	Spark	Apache License 2.0	Livy server	NO	• Automatic quality validation of the data. • Data profiling and anomaly detection. • Data quality lineage from upstream to downstream data systems. • Data quality health monitoring visualization. • Shared infrastructure resource management.	• Model driven data quality solution for modern data systems. It provides a standard process to define data quality measures, execute, report, as well as a unified dashboard across multiple data systems. • Big data ready. • Active community.
OpenRefine [31]	18 Nov 2017	Java supported platforms	Google 2010	YES (Python, R)	YES (web)	• Remove duplicate records. • Separate multiple values contained in the same field • Analyze the distribution of values throughout a data set. • Group together different representations of the same reality. • Extensions allow the identification of concepts in unstructured text (NER) and can also reconcile data with existing knowledge bases.	• Active community. • Not big-data ready.
DataQuality [32]	2 Mar 2018	Linux (Scala)	GPL v3	YES (scala)	YES	• Load heterogeneous data from different sources (HDFS, etc.) and various formats. • Select, define and perform metrics with different granularity. • Compose metrics and perform checks on them. • Evaluate quality and consistency of data, defining constraints and properties, both technical or domain dependent. • Save check results and historical metrics on multiple destinations (HDFS, MySQL, etc.).	• DQ is a framework to build parallel and distributed quality checks on big data environments. • Big data ready. • Active community.
Talend Open Studio for Data Quality [33]	19 Jan 2018	Linux, Windows, MacOS	Apache License 2.0	YES	YES	• Fraud pattern detection using Benford Law. • Advanced statistics with indicator thresholds. • Column set analysis.	• Open source version lacks features and has limitations. • Lack of support for Linux/Unix.

Name	Last release	Platform	License	API	GUI	Features	Remarks
						• Advanced matching analysis • Time column correlation analysis.	• Lack of support for NoSQL databases.

Table 2-2: Data Quality Software and their main aspects

2.1.3 Semantic Enrichment and Annotation

Semantics is often used in combination with terms such as enrichment, tagging, markup, indexing, fingerprinting, classification, and categorization. Although there can be important distinctions among these terms, they tend to be used loosely and interchangeably [34].

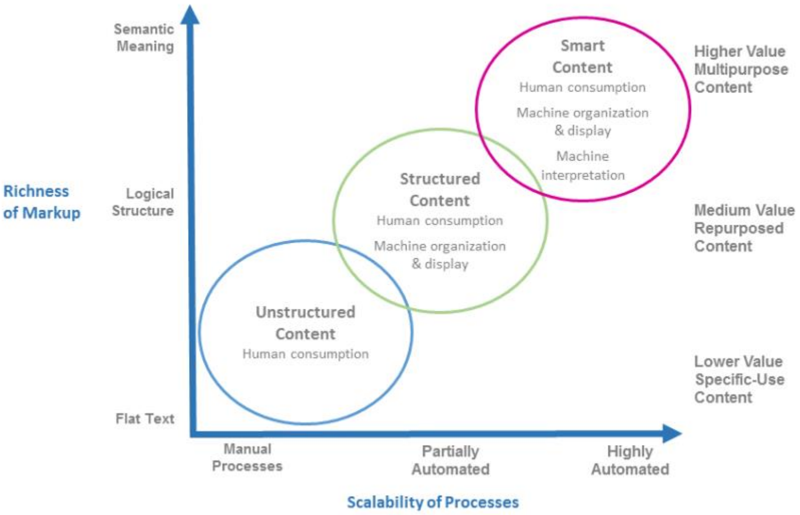


Figure 2-2: Mechanisms for Tagging Content [35]

Semantic enrichment is the process of adding a layer of topical metadata to content so that machines can make sense of it and build connections to it. The addition of semantic metadata to content is also called semantic tagging. A variety of technologies, methods, and practices can be used to enrich content with semantic metadata: tagging can be embedded directly in XML files or can be held externally in databases or content-management systems that reference elements in the content. For multimedia content, such as videos and images, tagging can be placed in metadata headers.

Mechanisms for tagging content vary from fully manual to fully automated (Figure 2-2). In manual tagging, a person who has the appropriate expertise (a domain expert) reads the content and applies tags; this process is sometimes referred to as semantic indexing. Manual tagging is the best solution when a high degree of precision of tagging is required. In automated tagging, software analyzes content, adding tags on the basis of concept matching, statistical patterns, and linguistic analysis. Most automated systems include a “training” and “evaluation” phase during which humans compare the algorithms used for tagging to increase the level of precision and accuracy that can be achieved through automation. Automated tagging is highly scalable and sometimes is the only option for very large content sets. However, automated approaches can lead to false positives (incorrect applications of a tag), missed concepts, and other inaccuracies. An automated process can be followed by manual review and modifications to improve the reliability of tags (hybrid process).

Semantic enrichment can be done at different levels of granularity in content. Tagging should be just granular enough to “atomize” content at the appropriate and useful level. Tagging can be done at the “top”, for example, at the article level. The right level of granularity will depend on how tagged results are used. Topic classification tagging is one example of top-level semantic tagging. Tagging can also be applied deeper within a work; some systems tag major sections of a work, tables, and figures. Some go even deeper, tagging at the paragraph or even the sentence level. Named-entity recognition (NER) is a granular form of semantic tagging that is used to identify predefined entities, such as persons, places, companies, clinical trials, drug names, gene sequences, and proteins. Another type of semantic tagging is the extraction of concepts and their linking to Wikipedia or Domain ontologies (LOD based semantic tagging). For NER and LOD based semantic tagging it is possible to use domain specific training sets, thesauri or ontologies. Keller reports recent advances on ontologies for aviation introducing ontologies as an alternative type of data model to be compared and contrasted with the conventional, UML-based aviation data models that have been under development by government and industry over the past decade [36].

2.1.3.1 Named entities recognition

Named-entity recognition (NER) is a subtask of information extraction that allows to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

NER methods include linguistic grammar-based techniques, supervised, unsupervised and weakly supervised machine learning, heuristics and hand-crafted rules. The different methods can vary significantly in terms of precision and recall, but also in the time and effort required to create a NER system. A very important feature for a NER system is the ability to recognize previously unknown entities [35]: while early studies were mostly based on handcrafted rules, NER is moving towards supervised machine learning as a way to automatically induce rule-based systems or sequence labelling algorithms starting from a collection of training examples, with conditional random fields being a typical choice. Recent advancement in NER are linked to the use of Deep Learning architectures [37].

2.1.3.2 LOD-based Methods for Semantic Enrichment

There are a number of state-of-the-art methods for semantic annotation and linking to DBpedia (e.g. DBpedia Spotlight, YAGO and MusicBrainz). These LOD-based entity-linking approaches have their roots in methods that enrich documents with links to Wikipedia articles. In addition, commercial web services such as AlchemyAPI, OpenCalais, and Zemanta are relevant. A recent evaluation of all state-of-the-art LOD-based methods and tools, showed that DBpedia Spotlight and Zemanta have the best accuracy on annotating texts with the corresponding URIs from DBpedia.

GERBIL[1] is a general entity annotation system based on the BAT-Framework (Blackbox Automated tests). GERBIL offers an easy-to-use web-based platform for the agile comparison of annotators using multiple datasets and uniform measuring approaches. To add a tool to GERBIL, all the end user has to do is to provide a URL to a REST interface to its tool which abides by a given specification. The integration and benchmarking of the tool against user-specified datasets is then carried out automatically by the GERBIL platform. Currently, the platform provides results for 9 annotators and 19 datasets with more coming. Internally, GERBIL is based on the Natural Language Programming Interchange Format (NIF) and provides APIs for datasets and annotators to NIF.

2.1.3.3 Topic Classification of documents

Supervised text classification is currently a challenging research topic, particularly in areas such as information retrieval, recommendation, personalization, user profiles etc. The most popular text classification methods are Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio, and K-Nearest Neighbors (kNN). These methods, using BOW (Bag of Words) for text representation, suffer the lack of semantics in text representation

[1] http://www.aksw.org/Projects/GERBIL.html

and in the rest of the classification process; they ignore all semantics included in the original text that can be deployed in text classification. Nevertheless, it’s possible to replace, in these methods, the classical BOW by BOC (Bag of concepts) through "conceptualization" that enriches document representation model using semantic resources to extract named entities or concepts. Most of the general purpose scientific software (like R or Python) include add-ons to perform supervised classification on text.

2.1.3.4 Related software

SpaCy [38] features new neural models for tagging, parsing and entity recognition. The models have been designed and implemented from scratch specifically for spaCy, to give you an unmatched balance of speed, size and accuracy. A novel bloom embedding strategy with subword features is used to support huge vocabularies in tiny tables. Convolutional layers with residual connections, layer normalization and maxout non-linearity are used, giving much better efficiency than the standard BiLSTM solution. Finally, the parser and NER use an imitation learning objective to deliver accuracy in-line with the latest research systems, even when evaluated from raw text.

GATE [39] is a Java suite of tools originally developed at the University of Sheffield beginning in 1995 and now used worldwide by a wide community of scientists, companies, teachers and students for many natural language processing tasks, including information extraction in many languages. GATE includes an information extraction system called ANNIE (A Nearly-New Information Extraction System) which is a set of modules comprising a tokenizer, a gazetteer, a sentence splitter, a part of speech tagger, a named entities transducer and a coreference tagger. ANNIE can be used as-is to provide basic information extraction functionality or to provide a starting point for more specific tasks.

OpenNLP [40] is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as language detection, tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing and coreference resolution. These tasks are usually required to build more advanced text processing services.

Stanford CoreNLP [41] provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.

It includes:

- An integrated NLP toolkit with a broad range of grammatical analysis tools
- A fast, robust annotator for arbitrary texts, widely used in production
- A modern, regularly updated package, with the overall highest quality text analytics
- Support for a number of major (human) languages
- Available APIs for most major modern programming languages
- Ability to run as a simple web service

Cogcomp-NER [42] tags entities with either the "classic" 4-label type set (people / organizations / locations / miscellaneous), and entities with a larger 18-label type set (based on the OntoNotes corpus). It uses gazetteers extracted from Wikipedia, word class models derived from unlabeled text, and expressive non-local features.

OpeNER [43] is a language analysis toolchain helping (academic) researchers and companies make sense out of "natural language analysis". It consists of easy to install, improve and configure components to:

- Detect the language of a text
- Tokenize texts
- Determine polarization of texts (sentiment analysis) and detect what topics are included in the text.
- Detect entities named in the texts and link them together.

DBpedia-spotlight [44] is a tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia.

Gerbil [45] is a general Linked Data benchmarking system (formerly used for entity annotation systems based on the BAT-Framework). GERBIL offers an easy-to-use web-based platform for the agile comparison of annotators using multiple datasets and uniform measuring approaches.

Name	Last release	Platform	License	API	GUI	Function	Remarks
SpaCy [38]	22 Feb 2018	Python supported platforms	The MIT License (MIT)	Cython	NO	NER and other linguistic tools	• Easy Deep Learning integration. Trainable for different languages. • Active community
GATE [39]	9 Jun 2017	Java supported platforms	Creative Commons	Java	YES	NER and other linguistic tools	• Trainable for different languages. • Active community
OpenNLP [40]	21 Dec 2017	Java supported platforms	Apache License 2.0	Java	NO	NER and other linguistic tools	• Includes rule-based and statistical named-entity recognition. • Trainable for different languages. • Active community
Stanford CoreNLP [41]	31 Jan 2018	Java supported platforms	GPL v3	Java	NO	NER	• Trainable for different languages. • Active development
Cogcomp-NER [42]	18 Feb 2018	Java supported platforms	Research and Academic Use License Cognitive Computation Group	Java	NO	NER	• Based on the OntoNotes corpus. • Trainable for different languages. • Active community
OpeNER [43]	18 March 2017	Java supported platforms	Apache License 2.0	Java	NO	NER, Concept extractor and other linguistic tools	• Trainable for different languages. • Active community
DBpedia-spotlight [44]	3 Nov 2017	Java supported platforms	Apache License 2.0	Java	NO	Concept extraction	• Annotation based on DBpedia resources. • Trained in different languages.
Gerbil [45]	12 Feb 2018	Java supported platforms	GNU AFFERO v3				• GERBIL offers an easy-to-use web-based platform for comparison of annotators. • Active development

Table 2-3: Semantic Enrichment and Annotation Software and their main aspects

2.2 Data Processing and Management Services

Under the prism of the novel aviation-driven data value chain of the ICARUS project and during the Data Collection step, a variety of data will be collected. This data is highly heterogeneous, received in multiple formats, at different velocities and is in most of the times multilingual and unstructured. As a consequence, in the step of "Data Curate and Link", several challenges are introduced concerning the curation, the linking, the storage of the data as well as the effective query execution. Thus, it is crucial to identify the techniques and existing frameworks and tools that will allow the data within the ICARUS data lake to be properly processed and managed towards the assurance that the provided datasets originating from several data sources will produce the maximum value for the ICARUS data value chain.

In this section the main aspects of the data processing and management services are described by presenting the state-of-play techniques and tools in data curation, data linking, data storage and query processing. The analysis conducted is focusing on the main functionalities of each tool and the ease of integration within the ICARUS project.

2.2.1 Data Curation

One of the key principles of data analytics is that the quality of the analysis is dependent on the quality of the information analyzed. The increasing availability of open data on the web, driven by the emergence of new platforms creating data in a decentralized manner such as sensors and mobile platforms, as well as the increase of the available data sources within organizations [46] brings up the problem of managing an unprecedented volume of data. Due to the nature of these data sources, data is created within different contexts and with different requirement, adding the problem of data variety in addition to the data volume.

Data curation provides the methodological and technological data management support to address data quality issues, maximizing the usability of the data. Data curation can be defined as the active management of data over its life cycle to ensure it meets the necessary data quality requirements for its effective usage [47]. Data curation activities enable data discovery and retrieval, maintain quality, add value, and provide for reuse over time [48]. Eventually, data curation emerges as a key data management process nowadays due to the increase in the number of data sources and platforms for data generation. The main goal of the data curation is to enable more complete and high-quality data-driven models that facilitate the reuse of data in different contexts and reduce the barriers of generating high quality analysis.

The major activities to be named under data curation are the following: (a) data cleaning, (b) data presentation, (c) data description, (d) data evaluation, (e) data publication, and (f) data access, use and security. Although there is a variety of tools in the field promising effective data curation, only a small portion is addressing all the major activities listed above. With this in mind, in the following paragraphs the tools that are currently widely used for data curation are described, focusing on the main functionalities in the context of a big data ecosystem.

OpenRefine [31] (formerly Google Refine) is a standalone open source tool suitable for data curation, cleaning and transformation. While it is similar to spreadsheets applications, it behaves more like a database. More specifically, it operates on rows of data which have cells under columns, which is very similar to relational database tables. OpenRefine goes beyond basic transformations by offering a set of advanced cell transformations, as well as the ability to create transformation scripts that can be applied to multiple datasets. It also supports dataset linking and extension via external web services, for example geocoding addresses to geographic coordinates. With OpenRefine cleaning operations like transformations, facets and clustering can be easily executed towards the aim of cleaning the data structure. OpenRefine also supports transformation of data to various formats, as well as normalizing and de-normalizing operations. Open Refine is actively supported by an open source community and is offered with an easy installation and configuration process. OpenRefine supports a variety of data files format as input like CSV, XML, JSON, Excel, RDF as XML and RDF N3 triples or fetch a data file via URL, however it lacks direct integration with relational databases. One major drawback is that the tool is not optimized for large datasets and there are limitations for very large files. OpenRefine comes with a broad list of extensions supported by the community offering additional functionalities to the tool making OpenRefine a powerful tool towards the data curation and data cleaning goal.

Datacleaner [49] is a data cleaning tool by the newly created business division of Neopost, namely the Quadient. It supports data cleansing, transformations, enrichment, deduplication, matching and merging for performing data profiling, data wrangling and data quality operations, as well as low-level data analytics. It provides an easy-to-use graphical interface in order to define and perform transformations and data analytics. Through a list of plug-ins and adapters, Datacleaner is integrated with Hadoop and Spark, as well as Pentaho. It supports many data source types including common file formats, like CSV and Excel files, Relational Databases (RDBMs) and NoSQL databases. The commercial version of Datacleaner provides additional support for more data source types, more data storage types (SQL, NoSQL) and integration with additional big data storage platforms.

Trifacta Wrangler [50] is a data curation and cleaning tool that is facilitating data exploration, transformation and enrichment in order to provide clean and structured data. Trifacta Wrangler is a connected desktop application capable of transforming and preparing data ready to be used for data analytics and visualizations by using machine learning and parallel processing. Additionally, it supports data export to various tools and platforms like Hadoop, Tableau or MS Excel. Recently a community edition was released, however with limited functionalities and the lack of Hadoop integration.

Talend [51] provides a suite of tools for data preparation, data management, master data management, data integration and data quality. The suite consists of a list of open-source as well as commercial tools. The Talend Open Studio for Data Integration is an open source ETL tool include in this suite with an Eclipse-based developer tooling and job designer, versioning support and with a wide range of connectors for RDBMS, SaaS, packaged applications and technologies. Talend Open Studio for Data Preparation provides cleansing and enrichment functions, import and data combining from any Excel and CSV file, export to Tableau as well as auto-discovery, standardization, profiling, suggestion and visualization functionalities. Both tools provide a well-documented API.

WinPure [52] is a data cleansing, matching and duplication removal commercial tool. The tool is primarily focusing on basic data cleansing and duplication. The tool provides a graphical interface and an API with Visual Basic and C# support. The major drawbacks of the tool are the lack of available transformation options and that it is not optimal for big data volumes.

Factual/Drake [53] is text-based command line data workflow tool that organizes command execution around data and its dependencies. With Drake the data processing steps are defined along with their inputs and outputs as workflows. These workflows can be defined using its own specific format, Drake document, or through the APIs provided. While focusing on resolving data dependencies automatically, it also provides a rich set of options for controlling the workflow with user-defined steps or with integration of other tools. It supports multiple inputs and outputs and has HDFS and Amazon S3 support built-in.

The following table summarizes the main functionalities, describes the integration aspects, documents the license information and provides a list of remarks for these data curation tools.

Name	Functionalities	Integration	License	Remarks
OpenRefine	• Data cleaning, and transformation • Reconciliation with web services • Transformation scripts can be generated and applied to multiple datasets • Easy-to-use interface • Supports for rollbacks	• Standalone desktop Java application • Easy installation and configuration • HTTP API available	BSD	• Active community • Not big-data ready
DataCleaner	• Data cleaning and transformations • Supports data profiling, data wrangling and data quality operations • Supports low-level data analytics • Easy-to-use interface • Supports integration with Hadoop, Spark and Pentaho • Support for most popular SQL and NoSQL databases (commercial version) • Native support for big data storages (commercial version)	• Standalone desktop Java application • Can be used as a command-line tool also • Easy installation and configuration	Open source version under LGPL, Commercial version	• Lack of functionalities in the Open source version
Trifacta Wrangler	• Data cleaning and transformations • Data enrichment through recommendations • Support for scripts (recipes) that can be reused in multiple datasets • Big data - ready with support for Hadoop • Integration with Tableau	• Standalone application • Easy to install via installer (Windows/OSX)	Commercial	• Open source version lacks of features and has limitations • Lack of API • Lack of support for Linux/Unix
Talend	• Data Integration and • Preparation; • Supports transformations and cleansing operations; • Easy-to-use interface; • Supports creation and execution of complex data workflows;	• The suite consists of several tools • Each tool included in the suit can be installed and configured easily • Talend Open Studio for Data Integration and Data	Open source version under Apache license v2.0, Commercial version available.	• Open source version lacks of features and has limitations • Lack of support for Linux/Unix • Lack of support for NoSQL databases

Name	Functionalities	Integration	License	Remarks
	• Support for most prominent RDBMs;	Preparation are the best candidates.		
WinPure	• Data cleansing • Data duplication • Available API with C #/VB only	• Tool is available in Windows only as a standalone tool	Commercial version	• Limited set of features • Not big-data ready • Easy-to-use for simple tasks
Factual/Drake	• Supports data workflow creation and execution • HDFS and Amazon S3 native support	• Java standalone command-line tool for Windows, Linux, OSX	Eclipse Public License	• Inactive community • No native support for SQL or NoSQL databases

Table 2-4: Data Curation Software and their main aspects

2.2.2 Data Linking

In the current Web of Documents, query processing is performed via keyword search over unstructured web pages lacking of semantics. Semantic Web is the emerging extension of the Web of Documents into distributed Web of Data [54]. The purpose of the Web of Data is to publish structured information on the web and interlink this information with other data sources. In particular, the Web of Data contains open and interlinked data which can be reused and shared. However, the Web of Data practically shares many characteristics with the Web of Documents. As on the Web of Documents, the quality is highly varying on the Web of Data [55]. It is very common that an entity may be described and published with different identifiers from different publishers, introducing barriers preventing the interlinking of the data towards the creation of comprehensive information that can be used during the query processing.

Towards this end, Linked Data was introduced as an initiative to publish, share and connect data in the Semantic Web with aim of transforming the web of documents into the web of interlinked data. Linked Data is a way to construct a global data space, through machine readable data format that interconnects structured data sources via typed links [56]. The basic principles of Linked data [57] are the following:

- Use URIs for naming things on the web
- Use HTTP URIs to name things so that internet users can look up these names
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs, so that they can discover more things

Although the concept of Linked Data is evolving over the last years and the number of structured and semi-structured data available is growing, there is a lack of discovery techniques and tools to properly address the semantic linking between related instances. In general, data linking can be described as the task of determining whether two entity descriptions can be linked one to the other in the case both descriptions refer to the same real-world object within a specific domain or if there is an acknowledgeable relationship between them. The construction of the identity link that determines the data linking of the entities is referred as instance matching. To perform instance matching, three basic techniques are followed [58]:

- **Value matching**: These techniques are focused on the identification of the equality of the property values of the two entity descriptions.
- **Individual matching**: These techniques are focused on the description of two different entities with multiple property values and by utilizing the results of the value matching technique on the property values, an aggregation of similarities is performed.
- **Dataset matching**: These techniques are focused on all the individuals of the datasets under investigation and try to construct an optimal alignment between these whole set of individuals. These techniques receive as input the results of individual matching and introduce several refinements by applying optimization algorithms, similarity propagation and more.

In the course of providing Linked Data, several steps must be followed which includes data source transformation and semantic enrichment, in order to provide the structuring of data in a format suitable for automatic processing and data linking. In the following paragraphs the tools that perform actions towards the aim of providing Linked Data are described.

Apache Marmotta LDClient [59] is an independent Linked Data Client library provided by the Apache Marmotta platform which converts data in various formats into RDF that can be used with RDF tools and for data linking. It is rather flexible and modular, providing the necessary infrastructure for remote resource retrieval via different protocols and comes with a number of pre-defined data providers for different services that can be wrapped as Linked Data resources.

Virtuoso Sponger [60] is part of the Virtuoso open source platform (see also 0) and is the Linked Data middleware capable of generating Linked Data from a variety of data sources while also supporting a wide variety of data representation and serialization formats. Sponger is integrated into Virtuoso's SPARQL Query Processor where it delivers URI de-referencing within SPARQL query patterns, across disparate data spaces. Sponger is enabling generation of quality linked data from unstructured or semi-structured data sources. Sponger is also offered as a full-fledged HTTP proxy service, directly accessible via SOAP or REST interfaces.

morph-RDB [61] (formerly called ODEMapster) is an RDB2RDF engine following the R2RML specification, supporting data upgrade by generating RDF data from a relational database in accordance to the R2RML mapping descriptions and query translation of SPARQL queries over the RDF datasets into SQL based on the R2RML mapping descriptions. morph-RDB performs a variety of optimizations during query rewriting process and it is currently supporting RDBMS like MySQL, PostgreSQL and MonetDB.

LODRefine [62] is based on the OpenRefine (see also section 2.2.1) supplemented with various integrated extensions of OpenRefine that facilitate transition from tabular data to Linked Data. The list of integrated extensions contains RDF extension, DBpedia extension, Crowdsourcing extension, Stats extension. LODRefine is the LOD-enable version of OpenRefine, with purpose data curation, semantic enrichments and data linking via RDF vocabularies.

Silk [63] is an open source framework for integrating heterogeneous data sources and maintaining the links. Silk is facilitating the link generation between related data items within different Linked Data sources. RDF links from the datasets of the publishers to other open data sources on the Web can be set using Silk. Silk defines a link specification language: Silk Link Specification Language (Silk-SLSL). Using Silk-SLSL publishers can define the RDF link types that should be discovered between data sources and the conditions that should be met in order to perform interlinking. The tool uses various string matching methods, but also numeric equality, date equality, taxonomical distance similarity, and sets similarity measure and matching algorithms can be combined. The data sources that will be interlinked can be accessed via the SPARQL protocol either locally or remotely and the link specification can be created via the graphical interface of the tool or can be set in XML. Silk also offers data transformations to structured data sources via lightweight transformation rules mainly for data cleaning, mapping between properties and conversions between various data formats.

LIMES [64] is a link discovery framework for the Web of Data. Limes implements time-efficient approaches for large-scale link discovery and its originality lies in the usage of the properties of metric spaces. Using different approximation techniques in order to compute estimates for the similarity of instances, is able to reduce the number of comparisons needed during the mapping process by several orders of magnitude by utilizing the estimates to rule out the matching pairs that do not fulfil the mapping conditions. Limes is implementing a large list of algorithms for the processing of structured data, as well as supervised and unsupervised machine-learning algorithms, including the supervised, active and unsupervised versions of EAGLE and WOMBAT, for accurate link specification discovery. Lime can be used as a standalone tool or as a Java library and can be easily configured via a configuration file or through the graphical interface.

CSV2RDF4lod [65] is a simple and easy-to-use tool for producing RDF encoding of data available in Comma-Separated-Values (CSV). The tool can handle tabular data from well-structured RDBMS while also being able to aggregate and integrate multiple versions of multiple datasets of multiple source organizations in an incremental and backward-compatible way.

Anything to Triples (Any23) [66] is a library, a web service and a command line tool that extracts structured data in RDF format from a variety of Web documents. In the latest version it supports the following input formats:

- RDF, Turtle, Notation 3
- JSON-LD which is a lightweight Linked Data format based on the JSON format that provides a way to help JSON data interoperate at Web-scale
- RDFa with RDFa1.1 prefix mechanism
- HTML5 Microdata
- Microformats1 and Microformat2: hAdr, hCard, hCalendar, hEntry, hEvent, hGeo, hItem, hListing, hProduct, hProduct, hRecipie, hResume, hReview, License, Species, XFN
- YAML
- Extraction support from a variety of vocabularies like Dublin Core Terms, XHTML, Description of a Career, Description of a Project, GEO Names, VCard, Friend of a Friend, ICAL, BBC Programmes Ontology, Open Graph Protocol, RDF Review Vocabulary and schema.org

D2RServer [67] is part of the D2RQ platform and is a tool that enables RDF and HTML browsers to navigate the content of the database while also enabling query execution via the SPARQL query language. D2RServer maps the database content using its own custom mapping language (D2RQ) that allows to browse and query the RDF data. More specifically, the requests from the Web are rewritten into SQL queries using the custom mapping language. By performing on-the-fly translation the contents of the database can be published as RDF without the need to generate RDF and store it in a dedicate RDF triple store. D2RServer offers browsing of the database content, semantic enrichment of the entities with resolvable URIs, SPARQL endpoints and explorer as well as publishing metadata.

Sparqlify [68] is a SPARQL-SQL rewriter that enables defining RDF views on relational databases and query them with SPARQL. The tool is currently in alpha state and powers the Linked-Data Interface of the LinkedGeoData Server. Sparqlify rewrites SPARQL queries into a single SQL statement giving all control over query planning to the underlying database system. For the moment, only the PostgreSQL database system is supported with support of geo-spatial functions.

LinDA [69] is a complete open-source suite of Linked Data tools facilitating the mapping and publishing of linked data, data linking with private and public data, as well as analysis and visualization of the Linked Data produced. It includes a lightweight transformation to linked data tool, a vocabulary repository, a conversion tool from RDF to conventional data structures, a query designing and processing tool and a linked data visualization and analytics tool.

Table 2-5 compiles the necessary information (the supported sources formats, the main functionalities, the semantic enrichment capability, the integration aspects, the license information and a list of remarks) for these data linking tools.

Name	Supported source formats	Functionalities	Semantic Enrichment	Integration	License	Remarks
Apache Marmotta LDClient	RDF, XML, HTML, Freebase API, Facebook API, YouTube API, Vimeo API, RDFa, MediaWiki API, PHPBB, LDAP	• Retrieve resources via various protocols • Retrieve Linked Data resources from various data providers • Conversion to RDF • Limited ontology mapping support	Yes	• Java Library that can be integrated • Can be extended	Apache v.2 License	Active community
Virtuoso Sponger	• Variety of data sources including in structured and semi-structured formats	• SPARQL query processing • Ontology mapping • Support for R2RML • Full-fledged HTTP proxy service (SOAP, REST)	No	• Not standalone, Part of the Virtuoso hybrid server	GNU General Public License (GPL) Version 2	Active community
Morph-RDB	• RDBMS • CSV files	• Generate RDF data from relational DB • Support for R2RML • SPARQL to SQL query translation	Yes	• Java Library that can be integrated	Apache v.2 License	Active community
LODRefine	• Any tabular data	• Based on OpenRefine • Transformation of tabular data to linked data • Data curation • Semantic enrichment • Data linking via RDF vocabularies	Yes	• Standalone application • Easy installation and configuration • Minimal requirements (Java/JRE)	BSD	Inactive community
Silk	• SPARQL endpoints • RDF	• Link generation between related data items with different linked data sources	No	• Java standalone application • Extensible via plug-in	Apache v.2 License	Active community

Name	Supported source formats	Functionalities	Semantic Enrichment	Integration	License	Remarks
	• XML • CSV	• Set RDF links from the user's data to external data sources • Access data sources through SPARQL endpoints • Silk Link Specification Language (Silk-SLSL • Supports data transformations to structured data sources		development		
LIMES	• SPARQL endpoints	• Time-efficient approaches for large-scale link discovery • Implements machine learning algorithms like EAGLE and WOMBAT	No	• Java standalone application or Java Library that can be integrated	GNU Affero General Public License v3.0	Active community
CSV2RDF4lod	• CSV files	• Easy-to-use command line tool • RDF generation • Aggregation of datasets in an incremental and backward-compatible way	No	• Command-line tool	Apache License, Version 2.0	Active community
Any23	• RDF, Turtle, Notation 3 • JSON-LD • RDFa with RDFa1.1 prefix mechanism • HTML5 Microdata • Microformats1 and Microformat2 • YAML	• Extracts structured data in RDF format	No	• Written in Java • Offered as a Java library, web service and command line tool	Apache License, Version 2.0	Active community

Name	Supported source formats	Functionalities	Semantic Enrichment	Integration	License	Remarks
D2RServer	• RDBMS	• Enables RDF and HTML browsers to navigate the content of the database • Query execution via the SPARQL query language • Custom mapping language (D2RQ) • Rewrites SPARQL queries to SQL • On-the-fly translation the contents of the database can be published as RDF • SPARQL endpoint publication	Yes	• Written in Java • Standalone application	Apache License, Version 2.0	Inactive community
Sparqlify	• RDBMS (PostgreSQL) • CSV	• Rewrites SPARQL queries into a single SQL statement • enables defining RDF views on relational databases	No	• Written in Java	No information provided.	Active community
LinDA	• CSV, RDBMS	• mapping and publishing of linked data • data linking with private and public data • conversion from CSV/RDBMS to RDF	Yes	• Standalone tool • Docker image available	MIT License	Inactive community

Table 2-5: Data Linking Software and their main aspects

2.2.3 Data Storage

In every Big Data ecosystem, one of the core challenges is the problem of data storage. Data storage addresses the need for storing and managing data in a scalable way, satisfying the needs of the rest of the components of the system that require access to the data. Thus, there is no doubt that data storage holds a key role in every system and the requirements linked to data storage are more than crucial for the overall success of the system. In the case of Big Data projects, the ideal big data storage system would allow storage of a virtually unlimited amount of data, cope both with high rates of random write and read access, flexibly and efficiently deal with a range of different data models, support both structured and unstructured data, and for privacy reasons, only work on encrypted data. In general, characteristics like scalability, high availability, consistency, security, flexibility and efficiency are those that describe a big data storage system. Obviously, all these needs cannot be fully satisfied. However, over recent years, many new storage systems have emerged embracing different storage technologies that at least partly address these challenges.

Big data storage technologies are referred to as storage technologies that in some way specifically address the volume, velocity, or variety challenge and do not fall in the category of relational database systems. This does not mean that relational database systems do not address these challenges, but alternative storage technologies such as columnar stores and clever combinations of different storage systems, e.g. using the Hadoop Distributed File System (HDFS), are often more efficient and less expensive [70].

In the case of big data storage, the volume challenge is typically addressed by utilizing techniques based on distributed architectures. In this way, scalability is achieved by introducing new nodes offering additional storage and computational power to the system in order to address possible increased storage requirements. Using specialized techniques, the new nodes are seamlessly added to the existing storage cluster and the storage system takes care of distributing the data between individual nodes transparently.

In addition to the volume challenge, the velocity and variety of data challenges should be addressed. Concerning the velocity challenge, the storage solution should be highly performant to handle multiple concurrent requests with high performance querying and indexing capabilities and it should be able also to provide input/output operations per second (IOPS) necessary to deliver data to analytics tools that will be used. The variety challenge relates to the level of effort that is required to integrate and work with data that originates from a large number of different sources and with diverse formats. The storage solution it should be able to cope with large amount of unrelated and complex data originating from various heterogeneous data sources.

2.2.3.1 Data Storage Technologies

Over the last decade, the rapid increase in the amount of generated data created the need to deal with the data explosion [71] and in conjunction with the hardware shift from scale-up to scale-out approaches led to an explosion of new big data storage systems that shifted away from traditional relational database models. It should be noted that these approaches are mainly focusing on properties such as speed, efficiency and handling unstructured and semi-structure data at large, while sacrificing other properties such as data consistency.

The following list assesses the different type of storage solutions:

- **Relational DBMS:** Relational database management systems support the relational (table-oriented) data model. The schema of a table is defined by the table name and a fixed number of attributes with fixed data types. A record corresponds to a row in the table and consists of the values of each attribute. It supports classical set operations (union, intersection and difference), selection operations (selection of a subset of records based on defined filter criteria), projection operations (selection of a subset of attributes of the table) and join operations (special conjunction of multiple tables as a combination of the Cartesian product with selection and projection). Popular examples are Oracle [72], MySQL [73], Microsoft SQL Server [74] and PostgreSQL [75].

- **NoSQL Databases**: The most important and widely used category nowadays. NoSQL databases are designed with scalability in mind, sacrificing the level of consistency. Comparing to relational databases the transactional properties of atomicity, consistency, isolation and durability (ACID) are not necessarily honoured. NoSQL databases usually utilize data models different from the relational databases which includes Key-Values Stores, Columnar Stores, Documents and Graph structures. Popular examples are MongoDB [76], HBase [77], Cassandra [78], Blazegraph [79], Neo4j [80] and OrientDB [81].

- **Distributed File Systems**: Another popular and widely used category. File systems such as Hadoop File System (HDFS) [82] are utilized to store large amount of unstructured data in a reliable way, capable for coping with bulk processing and quickly ingesting large data files. HDFS which is an integral part of the Hadoop Framework [83] is currently considered a de-facto standard.

- **NewSQL Databases**: These are the modern form of relational databases aiming for the scalability property of NoSQL databases and the transactional guarantees of relational databases. The main characteristics of NewSQL databases [84] are the following: (a) SQL is the primary mechanism for application interaction, (b) ACID support for transactions, (c) a non-locking concurrency control mechanism, (d) an architecture providing much higher per-node performance, (e) a scale-out, shared-nothing architecture, capable of running on a large number of nodes without suffering bottlenecks. NewSQL databases are expected to be about 50 times faster than RDBMS. The most popular solution is VoltDB [85] which provides ACID with linear scaling in the case of non-complex (single partition) queries.

- **Big Data Querying Platforms**: Big Data query platforms provide query facades on top of underlying big data stores. Their purpose is to simplify as much as possible querying on the underlying data stores usually via an SQL-like query interface in order to achieve low query latencies. An example is Hive [86] (or similarly Apache Pig [87]) that allows structured files in HDFS to be queried via an SQL-like query language by translating the composed queries in MapReduce jobs in Hadoop. One other benefit of Hive is the flexibility on the schemas evolvement by adapting the schema-on-read approach. Another example is Impala [88] which is designed low latency execution of queries on top of HDFS.

In the following paragraphs, the data storage solutions that are currently widely used for data storage are described, focusing on their storage type and main functionalities.

Oracle Database is probably the most popular enterprise relational database management system. It is a database management system supporting multiple data models, ideal for online transaction processing (OLTP), data warehousing (DW) as well as mixed (OLTP and DW) database workloads. It offers on-premises, on-cloud and hybrid-cloud deployment options. It comes with a variety of editions and releases depending on the user needs and support for all major operating systems. Oracle database offers a variety of features in the availability, scalability, analytics, performance, security, management, development and integration areas. The latest release, Oracle Database 18c, added new functionalities and features like multitenant architecture, in-memory column store, native database sharding, enhanced database performance, availability and security.

MySQL is relational database management system utilizing a pluggable storage engine architecture and offering multiple storage engines like InnoDB, Memory, Merge, MyISAM and Archive. MySQL is offering two version, an open-source version and licensed version. MySQL provides a wide range of features like cross-platform support, stored procedures, triggers, cursors, updatable views, performance schema, ACID compliance, SSL support, query caching and full-text indexing.

Microsoft SQL Server is a relational database management system with the primary function of storing and retrieving data as requested by other applications supporting different workloads, from small single machine applications to large scale applications. It supports ACID properties and has an easy to use interface with support for all major programming languages.

PostgreSQL is an open-source relational database. It is an ACID-compliant and transactional database with emphasis on extensibility and standards compliance. PostgreSQL provides full support for foreign keys, joins, views, triggers, and stored procedures while also provides native interface for most popular programming languages. PostgreSQL provides a list of sophisticated features such as Multi-Version Concurrency Control (MVCC), asynchronous replication, advanced query planner and optimizer. It supports various workloads, from small to large scale applications. It is very well known for the great support for geospatial data with the PostGIS extension.

MongoDB is an open-source cross-platform document-oriented NoSQL database. MongoDB is designed to offer horizontal scalability, high availability and flexibility. MongoDB supports efficiently large unrelated complex and nested data. MongoDB also provides high performance querying and indexing capabilities with the support of dynamic schemas, where each document can be indexed with primary and secondary indices. It supports load balancing while also ensuring accessibility and high availability without compromising advance data protection. Moreover, MongoDB also offers intra-cluster network compression for better performance.

Originating form Google's Bigtable, **HBase** is an open-source, non-relational distributed database currently under the umbrella of Apache Foundation. HBase runs on top of HDFS providing a fault-tolerant way of storing large quantities of sparse data while also offering faster read and write operations with high throughput and low input/output latency. HBase provides in-memory operation, compression and can serve as input for MapReduce jobs in Hadoop.

Cassandra is an open-source NoSQL distributed database management system designed to operate across several commodity servers. Cassandra is capable of handling large amount of data in a distributed way, while also offering high availability with no single point of failure. The architecture of Cassandra is following a masterless design providing advanced clustering capabilities with clusters spanning across multiple data centers with asynchronous replication resulting in low latency operations, easy scale-out and operational simplicity. It is fault-tolerant as data are automatically replicated to multiple nodes or multiple data centers. Cassandra is a highly performant database with linear scalability.

Blazegraph is an open-source high performance, scalable graph database with support for Blueprints and RDF/SPARQL APIs. Blazegraph supports extensions for durable named solutions sets, efficient storage and querying, as well as scalable graph analytics. It also supports full-text search and integration with Apache Solr and Apache Tinkerpop besides SPARQL query processing. Blazegraph supports multi-tenancy, can be deployed as standalone server or as a highly available replication cluster. Moreover, it can be deployed as a horizontally-sharded federation of services.

Neo4j is a graph database management system. Neo4j is an ACID-compliant transactional database with native graph storage and processing. Neo4j is built around performance, scalability and high availability. Neo4j is designed to store, query, analyse and manage highly connected data efficiently while providing optimized mechanisms to work with complex data, analysis and use-cases. Neo4j was designed to leverage not only data but also its relationships, with support for highly-scalable real-time queries of relationships using the declarative graph query language Cypher.

OrientDB is an open source multi-model NoSQL database management system that provides a distributed database engine that supports a combination of graph, document, key/value, object-oriented and geospatial models. OrientDB is an ACID-compliant, scalable, secure and high-performance database. OrientDB can embed documents where the relationships are handled as in graph databases. More specifically, OrientDB handles each record or document as an object and the relationships between the objects or documents is managed through fast and persistent pointers between the records. OrientDB provides several indexing mechanisms which provide fast traversal and quick retrieval of related data. It has also an extended SQL query language in order to support the SQL query execution against the database engine without SQL join, manage trees and graphs of connected documents. OrientDB offers features like horizontal scaling, fault tolerance, clustering, sharding and replication. It is designed to reduce operational complexity, provide flexibility and maintain data consistency.

Apache Hadoop is an open-source distributed file-system. It is considered a de-facto standard for storing and managing large amount of data offering also advanced data processing capabilities. In its core lies the Hadoop File system (HDFS), where files are split into large blocks and distributed across several nodes in a cluster. Hadoop support parallel processing via its implementation of the MapReduce programming model for large-scale data processing. Hadoop also offers a resource management framework, YARN, and a plethora of parallel data processing frameworks like Spark and Flink that are also performing the advanced query processing on top of the stored data.

Apache Accumulo [89] is an open-source sorted, distributed key-value database, based on Google's Bigtable. Accumulo utilizes Hadoop's HDFS for distributed storage and Apache Zookeeper for consensus. Accumulo is highly scalable, running on a cluster of HDFS instances. Accumulo offers a programming mechanism, called Iterators, to handle key/value pairs at various points in the data management process. Accumulo introduces cell-level security where each key-value pair obtains its own security label, limiting query results during query execution based on authorization policies. This allows multiple security requirements to be applied within the same table and different authorization policies to be applied per user.

VoltDB is an open-source in-memory, distributed NewSQL RDBMS. VoltDB is ACID-compliant and uses SQL for application interaction. VoltDB supports horizontal partitioning and active-active redundant clustering. VoltDB is utilizing shared nothing architecture which is a distributed-computing architecture where each node is independent and self-sufficient, and there is no single point of contention across the system. VoltDB provides high availability via synchronous replication and full disk persistence.

Apache Hive is a big data query platform designed to facilitate reading, writing, and managing large datasets residing in distributed storage and queried using SQL-like syntax. Hive is built on top of Hadoop providing the tools to easily access the stored data via an SQL-like query language, called HiveQL, for ETL, reporting and data analysis purposes. Hive is able to access files stored in HDFS or other data storage systems that integrate with Hadoop like HBase. Hive offers an SQL-like query interface where traditional SQL queries are translated to MapReduce and Spark jobs in order to execute SQL applications and queries over distributed data enabling the portability of SQL-based applications to Hadoop. Moreover, it provides indexing mechanism and adding metadata storage traditional relational database management systems.

Apache Impala is an open-source massive parallel processing query engine that runs on top of Hadoop. Impala provides a mechanism to execute low-latency SQL queries to data stored in HDFS or Apache HBase in order to perform analytics via SQL. Impala is integrated with Hadoop sharing the same file and data formats, metadata, security and resource management. Impala is utilizing the SQL mechanism provided by Hive.

The following table documents the storage types, summarizes the main functionalities, describes the integration aspects, documents the license type and provides a list of remarks for these data storage solutions.

Name	Storage Type	Functionalities	Integration	License	Remarks
Oracle Database	Relational Database Management System	• Multi-model database • Large variety of features • On-premises, on-cloud and hybrid cloud deployment • Mature product • Extensive support	• Easy to install • Multiple programming languages are supported	Commercial	• A variety of pricing options depending on the deployment type. • Free version has very limited capabilities and performance mainly for testing purposes.
MySQL	Relational Database Management System	• Pluggable storage engine architecture with multiple storage engines support • Provides a wide range of features like cross-platform support, stored procedures, triggers, cursors, updatable views, performance schema, ACID compliance, SSL support, query caching and full-text indexing.	• Easy to install • Multiple programming languages are supported	Open-source under GPL version 2 license Commercial version available also	• Very active community and support
Microsoft SQL Server	Relational Database Management System	• ACID properties • Easy to use interface • Mature product	• Easy to install • Multiple programming languages are supported	Commercial	• Limited performance on big-data operations • Horizontal scaling only supported
PostgreSQL	Relational Database	• ACID properties	• Easy to install	Open source under BSD	• Active community

Name	Storage Type	Functionalities	Integration	License	Remarks
	Management System	• Easy to use interface • Mature product • Extensive support of geospatial data (PostGIS)	• Multiple programming languages are supported • Partitioning is supported but is rather complex	license	• Limited performance on big-data operations
MongoDB	Document-based NoSQL Data Storage	• Big-data enabled • Support for large unrelated complex and nested data • Horizontal scalability and high availability • Support for geospatial data • Mature product	• Multiple programming languages are supported • Easy to install	Open source version under GNU AGPL v3.0 Commercial version available also	• Very active community • No ACID properties (atomic operations only on single document level)
Apache Hadoop	Distributed File System – (HDFS)	• Big-data enabled • Advanced storing and managing capabilities for large amount of data • Advanced data processing capabilities • Support for MapReduce, Spark and Flink • Provides a resource management framework (YARN)	• Multiple programming languages are supported • Easy to install	Open source under Apache license v2.0	• Very active community
HBase	Column-based Distributed Storage;	• Big-data enabled • Powerful and reliable processing of large files • In-memory operation • Supports collections of one or more key / value pairs that match a record • Linear and modular scalability • Automatic and configurable sharding of tables	• Multiple programming languages are supported • Easy to install	Open source under Apache license v2.0	• Active community • No ACID properties • Single point of failure problem
Blazegraph	NoSQL Graph database	• High-performance graph database supporting Blueprints and RDF/SPARQL APIs • Support for APIs	• Multiple programming languages are supported • Easy to	Open source through GPLv2, Commercial version available also	• Partial integration with Hadoop • High availability and scalability

Name	Storage Type	Functionalities	Integration	License	Remarks
		• High Availability and great scalability performance (commercial version only) • Java memory management • Integrated query functionality	install		only in the commercial version
Neo4j	NoSQL Graph database	• High-performance graph database • ACID-compliant • Performance, scalability and high availability • Designed to leverage data and its relationships • Provides highly-scalable real-time queries of relationships • Support for Cypher query language	• Multiple programming languages are supported • Easy to install	Community edition under GPL v3 license Enterprise and Government edition under commercial license	• Community edition lack of clustering support (single node deployment) • Enterprise edition can be obtained if full-application stack is open-source.
OrientDB	NoSQL database supporting a combination of document, graph, key/value, object-oriented models	• ACID-compliant • Highly scalable, secure and high-performance • Supports documents where the relationships are handled as in graph databases • Provides indexing mechanisms for fast traversal and quick retrieval of related data • Supports fault tolerance, clustering, sharding and replication • Provides an extended SQL query language	• Easy to install • Multiple programming languages are supported • Extensive Java and RESTful API	Open source under Apache license v2.0 Enterprise Edition is also available with commercial license	• Very active community • Open source version is not lacking of features
Apache Accumulo	NoSQL key-value store	• Sorted, distributed key/value database • robust, scalable data storage and retrieval • Supports very large rows with very large numbers of columns; • Offers a programming mechanism called Iterators • Multiple security requirements to be	• Not standalone; • Requires Apache Hadoop, Apache Zookeper and Apache Thrift • Easy integration with a variety of	Open source under Apache license v2.0	• Active community

Name	Storage Type	Functionalities	Integration	License	Remarks
		applied within the same table	tools from Apache Foundation		
VoltDB	Distributed In-Memory NewSQL RDBMS	• ACID-compliant • Supports horizontal partitioning and active-active redundant clustering • Designed with shared-nothing architecture • Provides high availability via synchronous replication and full disk persistence;	• Multiple programmin g languages are supported	Open source version with AGPL license, Commercial version available also	• Partial integration with Hadoop
Hive	Big data query platform	• Reading, writing, and managing large datasets residing in distributed storage and queried using SQL-like syntax • Built on top of Hadoop providing access to files stored in HDFS or other data storage systems that integrate with Hadoop like HBase • SQL-like query interface where traditional SQL queries are translated to MapReduce and Spark jobs Provides via SQL-like query language, called HiveQL	• Requires Apache Hadoop • Easy to install	Open source under the Apache License v2.0	• Active community
Impala	Big data query platform	• Massive parallel processing query engine • Integrated with Hadoop; • Execute low-latency SQL queries • Utilizes Apache Hive SQL query mechanism	• Requires Apache Hadoop • Easy to install	Open source under the Apache License v2.0	• Active community

Table 2-6: Data Storage Software and their main aspects

2.2.4 Query Processing

Query processing is the procedure of transforming high-level queries into a correct and efficient low-level expressions in low-level language that will be used to perform the requested action. In general, there are four phases in a typical query processing:

- parsing and translation where the query is parsed, checked for validity and converted to low-level language

- query optimization where an optimal evaluation plan is generated with the minimum cost for execution
- the evaluation where the optimal evaluation plan is converted into the optimal query
- the execution where the optimized query is executed and the results are returned back to the requestor

An effective and efficient query processing engine shall implement all four phases, yet the execution of the query processing engine depends on the different infrastructures and data models available. In the course of the ICARUS project, large amounts of data will be collected from a variety of heterogeneous data sources in different formats and structures, at different volumes and velocities. Besides effectiveness and efficiency, one key aspect that needs to be taken under consideration is the scalability that should be ensured.

In big data ecosystems, the need to support large-scale query processing is mandatory. In the following paragraphs the available query engines suitable for big data ecosystems are described, taking into consideration all the aspects described above:

Apache Spark [90] is one of the most popular tools when it comes for data processing and query processing of large files. Spark is designed using many principles from Hadoop MapReduce engine but focuses primarily on batch processing workloads with full in-memory computation and processing optimization. Spark is providing advanced stream processing capabilities for batch-oriented workloads with the concept of micro-batches, where streams of data are treated as very small series of batches that are handled by the batch engine. Spark offers an ecosystem of libraries that can used to execute advanced, complex or even interactive queries in many programming languages and it is ideal for query processing over large files. Spark offers speed advantages with the cost of high memory usage for batch processing. From stream processing Spark provides increased overall throughput over latency.

Besides a document-based NoSQL storage solution, **MongoDB** also provides a scalable document management engine with document repositories management, as well as an API suitable for advanced database query execution and for targeted query execution, such as geospatial query execution.

ElasticSearch [91] is an open-source, broadly-distributable, readily-scalable, enterprise-grade RESTful search engine. Accessible through an extensive and elaborate API, ElasticSearch can power extremely fast searches over datasets of different format, both structured and unstructured as well as full-text search. ElasticSearch offers integration with Hadoop and several document-oriented databases.

SparqlMap [92] is offering the execution of SPARQL query inside a relational database by supporting mappings between RDF data and the relational schema of a RDBMS. SparqlMap utilizes R2RML to express the mappings used both for extracting RDF from a relational database and for rewriting SPARQL queries into SQL.

FedX [93] is a query engine that enables easy setup of on-demand federations by specifying a list of relevant endpoints (e.g. from the LOD cloud) and querying these federations via SPARQL in a transparent and efficient way.

Virtuoso [94] is a modern enterprise-grade solution for data access, virtualization, integration and multi-model relational database management (SQL Tables and/or RDF Statement Graphs). Virtuoso offers an RDF triple store supporting SPARQL query execution while also offering the publication of SPARQL endpoints.

Apache Solr [95] is an open source enterprise search platform offering features like faceted search, near real time indexing, full-text search and hit highlighting. It also supports spatial search via a built-in mechanism. Solr offers extendable interfaces like XML, JSON and HTTP. Moreover, Solr supports integration with the majority of SQL and NoSQL databases and rich document (PDF files) handling. The query processing engine is optimized for high volume traffic, highly reliable, highly scalable and fault-tolerant.

Sansa-Stack [96] is a processing data flow engine enabling data distribution, communication and fault tolerance for distributed computation on top of RDF large-scale datasets. Sansa-stack is enabling read/write operations of native RDF or OWL data from HDFS or a local drive in the native distributed data structures. Sansa-stack is providing a querying library for query execution on an RDF graph to facilitate browsing, searching and exploring the structured information available in a fast and user-friendly way. Sansa-stack is providing and inference library in order to perform rule-based reasoning on the RDF and OWL data. Finally, Sansa-stack offers a machine-learning library to enable graph structure and semantics exploitation using the RDF and OWL standards. SANSA can be easily integrated with open source systems both for data input and output (HDFS) and is built on top of Spark and Flink.

Apache Flink [97] is an open source streaming processing framework, supporting also batch processing tasks. Flink offers a distributed streaming dataflow engine enabling the data-parallel and pipelined execution of dataflow programs. Flink supports also the execution of iterative algorithms. Flink ensure accurate results, is stateful and fault-tolerant while also ensuring performance at large scale. Flink is capable of high throughput and low latency, offering a state versioning mechanism and is designed to run on large-scale clusters. Flink is performing well in stream processing with real entry-by-entry processing, data partitioning and caching. Flink offers also SQL-style querying, graph processing and machine learning libraries, and in-memory computation. Flink supports also complex event processing via the FlinkCEP library.

The following table summarizes the main functionalities, the integration aspects, the license information and a list of remarks for the query processing tools mentioned above.

Name	Functionalities	Integration	License	Remarks
Apache Spark	• Support for batch processing • Support for stream processing • Full in-memory computation and processing optimization • Supports advanced, complex or interactive queries in many programming languages • Ideal for query processing over large files • Great performance in batch and stream processing • Supports machine learning • Integration with Hadoop	• Works as standalone or on top of Hadoop • Integrates with various data storage solutions (HDFS, MongoDB, Cassandra, HBase)	Open source under Apache License 2.0	• Active community • Due to the in-memory processing architecture high computational resources are required.
MongoDB	• Supports advanced query	• Multiple programming	Open source version under	• Active community

Name	Functionalities	Integration	License	Remarks
	processing • Supports field, range queries, regular expression searches • Batch processing and streaming processing • Load balancing • Support for large unrelated complex and nested data • Support for geospatial data	languages are supported • Easy to install	GNU AGPL v3.0 Commercial version available also	• No ACID properties. • High memory consumption problems • Limitations on queries • Queries against an index are not atomic
ElasticSearch	• Powerful broadly-distributable, readily-scalable, enterprise-grade RESTful search engine • Batch processing and aggregation operations • Integration with Hadoop and document-oriented databases for query execution • Supports query DSL • Load balancing • Horizontal scaling (sharding) • Advanced search capabilities for full text search • Supports query parsers	• Multiple programming languages are supported • Easy to install • Extensive RESTful API	Open source under Apache License 2.0	• No support for distributed transactions
SparqlMap	• Enables SPARQL execution on top of relational databases by a SPARQLtoSQL interpreter	• Standalone command line tool	Unspecified	• Not active community • R2RML knowledge is mandatory
FedX	• SPARQL query engine; • Good performance in SPARQL query execution • Support for federation of SPARQL endpoints	• Standalone Java application	Open Source under GNU Affero General Public License (AGPL)	• Lack of API • Sometimes query results are incomplete due to SPARQL endpoint availability
Virtuoso	• Hybrid database engine offering the functionalities of RDBMS, ORDBMS, RDF, XML, free-text data management • Supports OWL reasoning • Supports publication of SPARQL endpoints • Provides high-performance query processing in both	• Standalone hybrid server	Open source under GNU General Public License (GPL) Version 2	• Active community • Sometimes query results are incomplete due to SPARQL endpoint availability

Name	Functionalities	Integration	License	Remarks
	SQL and SPARQL			
Solr	• Provides an engine optimized for high volume traffic, highly reliable, highly scalable and fault-tolerant. • Multiple features like faceted search, near real time indexing, full-text search and hit highlighting. • It has built-in support for spatial search • Offer interfaces like XML, JSON and HTTP. • Supports integration with the majority of SQL and NoSQL databases. • Supports rich document (PDF files) handling.	• Multiple programming languages are supported • Easy to install • Extensive RESTful API	Open source under Apache License 2.0	• Active community
Sansa-Stack	• Processing data flow engine for distributed computation on top of RDF large-scale datasets • Data distribution, data communication and fault tolerance • Read/write operations of native RDF or OWL data from HDFS or local drive • Query execution on an RDF graph • Rule-based reasoning on RDF and OWL data • Machine-learning library for graph structure and semantics exploitation using the RDF and OWL standards	• Runs on top of Spark or Flink • Easy to install	Open source under Apache License 2.0	• API availability • Not very active community
Apache Flink	• Powerful streaming processing framework • Supports also batch processing • Supports the execution of iterative algorithms • Great performance at large scale • Provides high throughput and low latency • Offers SQL-style querying, graph processing, machine	• Standalone or Hadoop installation provided • Does not provide storage system but provides data source and sink connector for various systems such as Kafka, HDFS, Cassandra and ElasticSearch	Open source under Apache License 2.0	• Active community

Name	Functionalities	Integration	License	Remarks
	learning libraries, and in-memory computation. • Supports complex event processing via the FlinkCEP library.			

Table 2-7: Query Processing Software and their main aspects

2.3 Data Analytics Services

Data Analytics [98] is the process of analyzing data in order to discover useful information, draw conclusions and make decisions. It is widely used by researchers or commercial industries for the sake of designing or verifying scientific models or hypotheses, so as to improve and optimize their operations. It enables businesses to make decisions for competitive reasons such responding faster to emerging market trends and gaining competitive edge over their competitors, with the purpose of increasing their revenue.

The area of Data Analytics involves various and more advanced subareas that are related. Descriptive analytics answers the question of “what happened”, that is analyzing data to give valuable insights into the past. For instance, airlines will learn how many travelers traveled last month; a retailer in the duty-free can learn the average weekly sales volume, etc. However, these findings simply signal that something is wrong or right, without explaining why.

The answer to “why something happened” is given by Diagnostic analytics. This type of analytics aims to find out dependencies and to identify patterns from the data in order to obtain deep insights into a particular problem. For example, a travel agency compares customers’ response to a travel package deal in different regions; a duty-free shop drills the sales down to subcategories, etc.

Instead of looking into the past like the previous two types, Predictive analytics focuses on the future and to “what is likely to happen”. It uses the findings of descriptive and diagnostic analytics to detect tendencies, clusters and exceptions, with the purpose of predicting future trends and behaviors, which makes it a valuable tool for forecasting. Thanks to predictive analytics and its proactive approach, airlines for instance, can identify the most popular potential destinations in a specific season period and thus, make the necessary plans to increase their revenue.

Finally, there is the Prescriptive analytics which aims to literally prescribe “what action to take” to eliminate a future problem or take full advantage of a promising trend. For instance, in the previous example of the airlines looking to increase their revenue, prescriptive analytics could suggest increasing the value of a ticket or add more direct flights to a destination in a specific month.

A broader area of Data Analytics is Machine Learning which involves advanced algorithms applied to mathematical models in order to automatically learn and improve in many tasks such as Classification, Clustering, Dimensionality Reduction and Feature Selection. Finally, the rapid progression of Data Analytics methods has given rise to the need of Data Visualization tools, as a means of visualizing the patterns and correlations of the data in a more human-understandable way.

2.3.1 Machine Learning

Nowadays, there is a high demand for application that offer Data Analytics services. Even though there are plenty of advanced algorithms and models that can accurately tackle many difficult problems, they still need huge amounts of time for training, especially when dealing with big data. This is the reason why some of the top organizations in this area like Google, Amazon and Microsoft, are spending tremendous amount of time and money, aiming to develop software frameworks and platforms that can optimize both accuracy and efficiency.

There are many machine learning libraries developed in different languages. One of the most popular machine learning libraries is **Scikit-learn** [99]. It is open source, offering a Python API and an underlying C/C++ implementation to achieve highly efficient performance. It includes a huge collection of machine learning models for classification, regression, clustering, dimensionality reduction and feature selection. Another two popular open source libraries for machine learning tasks are **Java-ML** [100] and **Weka** [101]. Both are collections of machine learning models with Java API. In addition, **mlpack** is a C++ library that also provides Python bindings. For Natural Language Processing (NLP), there is an open source Python library called **NLTK** (Natural Language Toolkit) [102] that supports many NLP tasks such as part-of-speech tagging, lexical analysis, named-entity recognition, n-gram and tree models.

Similarly, **Intel DAAL** (Intel Data Analytics Acceleration Library) [103] is a machine learning library, offering C++, Java and Python APIs and containing models for supervised and unsupervised learning. It is an open source library, optimized for Intel architecture processors and it integrates with big data platforms such as Hadoop and Spark for distributed usage. Comparably, **Spark MLlib** [104] is a scalable machine learning library, designed for Apache Spark with focus on the distribution of the execution on Hadoop clusters. It is open source and provides APIs for Java, Scala, Python and R.

Additionally, **H2O** [105] is a scalable open source platform for machine learning, able to integrate with distributed frameworks like Apache Hadoop and Spark. It can analyze massive amount of data and provides Java, Scala, R and Python APIs. H2O is also commercially supported. Another open source scalable framework built on top of Apache Hadoop is **Apache Mahout** [106]. Its main focus is on collaborative filtering, classification and clustering, using Java libraries. It supports Java and Scala. Moreover, **Apache PredictionIO** [107] is an open source machine learning framework that supports event collection, deployment of algorithms, evaluation and querying predictive results via REST APIs. It is based on scalable open source services like Hadoop, HBase, Elasticsearch and Spark.

There are many proprietary platforms, offering Data Analytics services that are worth mentioning. **IBM Watson Data Platform** [108] is a complete Data Science platform that offers machine learning models, running on the IBM Cloud. **RapidMiner** [109] is also a Data Science platform that offers similar services. Both integrate with any Hadoop-based products like Apache Spark and offer a limited free edition, in addition to their commercial editions. Furthermore, **IBM SPSS** [110] and **Stata** [111] are software packages, widely used for statistical analysis.

Table 2-8 presents the above software and their main aspects.

Software	Last Release Date	License	Platform	Programming Language	Parallel execution (multi-node)	Remarks
H2O	28 Feb 2018	Apache 2.0	HDFS, AWS, Google Cloud, Azure	Java, Scala, R and Python	yes	• Widely used • Easy to use • Scalable • Integrates with Hadoop-based systems
IBM SPSS	8 Aug 2017	proprietary	macOS, Windows	-	N/A	• Statistical analysis • Not open source
IBM Watson Data Platform		proprietary	Linux, IBM AIX, Windows	-	yes	• Well documented • Not open source
Intel DAAL	16 Nov 2017	Apache 2.0	Linux, macOS, Windows	C++, Python, Java	yes	• Integrates with Hadoop-based systems • Scalable
Java-ML	7 Oct 2012	GNU GPL	Java Platform SE	Java	no	• Well documented • Not scalable
Mahout	17 Apr 2017	Apache 2.0	cross-platform	Java, Scala	yes	• Scalable • Integrates with Hadoop-based systems • Distributed computation
NLTK	24 Sep 2017	Apache 2.0	Linux, macOS, Windows	Python	no	• Advanced capabilities for NLP
PredictionIO	28 Sep 2017	Apache 2.0	HDFS, Spark	-	yes	• Scalable • Web service via REST API
RapidMiner	3 May 2017	AGPL	cross-platform	-	yes	• Easy to use • Not open source
Scikit-learn	22 Oct 2017	BSD 3-Clause	Linux, macOS, Windows	Python	no	• Easy to use • Well documented • Widely used • Not easy to customize low-level configurations
Spark Mllib	28 Feb 2018	Apache 2.0	HDFS	Java, Scala, R and Python	yes	• Distributed computation

Software	Last Release Date	License	Platform	Programming Language	Parallel execution (multi-node)	Remarks
STATA	6 Nov 2017	proprietary	Linux, macOS, Windows	-	N/A	• Statistical analysis • Not open source
Weka	22 Dec 2017	GNU GPL	Java Platform SE	Java	no	• Not scalable

Table 2-8: Typical Machine Learning Software and their main aspects

2.3.2 Deep Learning

Nowadays, there is a lot of research and investment in Deep Learning, a subfield of Machine Learning. Deep Learning architectures refer to deep Neural Networks that have a large number of layers, enabling them to automatically learn complex features at multiple levels of abstraction. Theoretically, these models can outperform typical machine learning models as they can be fitted with tremendous amount of training examples without "overfitting" (i.e. learn the training data too well and fail to generalize if fitted with new data) and can automatically learn and use more complex features. However, they have high time and space complexity, demanding exponential computational time and large amount of storage for training due to their deep architecture.

The data science community has found a way to overcome this problem or at least, reduce the training time to a great extent. The distribution of the execution to multiple machines has made a significant impact to the problem. The MapReduce programming model and the development of so many cloud-based systems such as Apache's Hadoop, Spark and Kafka that are able to process large datasets in parallel on a cluster. Furthermore, NVIDIA has developed the **CUDA Toolkit** [112], a high-performance programming model for general computing that runs on graphical processing units (GPUs). With CUDA, the computation time can be reduced dramatically by taking advantage of the GPUs computing power. This is done by assigning the sequential part of the workflow to the CPU and leaving the compute intensive part to run on the GPU cores in parallel. The toolkit is programmable in C/C++, Fortran, Python and MATLAB and its applications can run across all NVIDIA GPU families available, scaling from a single GPU on desktop workstations to cloud-based platforms having multiple GPUs.

In recent years, there has been a gigantic increase in the number of deep learning software developed. Among the various deep learning libraries, **Keras** [113] is one of the most popular. Keras is a fast-growing deep learning library that supports CNNs and RNNs. It provides a consistent and simple Python API that enables the distribution of training onto clusters of CPUs or GPUs. At the moment, the official Keras release runs on top of **Google's TensorFlow** [114], **Microsoft's CNTK** [115] and **Theano** [116], using them as backends. Bearing in mind that Keras can use different backends implementations by writing simple high-level code which can be easily distributed through multiple GPUs, as well as its large fast-evolving community, it is definitely a framework for deep learning to consider.

TensorFlow and Theano are open source low-level numerical computing libraries. Using Theano, the computation can be deployed on either single CPU or GPU architectures. On the other hand, TensorFlow is using data flow graphs that can run efficiently on multiple CPU or GPUs in a desktop, server or mobile device. The **Microsoft Cognitive Toolkit**, also known as CNTK, is an open source deep-learning toolkit. It contains deep learning models such as CNNs, RNNs and LSTMs (Long Short-Term Memory Networks) and it can be integrated with Azure and parallelize its execution to multiple machines or GPUs as there is support for CUDA. All of them can be used for supervised and unsupervised learning, however, CNTK seems to be more efficient for LSTMs and TensorFlow more efficient for CNNs. Theano's development has stopped after the 1.0 release.

Besides the above, another Keras' backend is **Apache MxNet** [117], but it is still in experimental Beta phase. MxNet is a modern machine learning framework that is lightweight and can scale effectively to multiple machines, including GPUs. It is adopted by Amazon and it is directly compatible to Amazon S3, HDFS, and Microsoft Azure. It offers APIs for Python, R, Scala, C++ and Julia, as well as pretrained models. Even though it makes the deployment of deep learning models easy, it leaves the responsibility to optimize and parallelize the execution to the developer. In order to solve this inefficiency, **Gluon** [118] was developed. Gluon is a flexible

Python API over Apache MxNet which simplifies process of creating and training deep learning models, without affecting the performance. However, MxNet adoption by the community is still far away from the levels of TensorFlow.

Likewise to TensorFlow (and Theano), **Torch** [119] is an open source deep learning framework that has its own script language based on Lua programming language. Its main focus is to speed up training with GPUs through an underlying CUDA implementation. Their main difference is that TensorFlow uses static computation graphs to allow the processing of complex inputs and outputs, while Torch uses dynamic computation graphs that allow it to process variable length inputs and outputs which is a great advantage. The downside of the latter is that the flexibility of variable length processing adds more parameters to the model making it slower. PyTorch [120] is an open source deep learning library for Python that is based on Torch.

Caffe (Convolutional Architecture for Fast Feature Embedding) [121] is an open source deep learning framework written in C++ with Python API, developed by the Berkeley Vision and Learning Center. It uses Convolutional Neural Networks (CNNs) very efficiently, but it is not quite good for Recurrent Neural Networks (RNNs). The fact that it is neither extensible nor lightweight for big networks, with slow development and not commercially supported, has led to the development of **Caffe2** [122]. Caffe2 is a lightweight, modular, scalable deep learning framework, which is built on the original Caffe. The main difference seems to be the claim that Caffe2 is more scalable and light-weight. Caffe2 focuses on large scale deployments, able to easily scale up or down and distribute training using multiple CPUs or GPUs. Furthermore, it excels on mobile deployment as it integrates with iOS and Android. However, the fact that is not efficient for RNNs, which are widely used for text processing, might be a significant flow for using these frameworks to the ICARUS platform.

Additionally, there are other deep learning software focusing on scalability. **BigDL** [123] is a distributed deep learning library for Apache Spark. High performance is attained using multi-threaded programming and Intel MKL (Math Kernel Library) [124], a math library for Intel-based systems. It can also load pretrained models of Keras, Caffe and Torch into Spark programs. However, BidDL does not support GPU-acceleration as other deep learning frameworks.

Deeplearning4j (DL4J) [125] is a portable distributed deep-learning framework that is based on JVM. Even though it is open source, it has commercial support and focuses on industry. DL4J is platform-independent, enabling it to integrate with Hadoop, Spark and Kafka, while it can work with multiple distributed CPUs or GPUs. It provides Java and Scala APIs, as well as Python API using Keras. DL4J is surely a framework worth considering to be used in ICARUS, as it takes advantage of the Java usage in industry, as well as its scalability and extensibility.

PaddlePaddle (PArallel Distributed Deep LEarning) [126] is a scalable deep learning platform that offers Python API. It supports CNNs for image processing, RNNs for sentiment analysis and deep learning on recommendation systems. It works on multiple CPUs or GPUs and can run distributed training jobs on Kubernetes and MPI clusters. It is a solid deep learning framework; however, it is not adopted by a community as large as other popular frameworks. This is a main drawback for its future maintenance and adoption.

Table 2-9 sums up the deep learning software and their main aspects.

Software	Last Release Date	License	Platform	Programming Language	CUDA Support	Parallel execution (multi-node)	Remarks
BigDL	4 Jan 2018	Apache 2.0	Apache Spark	Scala, Python	no	yes	• Scalable using multiple CPU • Not supporting GPU acceleration
Caffe	18 Apr 2017	BSD 2-Clause	Linux, macOS, Windows	C++, Python, MATLAB	yes	no	• Good for image processing (CNN), but not good for other tasks (e.g. RNNs) • Fast using GPU acceleration • Configurable at source code level • Large community • Not well documented • Not easily integrating with other systems • Multi-GPU is only partially supported • Supports very few input and output formats
Caffe2	9 Aug 2017	Apache 2.0	Linux, macOS, Windows, iOS, Android	C++, Python	yes	yes	• Similar to Caffe, but more scalable and lightweight • Good for CNNs, but not for RNNs
Deeplearning4j	8 Dec 2017	Apache 2.0	cross-platform	Java, Scala, Python	yes	yes	• Platform independent • Integrates with Hadoop-based systems • Commercially supported
Keras	14 Feb 2018	MIT	Linux, macOS, Windows	Python	yes	yes	• Widely used • Well documented • Can use many different backends (TensorFlow, Theano, CNTK, MxNet) • Difficult to customize low level configurations
Microsoft Cognitive Toolkit (CNTK)	1 Feb 2018	MIT	Linux, macOS, Windows	C++, Python	yes	yes	• Lightweight and high performance • Good for RNNs, but supports other models as well
MxNet / Gluon	20 Feb 2018	Apache 2.0	Linux, macOS, Windows, AWS, iOS, Android	Python, R, Scala, C++, Julia, Perl	yes	yes	• High performance • Gluon makes it easier to write code • Adopted by Azure
PaddlePaddle	9 Dec 2017	Apache 2.0	Linux, macOS, Android, Raspberry Pi	C++, Python	yes	yes	• Well documented • GPU acceleration

Software	Last Release Date	License	Platform	Programming Language	CUDA Support	Parallel execution (multi-node)	Remarks
							• Integrates with other systems • Not so big community
PyTorch	14 Feb 2018	BSD 3-Clause	Linux, macOS	Python	yes	yes	• Processing variable-length inputs and outputs • Good for RNN • Complex model architectures can be easily built • Low-level configurable • Need to write your own training code • Spotty documentation • Not lightweight
TensorFlow	28 Feb 2018	Apache 2.0	Linux, macOS, Windows	Python, C/C++, Java, Go, R	yes	yes	• Large community, used by Google • Supports GPU acceleration • Good for CNN • Low-level configurable • Not very fast for RNN • Not lightweight
Theano	7 Dec 2017	BSD 3-Clause	cross-platform	Python	yes	no	• Low-level configurable • Not lightweight • Development has stopped
Torch	27 Feb 2017	BSD 3-Clause	Linux, macOS, Windows, iOS, Android	Lua, LuaJIT	yes	yes	• Provides pretrained models • Lots of modular models that are easily combined • Support GPU acceleration • Low-level configurable • Complex model architectures can be easily built • Spotty documentation • Not lightweight

Table 2-9: Deep Learning Software and their main aspects

2.3.3 Data Visualization

Data visualization involves the visual presentation of data in a way that makes it easy to understand and interpret [127]. The advantages of understanding data are tremendous and data visualization has become a very critical part of the world today. Data visualization provides users with insights that helps them to take well-informed decisions over time. There are many ways to create a visual encoding of data and each visualization design is full of trade-offs. There is no visualization design suited for all possible tasks. Therefore, it is crucial to validate the effectiveness of a design so as to visualize information clearly and efficiently.

Modern visualization tools can be divided into three categories: **notebook-based visualization tools**, **code libraries** and **business intelligence frameworks**. Notebook-based visualization tools provide interactive, collaborative and exploratory environments for computing and documenting workflows. Furthermore, they combine code fragments that can be executed, text for the description of the application and figures depicting the data or the results on a single web document. These web documents provide a complete record of a workflow that can be converted into various formats (PDF, HTML, etc.) and shared with others. Furthermore, a single web page can contain a mix of programming languages. In addition to the ability to combine text, execute code right on a web page and create charts, the most important feature of notebooks is interactivity. In particular, they offer the capability of modifying part of the source code on runtime, updating the previous results and charts immediately. Furthermore, the interaction that notebooks provide, is suitable for analyzing and exploring large and dynamic data. However, a major disadvantage of the notebook-based visualization tools is the requirement of programming knowledge in order to interact with data.

Tools	Languages Support	Export Formats	Multi-User Environment	Big Data Frameworks Integration	Github Stars	Last Release	License
Apache Zeppelin [128]	Scala, Python, R SparkSQL, SQL, Hive, Shell Markdown	JSON	Yes	Spark, Flink, Ignite, Hive, ElasticSearch, Google BigQuery, Hadoop, HBASE	3476	Sep, 2017	Apache 2.0
Beaker [129]	Python, Python3, R, JavaScript, SQL, C++, Scala/Spark, Lua/Torch, Java, Julia, Groovy, Node, Ruby, HTML, Clojure	-	No	Spark	1624	Mar, 2018	Apache 2.0
Jupyter [130]	40 programming languages, including Python, R, Julia, and Scala	PDF, LaTeX, HMTL, Markdown, reST	Yes (JupyterHub)	Spark	3885	Mar, 2018	BSD-2

Table 2-10: Notebook-based visualization tools and their main aspects

Code libraries (mostly written in JavaScript) provide integration flexibility but require programming skills so as to be integrated in a system. On the other hand, Business Intelligence (BI) tools are enterprise ready solutions that enclose numerous visualization features in a user-friendly method. All BI tools provide an interface that enables non-technical users to create reports and perform exploratory analysis.

Each visualization tool can contain different types of charts. Therefore, depending on the use-case, a different visualization tool may be needed. The most common types of data visualization fall under the following categories [131][132]:

- 2D area (Geospatial): 2D area types of data visualization are usually geospatial, meaning that they relate to the relative position of things on the earth's surface (e.g. cartogram, choropleth, dot distribution map).
- Temporal: Temporal visualizations are similar to one-dimensional linear visualizations but differ because they have a starting and an ending time and items that may overlap each other (e.g. timeline, time series).
- Multidimensional: Multidimensional data elements are those with two or more dimensions. This category is considered home for many of the most common types of data visualization (e.g. pie chart, histogram, scatter plot, bar chart, line chart).
- Tree/Hierarchical: Hierarchical data sets are orderings of groups in which larger groups encompass sets of smaller groups (e.g. general tree visualization, dendrogram, radial tree, tree map).
- Graph/Network: Network data visualizations show how data sets are related to one another within a network (e.g. node-link diagram, dependency graph/circular hierarchy).

There are numerous of charting libraries available with different rendering technologies. A rendering technology can affect the loading time of a chart and also the interactivity. The rendering technologies can be broken into 2 categories: SVG-based (Scalable Vector Graphics) and Canvas-Based.

Canvas is a HTML element and it is used to draw graphics on a web page. It is a bitmap with an "immediate mode" graphics application programming interface (API) for drawing on it. Canvas is a "fire and forget" model that renders its graphics directly to its bitmap on the fly (with JavaScript) and once the graphic is drawn, it is forgotten by the browser; only the resulting bitmap stays around. Furthermore, Canvas is pixel-base (an image element with a drawing API) and it is resolution dependent. Generally, Canvas-based charts are well suited for real time high volume data presentations [133][134].

SVG is a language for describing 2D graphics and is based in XML. Thus, every element is available within the SVG DOM. SVG is known as a retained mode graphics model persisting in an in-memory model and similar to HTML, SVG builds an object model of elements, attributes, and styles. In SVG, each drawn shape is remembered as an object and if the attributes of an SVG object change, the browser can automatically re-render the shape. SVG-based libraries are resolution independent and are best suited for applications with large rendering areas and interactive sharp-looking charts [133][134].

There are many elements that can be equally important when comparing visualization solutions. The most critical **evaluation criteria** are the following:

- Visualization provided
- Rendering Technologies
- Framework Compatibility
- Customization options
- Input data format
- Supported Browsers
- Type of tool (code library or business intelligence tool)
- Library size
- License
- Library enrichment and maintenance over time (this can be assessed through the libraries' Github repository popularity and activity)

Table 2-11 describes the data visualization software and their main aspects.

Tools / Solutions	Supported Chart Categories					Rendering Tech.	Other features							Library enrichment and maintenance over time	
	2D Area	Temporal	Multi-dimensional	Tree / Hierarchical	Graph / Network	Canvas / SVG	Frameworks Compatib.	Customization options	Input Data Format	Supported Browsers	Type	Library Size	License	Github Stars	Last Release
amCharts [135]	Yes	Yes	Yes	No	No	SVG	Angular, React, Vue	Legend, annotation, zoom, interactive, export chart	JSON, CSV	All browsers, including IE6+	Code library	391KB	Linkware license	423	Dec, 2017
AnyChart [136]	Yes	Yes	Yes	Yes	No	SVG	Angular, Ember, React, Vue	Legend, annotation, interactive, export chart	XML, JSON, CSV	All browsers	Code library	1,26MB	Free for non-commercial uses	161	Dec, 2017
CanvasJS [137]	No	Yes	Yes	No	No	Canvas	-	Legend, annotation, zoom, interactive, export chart	JSON, XML, CSV	All browsers, including IE8+	Code library	277KB	Proprietary	-	Feb, 2018
Chart.js [138]	No	Yes	Yes	No	No	Canvas	Angular, Ember, React, Vue	Legend, annotation, zoom, interactive	JSON	All browsers, including IE6+	Code library	50.9KB	MIT	35444	Mar, 2018
Chartist.js [139]	No	Yes	Yes	No	No	SVG	Angular, Ember, React, Vue	Legend, zoom, interactive	JSON	All browsers, including IE9+, Safari7+, Android 4.3+, iOS Safari 6+	Code library	50,5KB	WTFPL or MIT	10487	Apr, 2017
Cytoscape.js [140]	No	No	No	No	Yes	Canvas	Angular, React	Zoom, interactive, export chart	JSON	All modern browsers (canvas support is required)	Code library	308KB	MIT	3976	Feb, 2018

Tools / Solutions	Supported Chart Categories					Rendering Tech.	Other features							Library enrichment and maintenance over time	
	2D Area	Temporal	Multi-dimensional	Tree / Hierarchical	Graph / Network	Canvas / SVG	Frameworks Compatib.	Customization options	Input Data Format	Supported Browsers	Type	Library Size	License	Github Stars	Last Release
D3.js [141]	Yes	Yes	Yes	Yes	Yes	Both	Angular, Ember, React, Vue	Legend, zoom, interactive, export chart	JSON, XML, CSV	All browsers, IE - 9 and above	Code library	217KB	BSD-3	72922	Jan, 2018
Datawrapper [142]	Yes	No	Yes	No	No	-	-	Legend, interactive, export chart	CSV	-	BI tool	-	MIT	884	Nov, 2017
dc-js [143]	Yes	Yes	Yes	No	No	SVG	Angular, React	Legend, zoom, interactive	JSON	All browsers, IE - 9 and above	Code library	89,9KB	Apache 2.0	6102	Mar, 2017
Dygraphs [144]	No	Yes	Yes	No	No	Canvas	Angular, React, Vue	Legend, zoom, interactive	CSV	Firefox, Chrome, IE - 9 and above	Code library	122KB	MIT	2367	Dec, 2017
Echarts [145]	Yes	Yes	Yes	Yes	Yes	Both	Angular, Ember, React, Vue	Legend, zoom, interactive	JSON	All browsers, including IE6+	Code library	690KB	BSD-3	25578	Feb, 2018
FusionCharts [146]	Yes	Yes	Yes	Yes	No	SVG	Angular, Ember, React, Vue	Legend, annotation, zoom, interactive, export chart	JSON, XML	All browsers, including IE6+	Code library	2,24MB	Free for non-commercial uses	-	Oct, 2017
Google Charts [147]	Yes	Yes	Yes	Yes	Yes	SVG	Angular, Ember, React, Vue	Legend, annotation, interactive	JSON	All browsers, including IE6+	Code library	107KB	Free for all usage.	-	June, 2017
Highcharts [148]	Yes	Yes	Yes	Yes	No	Both	Angular, React, Vue	Legend, annotation, zoom, interactive,	JSON, XML, CSV	All browsers, including IE6+	Code library	661KB	Free for non-commercial	7362	Feb, 2018

Tools / Solutions	Supported Chart Categories					Rendering Tech.	Other features							Library enrichment and maintenance over time	
	2D Area	Temporal	Multi-dimensional	Tree / Hierarchical	Graph / Network	Canvas / SVG	Frameworks Compatib.	Customization options	Input Data Format	Supported Browsers	Type	Library Size	License	Github Stars	Last Release
								export chart					uses		
Kibana [149]	Yes	Yes	Yes	No	Yes	SVG	-	Legend, zoom, interactive	-	All browsers, including IE11+	BI tool	-	Apache 2.0	8931	Feb, 2018
Knowage [150]	Yes	Yes	Yes	Yes	No	-	-	Legend, zoom, interactive, export chart	-	FireFox2+, IE8+, Chrome 11+	BI tool	-	AGPL 3.0	42	Nov, 2017
Leaflet [151]	Yes	No	No	No	No	Both	Angular, Ember, React, Vue	Legend, zoom, interactive, export chart	JSON, GeoJSON	All browsers, Safari 5+, Opera 12+, IE 7-11, Safari iOS 7+, Android 2.2+	Code library	136KB	BSD-2-Clause	20890	Jan, 2018
Mandola Dashboard [152]	Yes	Yes	Yes	No	No	Both	-	Legend, zoom, interactive, export chart	-	-	BI tool	-	Under request	-	Oct, 2017
Metabase [153]	Yes	Yes	Yes	No	No	-	-	Legend, interactive, export chart	-	-	BI tool	-	AGPL 3.0	8780	Feb, 2018
OpenLayers [154]	Yes	No	No	No	No	Both	Angular, Ember, React, Vue	Zoom, interactive, export chart	JSON, GeoJSON, XML	All browsers, IE - 9 and above	Code library	534KB	BSD-2-Clause	3319	Jan, 2018

Tools / Solutions	Supported Chart Categories					Rendering Tech.	Other features							Library enrichment and maintenance over time	
	2D Area	Temporal	Multi-dimensional	Tree / Hierarchical	Graph / Network	Canvas / SVG	Frameworks Compatib.	Customization options	Input Data Format	Supported Browsers	Type	Library Size	License	Github Stars	Last Release
Plotly.js [155]	Yes	Yes	Yes	No	No	Both	Angular, Ember, React, Vue	Legend, annotation, zoom, interactive, export chart	JSON, CSV	All browsers, including IE8+	Code library	2,3MB	MIT	7802	Feb, 2018
RAWGraphs [156]	No	Yes	Yes	Yes	No	SVG	-	export chart	JSON, CSV	All browsers, IE - 9 and above	BI tool	-	Apache 2.0	5053	Jan, 2017
Sigma.js [157]	No	No	No	No	Yes	Canvas	Ember, React	Zoom, interactive, export chart	JSON	All browsers, including IE9+	Code library	368KB	MIT	7803	Oct, 2017
Tableau [158]	Yes	Yes	Yes	Yes	Yes	Canvas	-	Legend, annotation, interactive, export chart	JSON, CSV	All browsers, including IE11+	BI tool	-	Proprietary	-	Feb, 2018
ZingChart [159]	Yes	Yes	Yes	Yes	No	Both	Angular, Ember, React	Legend, annotation, zoom, interactive, export chart	JSON, CSV	All browsers, including IE6+	Code library	687KB	Free or commercial	171	Jan, 2018

Table 2-11: Data Visualization software and their main aspects

2.4 Relevant EU Projects

ICARUS comes as a project that will highly exploit and go beyond the outputs of other relevant projects, hence, it is important to briefly present some of them. Currently, there are some ongoing projects that have received funding from the European Union's Horizon 2020 research and innovation programme and are relevant to ICARUS, as they deal with the domain of big data, data preparation and data analytics, as well as data privacy and data sharing.

BigDataOcean [160] (expected completion June 2019) deals with big data in the maritime domain. The main objective of BigDataOcean is to develop a multi-segment platform that will combine data of different velocity, variety and volume under an inter-linked, trusted, multilingual engine to produce a big data maritime repository for EU-based companies and organizations, acting as a resource of collaborative, data-driven intelligence. It can be seen that this project moves toward a direction which is similar to ICARUS, but in another domain.

AEGIS [161] (expected completion June 2019) aims to create an interlinked "Public Safety and Personal Security" Data Value Chain and deliver a novel platform that can revolutionize semantic technologies in big data, big data analytics and visualizations, as well as security and privacy frameworks. This project will help EU companies to adopt a more data-driven mentality, extending and/or modifying their individual data solutions and offering more advanced data services (e.g. data cleansing, data integration, semantic data linking), while at the same time attaching value to their datasets and introducing novel business models such as a business broker engine based on blockchain to support the data sharing economy, acting as a data marketplace.

TOREADOR [162] (expected completion December 2018) takes a model-based Big Data Analytics as-a-service approach, providing models of the entire Big Data Analysis process including handling big data opacity, diversity, security, and privacy compliance, and will support abstract modelling of the Big Data Analysis life cycle from distributed data acquisition/storage to the design and parallel deployment of analytics and presentation of results. TOREADOR is suitable for standardization models, supporting substantial automation and commoditization and can be adapt to the domain-specific requirements of the customer. Regarding the latter, ICARUS may benefit from this project as TOREADOR may be adapt in the aviation domain. Besides that, TOREADOR provides an architectural framework and a set of components for model-driven set-up and management of Big Data analytics processes which might also be useful for ICARUS.

UNICORN [163] (expected completion December 2019) aims to develop a framework for the deployment of secure big data applications on the cloud. More precisely, UNICORN aims to empower the European digital SME eco-system by delivering a novel and unified framework that simplifies the design, deployment and management of secure and elastic-by-design cloud applications that follow the micro-service architectural paradigm and can be deployed over multi-cloud containerized execution environments. Part of the UNICORN framework may be extended to generate the applications related to the demonstrators of ICARUS.

My Health - My Data (MHMD) [164] project (expected completion October 2019) aims to address the challenge of data subjects' privacy and data security in the biomedical sector by introducing a distributed, peer-to-peer architecture which will determine new mechanisms of trust and of direct, value-based relationships between people, hospitals, research centers and business. MHMD will develop mechanisms for data and identity protection and approaches for classifying sensitive data based on their informational and economic value, in order to foster the development of a true information marketplace, enabling individuals to exercise full control over their personal data and leverage their value.

SPECIAL [165] (expected completion December 2019) will address the contradiction between big data innovation and privacy-aware data protection by proposing a technical solution that makes both of these goals realistic. The project aims to allow citizens and organizations to share more data, while guaranteeing data protection compliance, thus enabling both trust and the creation of valuable new insights from shared data. It will provide technology that supports the acquisition of user consent and the recording of both data and metadata according to legislative and user-specified policies, provide secure privacy-aware workflows (including usage/access control, transparency and compliance verification) and robustness in terms of performance, scalability and security.

proDataMarket [166] is a relevant project to ICARUS that has been completed on August 2017 and it was funded by the EU Horizon 2020 programme. proDataMarket is a digital data marketplace for contextually related open and proprietary data that makes it easier for data providers to publish and distribute their data (share for free or trade) and for data consumers to easily access the data they need for their businesses. ICARUS may consider the concepts of proDataMarket, in order to develop a marketplace for digital data, incorporating a blockchain broker that will safeguard transactions between different parties.

3 Aviation Data Value Chain Requirements Analysis

Requirement analysis is the cornerstone activity of any successful project. It plays a key role for the successful scoping, defining, estimating and managing of a project right from the start. Successful requirements collection is typically unique in every project and circumstances, but it also can lead to many advantages. For instance, it can accommodate better resource management, design, improved quality in the product delivered, and minimize the risk for delays and overruns. Figure 3-1 depicts a high-level and abstract overview of the requirement analysis' process that this deliverable followed in order to derive the initial needs of the ICARUS target audience.

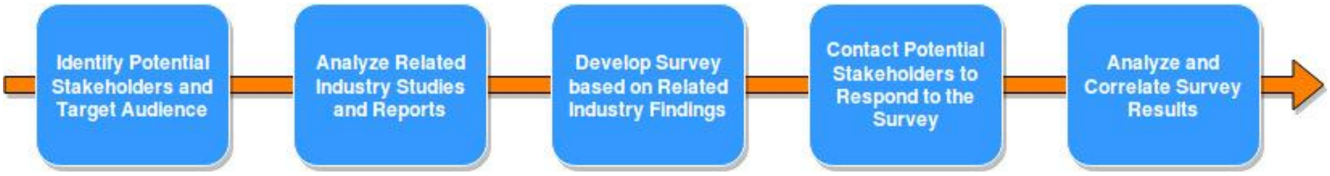

Identify Potential Stakeholders and Target Audience
Analyze Related Industry Studies and Reports
Develop Survey based on Related Industry Findings
Contact Potential Stakeholders to Respond to the Survey
Analyze and Correlate Survey Results

Figure 3-1: High-level process to derive ICARUS Stakeholders Requirements

The first task of this process involved identifying and clearly defining the stakeholders and target audience of the ICARUS platform. A comprehensive description of this task is found in Section 3.1. The next task involved trawling related industry studies and technology leaders' websites for global market and technology reports, relevant to the ICARUS identified stakeholders and target audience. A summary of key findings and points of interest from these studies are listed in Section 3.2. To this end, an online questionnaire was developed and disseminated to various potential stakeholders to ensure the generalization of the requirements beyond the ICARUS consortium. Having obtained multiple completed questionnaires, the final task involved analyzing, correlating and elaborating on the results in order to derive the key findings (Section 3.3).

This procedure helped us to have a more concise picture of the project and to better understand the goals and expectations of the users and stakeholders in a market like the one ICARUS aims to target. The responds to the questionnaire were clear and no further elaboration was needed.

3.1 ICARUS Stakeholders and Target Audience

ICARUS aims to promote data-driven collaboration between the domains that are directly or indirectly linked with the aviation sector, bringing together stakeholders from diverse domains such as Aerospace, Tourism, Health, Security, Transport, Retail, Weather, and Public sectors. Therefore, it is critical to identify which are the potential targeted groups that have specific interests in this project, in order to determine the project's requirements and increase the likelihood of adoption in the aviation sector.

This data-driven collaboration that ICARUS promotes is under the prism of an aviation-driven data value chain, which can be classified into three core tiers (Figure 3-2):

- **Data Tier 1**: **Primary Aviation Data** consists of aircraft sensor data, scheduled route plans, airport traffic, fuel emissions, passenger data that pile up in heaps of data in every flight. Typical data providers include airports, airlines and original equipment manufacturers (OEMs).

- **Data Tier 2**: **Extra-Aviation Data** features data collected by airport services providers and aviation-related service providers (e.g. drones, helicopters, etc.). Such data concern passengers' profiles (purchases in duty free shops, parking history, social media activities, etc.) which are complemented by Linked Open Data (indicatively weather, environment) and other historical data.
- **Data Tier 3**: **Aviation-derived and Aviation-combined Data** contains data and knowledge from businesses and organizations in other sectors such as Health, Tourism, Security industries and Public organizations (e.g. local municipalities), which can be combined with aviation data from tiers 1 and 2 to produce new derived data and create new knowledge that would be impossible to infer otherwise.

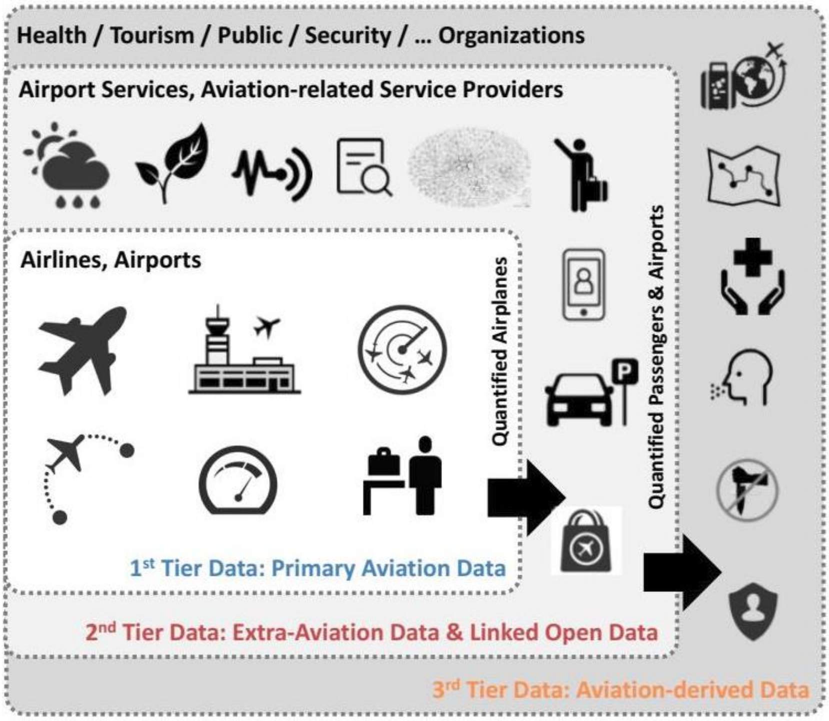

Health / Tourism / Public / Security / … Organizations
Airport Services, Aviation-related Service Providers
Airlines, Airports
Quantified Airplanes
Quantified Passengers & Airports
1st Tier Data: Primary Aviation Data
2nd Tier Data: Extra-Aviation Data & Linked Open Data
3rd Tier Data: Aviation-derived Data

Figure 3-2: ICARUS Data Value Chain

Data providers and consumers from the three tiers define a group of **aviation value chain industry stakeholders** that is directly linked to the project and can benefit with its outcomes. The collaboration and data exchange between airports, airlines and OEMs can help them optimize their operations and inspire innovative products and services addressed to passengers. For example, ICARUS can address the challenge of sharing their data and establishing collaboration principles between them, in a trusted and confidential manner. The combination of data from the 1st and 2nd tiers will enable airport and aviation-related services providers to create more personalized and smart user experiences which may lead to the increase of their sales before, during and after each flight. Businesses and organizations of the 3rd tier will obtain access to data that can be critical for their operations. For instance, in the domain of health, the access to travelling data of passengers will allow better evaluation and intervention for containing the spread of diseases and improving public health policies related to travel restrictions.

Another interested group involves the **IT industry players for the aviation value chain** such as IT companies, web entrepreneurs and software engineers. As ICARUS will consist state-of-the-art technologies for facing different tasks, the exploitation of its open source results can inspire new ideas and applications for this group.

Industry associations and technology clusters that considers European initiatives and clusters (such as SESAR 2020, Clean Sky, BDVA, AIOTI, FIWARE, ETP4HPC, I4MS, etc.) are also considered as a targeted group. This group can include ICARUS results to their collaborative research activities and deduce new knowledge. Furthermore, **participants in European Commission (EC) projects**, as well as the **project partners and relevant stakeholders** active in Horizon2020 are also considered a targeted group. Through ICARUS work, they can identify common topics and by combining its results with their own, they can enhance innovation.

In addition to the above groups, the project can be beneficial for **universities and research organizations** in general. The access to aviation related data will enable future research initiatives. The results of this project can be extent and reused for further advancements and for deployment of innovative technologies and applications.

Moreover, **policy-makers** such as regulatory agencies, ministries and governments, standardization organizations and so on, can be interested in this project. The project considers social, technological, economic, environmental and political aspects that this targeted group can evaluate in order to provide inputs for standardization activities, as well as define future research and innovation directions for the EC.

Finally, it can be beneficial for the passengers and the **general public**, as they will acquire new experiences in their interaction with the aviation industry players of the three tiers.

Table 3-1 summarizes the targeted audience and potential stakeholders of ICARUS and their interests in this project.

Target Audience	Description	Interest in ICARUS
Aviation Value Chain Industry Stakeholders	Data providers and consumers of data from: • 1st Tier: Airports, Airlines, OEMs • 2nd Tier: Airport Services Providers, Aviation-related Service Providers • 3rd Tier: Businesses and organizations in Health, Tourism, Security industries, Public Organizations	• Optimize their operations • Establish collaboration principles • Securely exchange aviation data with respect to their IPR • Strengthened innovation • Training on the project's outcomes
IT Industry Players for the Aviation Value Chain	IT companies, web entrepreneurs, software engineers of solutions for all three tiers of the ICARUS aviation data value chain	• Exploitation of ICARUS open source results • Inspiration for new ideas and applications
Industry Associations and Technology Clusters	European initiatives and clusters (like SESAR 2020, Clean Sky, BDVA, AIOTI, FIWARE, ETP4HPC, I4MS)	• Inclusion of project results to collaborative research activities (roadmap, white papers, etc.) • Dissemination of project results to their members

Target Audience	Description	Interest in ICARUS
		• Infer new knowledge by exchanging ideas
EC Big Data Value Public-Private Partnership Stakeholders	Participants, project partners and relevant stakeholders active in the H2020 projects funded under the EC BDV-PPP programme	• Identify common topics • Synergies and collaborations for results promotion • Enhancing innovation through results combination
Researchers and Academia	Individuals engaged in research initiatives and/or working in research/academic institutes conducting core or application research on big data and / or the aviation data value chain	• Extent project research for further advancements • reuse of the project's innovative technologies to other application domains • Inspiration for future research initiatives based on the project concept and results
Policy-makers	Policy-makers at any level like EC Directorates and Units, Ministries and Governments, Regulatory Agencies, Standardisation Organisations (CEN, ISO, ETSI, etc.) on Big Data technologies	• Evaluate the project's Social-Technological Economic-Environmental-Political (STEEP) aspects • Define future research and innovation directions for the EC • Inputs for standardization activities
General Public	Passengers and the general public who benefit from the project outcomes	• Acquire new experiences by interacting with the aviation industry players of the three tiers

Table 3-1: ICARUS Target Audience

3.2 Key Findings from Industry Studies

In order to properly capture the needs of the aviation sector beyond the consortium and the stakeholders engaged in ICARUS and the best practices currently adopted by the organizations and businesses, various related industry studies were analyzed. The purpose of this process is not to attain a detailed list of specific requirements and technologies related to ICARUS project particularly. The main purpose was to provide inputs to form the questionnaire, in order for the questionnaire to keep up with the specific domain of interest. Not only this, but also to complement and further support the key findings of the ICARUS stakeholders' requirements (identified in Section 3.3), as it seems to move toward the same direction.

The related studies considered, are not only for the aviation domain, but also for the domain of big data in general. The reason for considering big data studies that are not aviation-specific is quite rationale and is driven through the aviation industry studies, as the advancements in the big data domain are highly affecting the aviation domain. Recent studies [167] has shown that about 75% of airlines and aerospace companies consider Big Data as a significant opportunity to improve efficiency and dispatch reliability. More than that, 77% of aerospace companies and 69% of airlines have already a big data initiative underway or are planning to.

The analysis of the related studies has shown that the aviation sector organizations such as the airline industries are interested in new technology related to the field of big data and are eager to adopt the state-of-the-art technologies [168]. The advances in data analysis, processing power and cloud computing have created significant challenges for these organizations, challenges that can be beneficial for adapting to changes in supply and demand in real-time.

One of the most common challenges is the limited access to usable data [169], as eight out of ten respondents consider the process of collecting the necessary data difficult [170]. According to the respondents, shared platforms that provide open access to information are considered vital for the future. In addition, airlines not only struggle to collect data, but also to analyze them, as half of their time is spent on this process [171]. The respondents in [172] stated that one of their main issues is that they are dealing with noisy and complex data.

As a result, the adoption of big data application and cloud computing by commercial aviation has increased significantly, focusing on the implementation of aircraft health monitoring and predictive maintenance systems [173]. The findings have proved that predictive analytics, dynamic pricing and cognitive intelligence will have a great impact for the airlines in the future, as well as personalization [171].

Table 3-2 summarizes the main key findings from related industry studies.

Industry Studies	Key Findings
FlightGlobal - The Big Data Landscape: How the aviation and aerospace sectors view the Big Data opportunity [167] Year: 2018 • 300 industry professionals worldwide • Aerospace companies and airlines	• about 75% of airlines and aerospace companies consider Big Data as a significant opportunity to improve efficiency and dispatch reliability • 67% of aerospace companies have a big data initiative underway and 10% more are planning to, while only 44% of airlines have a big data initiative underway and 25% are planning to • about 75% of the respondents are currently using aircraft health data to make business decisions • airlines are currently investing sufficiently in Big Data: 79% in Asia-Pacific; 61% in North and South America; 55% in Europe, Middle East and Africa • 43% of respondents strongly believe that airlines can realize significant business benefits by sharing their data with OEMs/MROs, while 37% believe that it depends what the OEM/MRO offers in return
FUTURE OF THE AIRLINE INDUSTRY 2035 [168] Year: 2017 • 500 industry professionals • 16 interviews with industry representatives, sector experts and futurists	• The airline industry does not open the way to technological innovation, but it responds to new technology. Research in big data and data transparency, breakthroughs in energy, the creation of new manufacturing tools, planning for alternative transport modes and the evolution of quantum computing may disrupt the existing airline models. • Notable challenges for airlines are caused by the progress toward big data, predictive analytics, processing power, sensor technology, storage and connectivity. • Advances in big data and data analytics come to the aid of airlines in order to predict and adapt to changes in supply and demand in real-time. • Shared platforms that give open access to information is the future.
MRO BIG DATA – A LION OR A LAMB? INNOVATION AND ADOPTION IN AVIATION MRO	• More than 98 million terabytes of data could be generated by the newest aircrafts by 2026. • There is a significant increase in the adoption of big data applications in

Industry Studies	Key Findings
[173] Year: 2016 • Respondents: top executives from airline operations, procurement and engineering departments, captive and independent maintenance providers, OEM aftermarket divisions, and financing and leasing professionals • Global cross section of the industry	commercial aviation, with the majority announcing implementation of aircraft health monitoring (AHM) and predictive maintenance (PM) systems on at least a fair scale. • Instead of following a broad or comprehensive approach, 59% of airline respondents stated that they are planning to restrict AHM use to narrow subsets of data, either directly or through a third party. • 83% of those who are using PM work on small subsets, while only the 20% use all the available data for predictive techniques. • The majority of operators follow two trends: either entering big data wisely by building their analytics upon the most influential and feasible datasets that they can select, or not allocating the sufficient resources to handle the full volume of available data. • 27% of airline respondents gather and analyze data for PM using an in-house customized platform. • Respondents seem less willing to hire external support for data analysis and visualization of results than for storage and aggregation of data.
The Future of Airline Distribution, 2016 – 2021 [171] Year: 2016 • 49 airline participants • 21 telephone interviews with airline executives and managers • Interviewed 17 technology vendors and industry consultants. • 42% Director/Managing Director roles, 34% Vice Presidents/SVPs, 22% Supervisors/Managers, and 2% "C-level" roles	• Significant impact for the airlines by 2021: Dynamic Pricing (69%) Predictive Analytics (62%) and Cognitive Intelligence (33%) • Even though personalization is not considered critical at the moment, 88% of airline executives expect that it will be very important by 2021. • 50% of airlines' time is spent on data that they struggle to collect and analyze. • The average score of customer data quality which is given by airlines is 5.2 out of 10.
DZone: "Artificial Intelligence: Machine Learning and Predictive Analytics" [169] Year: 2017 • 463 respondents • 36% developers/engineers, 17% developer team leads, and 13% software architects • 33% Europe; 36% North America.	• The most common programming languages for developing applications for Artificial Intelligence (AI) / Machine Learning (ML) are Java (41%), Python (40%) and R (16%). • The most common frameworks for AI / ML are TensorFlow (25%), Spark MLlib (16%) and Amazon ML (10%). • The most common challenges that organizations face are the lack of data scientists (43%), achieving real-time performance in production (40%), developer training (36%), and limited access to usable data (32%) • Organizations that are not currently investing in AI stated that this is due to the lack of apparent benefit (60%), developer experience (38%), cost (35%) and time (28%).
DZone: "Databases: Speed, Scale, and Security" [174] Year: 2017 • 528 respondents	• The most popular databases used in production environments are MySQL (48%), Oracle (45%), Microsoft SQL Server (32%), PostgreSQL (32%), and MongoDB (28%). • The most popular databases used in non-production environments are

Industry Studies	Key Findings
- 38% developers/engineers, 17% developer team leads, and 15% software architects - 38% Europe; 33% North America	MySQL (51%), PostgreSQL (35%), Oracle (34%), MongoDB (31%) and Microsoft SQL Server (26%). - 40% of respondents use NoSQL databases in production environments (adoption grew from 30% in 2016 to 40% in 2017). - The most popular NoSQL databases for production environments are MongoDB (from 19% in 2016 to 28% in 2017) and Apache Cassandra (from 5% in 2016 to 12% in 2017). - 32% of respondents plan to adopt new database technology in the next six months and the most commonly preferred databases are MongoDB (38%), Cassandra (32%), Neo4j (22%) and Couchbase (17%). - There is a significant increase of using cloud storage by developers, from 17% in 2016 to 24% in 2017. - Regarding the partition of a database, 49% of the respondents use Vertical partitioning (split tables), 41% Horizontal partitioning (shared across multiple machines) and 37% Functional partitioning (by bounded context). - Regarding the security protocols of the databases, 89% of the respondents uses Authentication, 58% Encryption, 24% Userspace-level resource isolation, 24% Penetration Testing, 19% Data persisted on separate physical machines and 17% Data persisted on separate VMs
DZone: “Big Data: Data Science and Advanced Analytics" [170] Year: 2017 - 734 respondents - 33% developer/engineer; 22% developer team lead - 35% Europe, 35% North America	- The adoption of Apache Spark has grown to 45% in 2017, compared to 31% in 2016 - 65% of respondents use Apache Hadoop - 47% of respondents use Yarn for cluster resource management - 62% of respondents use Apache Zookeeper for node coordination - 55% of respondents use Hive for data warehousing - Concerning the process of collecting the necessary data, 69% of the responders stated that it is “somewhat difficult” to obtain sufficient training data from their source systems and 13% that is “very difficult”. Only the 18% of the responders consider the process easy - The most popular technologies for data visualization are D3.js (42%), Tableau (29%) and Chart.js (28%) - Tableau was popular specifically for real-time visualization (39%) and “Big Data” visualization (39%). Chart.js was recommended for “Big Data” (32%) and open-source visualization (37%). For the open source category, D3.js was the most popular (53%)
DZone: “Big Data: Stream Processing, Statistics, and Scalability" [172] Year: 2018 - 540 respondents - 42% developers/engineers, and 23% developer team leads - 39% Europe; 30% North America	- Working with noisy data (64%), working with deadlines (40%) and limited training and skills (39%) are the most difficult challenges in data science. - The focus of investing in analytics projects is mostly on the speed of decision-making (47%), the speed of data access (45%) and data integrity (31%). - Comparing to 2016, R usage for data science projects is decreased from 60% to 50% in 2017, while Python usage is increased from 64% to 70%. - When dealing with data of high volume, the most challenging data sources are files (47%) and server logs (46%), while the most challenging data types are relational (51%) and semi-structured (e.g.

Industry Studies	Key Findings
	JSON, XML; 39%). • The most challenging data sources when dealing with high velocity data are sensors/remote hardware and server logs (both 42%), while the most challenging data types are semi-structured (36%) and complex data (e.g. graph, hierarchical; 30%) • About data variety, 56% of the responders stated that the most challenging data sources are files, while 28-32% of the responders stated the same for sensor/remote hardware data, server logs, ERP and other enterprise systems, user-generated data and supply-chain/logistics/other procurement data.

Table 3-2: Related Industry Studies Key Findings

3.3 ICARUS Survey Key Findings

3.3.1 Respondents Profile

The online survey contains 44 questions in English and there were 31 experts from the aviation sector that participated from February 1st until March 12th, 2018. The survey will remain open so as to receive more responses and the updated results will be included in future deliverables of WP1. The participants who took the survey were promised complete anonymity.

The majority of participants (62%) hold “Management/Team leader” roles, 10% “Sales Marketing” roles, 3% “Data/Business Analyst” roles, 3% “Researcher” roles, 3% “Scientist” roles and 3% “Software Developer” roles. Furthermore, the majority of participants (76%) have “more than 15 years of experience” in their position, 14% have “4-10 years of experience”, 7% have “11-15 years of experience”, and 3% have “1-3” of experience. Most participants (66%) work in an organization that has more than 250 employees and most of the organizations (35%) are identified as “Transport Organizations” (e.g. Airports, Airlines) (Figure 3-3). These organizations operate in multiple and different fields in the aviation sector (Figure 3-4) such as “Airline Services” (61%), “Extra Aviation Services” (35%) and “Research/Learning Services” (32%). Furthermore, most participants (70%) state that their organizations already employ a data analyst. Particularly, all transport organizations and university/research organizations already employ a data analyst while most of the SMEs (66%) do not employ a data analyst.

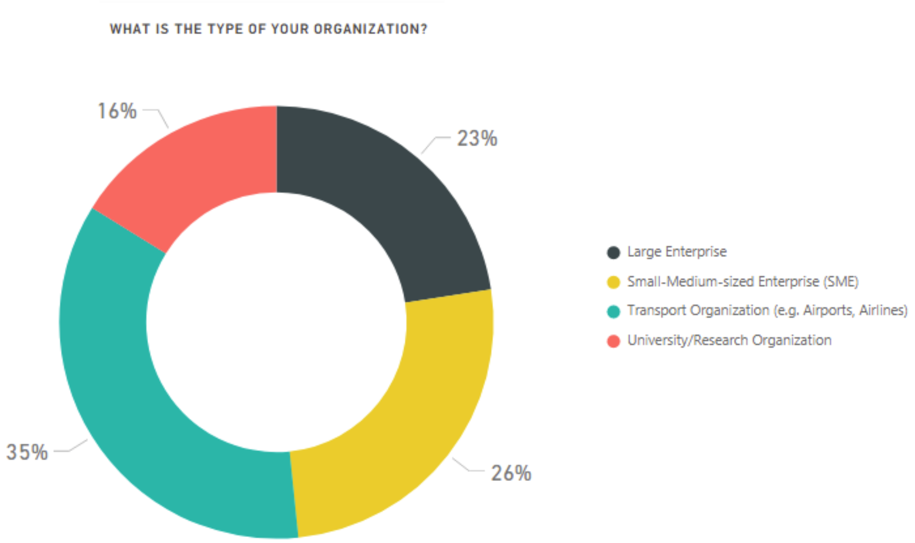


Figure 3-3: Type of Organization

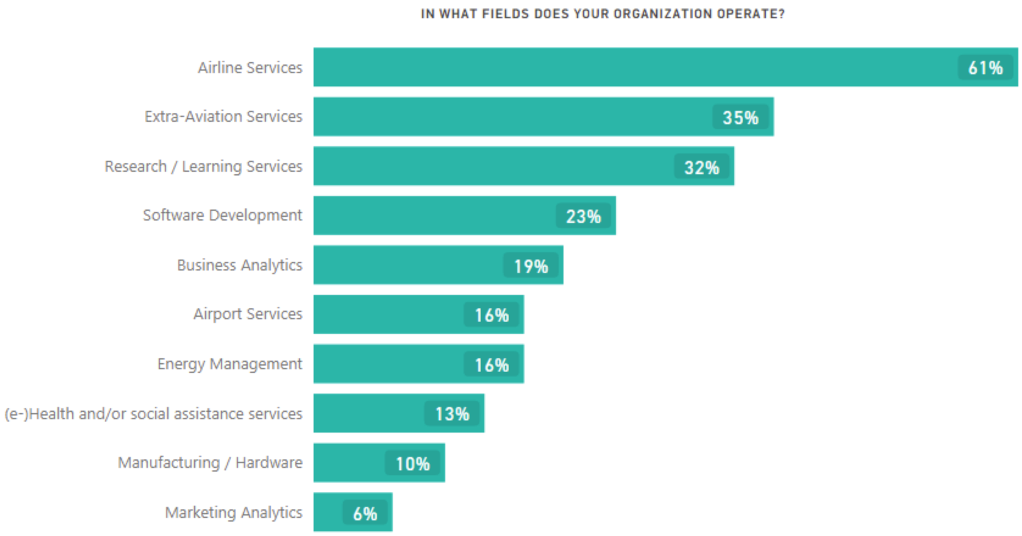


Figure 3-4: Organization Operating Business Domains

Figure 3-5 depicts the data challenges for the organizations. From this figure, we observe that **the most difficult processes for organizations are the data anonymization (33%) and data linking (38%) while the least difficult processes are the collection of data (9%), data analysis (19%), data curation (19%) and data visualization (19%)**. In particular, we observe the following regarding the data challenges:

- 100% of the transport organizations state that it is not easy to collect data and 50% of the transport organizations find it difficult to visualize data. This is also supported by the industry surveys [171].
- 40% of the university/research organizations find it difficult to curate data while 100% of the transport organizations state that it is not easy to curate data.
- 75% of the university/research organizations, 33% of the transport organizations and 50% of the large enterprises find it difficult to link data.

- 50% of the university/research organizations and 25% of the transport organizations state that it is difficult to trade/share their data.
- 40% of the university/research organizations, 30% of the SMEs and 50% of the transport organizations find it difficult to anonymize their data.

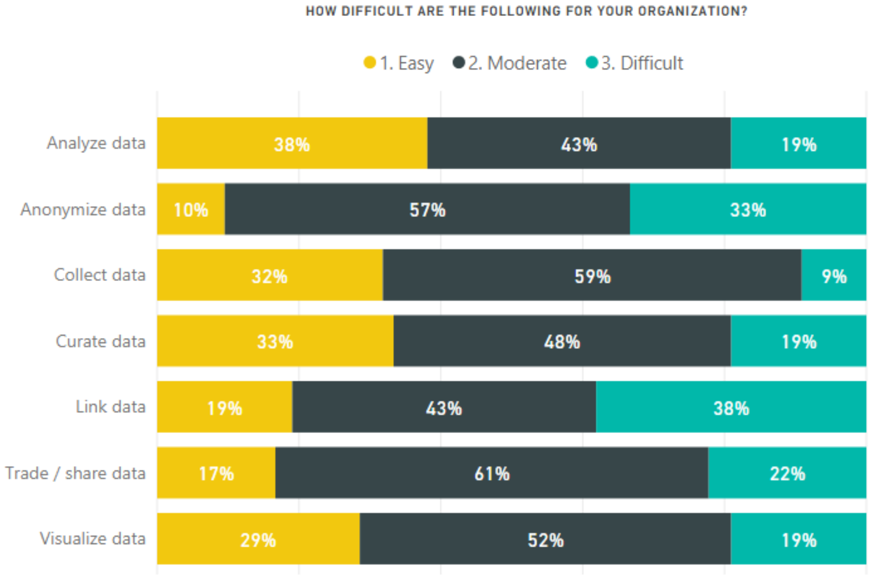


Figure 3-5: Data Challenges for the Organizations

3.3.2 Data Collection

Figure 3-6 depicts the aviation related domains of data that are been collected by the organizations. From this figure, we observe that the “Aircraft Data” (81%), “Airport Data” (77%), “Aviation Business Data” (77%) and “Environmental Data” (77%) are the most popular aviation related domains of data that are been collected. After a further analysis, we observe the following regarding the collection of data:

- 100% of transport organizations and 90% of large enterprises, that are related to the aviation sector, are collecting “Aircraft Data” and “Airport Data”.
- 90% of transport organizations and 85% of large enterprises are collecting “Environmental Data”.
- 70% of transport organizations are not collecting “Health and Epidemics Data”.
- The difficulty for linking and anonymizing data does not change depending on the data domain.
- The difficulty for visualizing data is increasing for the “Aircraft Data” and “Aviation Business Data”.

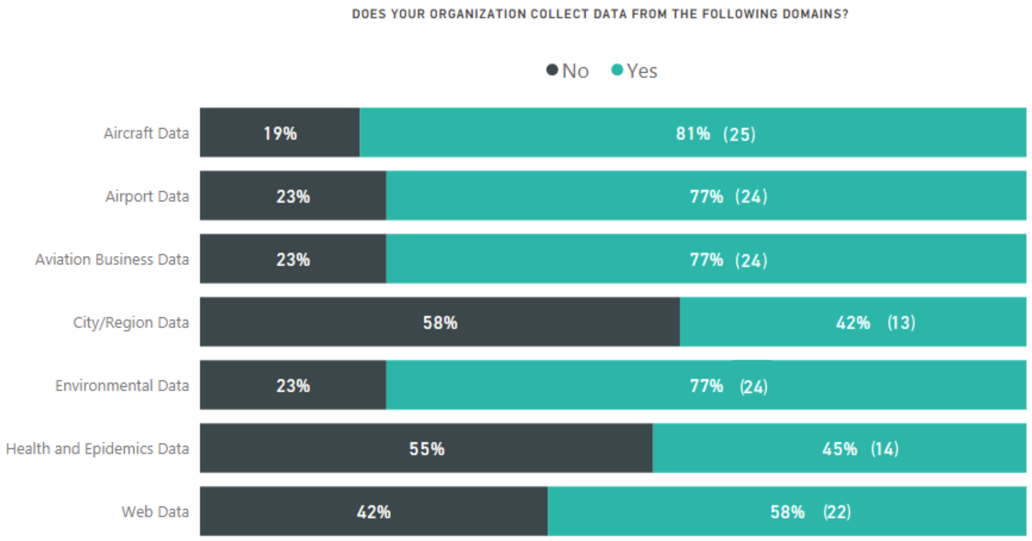


Figure 3-6: Domains of Data Collected by the Organizations

The majority of the participants that collect aircraft, airport, aviation business and environmental data mention that they collect them continuously in real time. On the other hand, city/region and health/epidemics data are collected mostly under request (Figure 3-7). Most of the data that are collected in real-time are either a few megabytes or gigabytes (Figure 3-8). Particularly, **the aircraft, airport, aviation business and environmental data, that are mostly collected in real time, are megabytes or even gigabytes while most of the city/region and health/epidemics data, that are collected under request, are megabytes.** Furthermore, we observe the following regarding the data velocity and the data volume:

- For most participants, a higher velocity of data affects the difficulty of curating and visualizing data, while a higher volume of data affects the difficulty of linking, analyzing, trading/sharing and anonymizing data. Therefore, **aviation related data that have high velocity and high volume, affect the difficulty of curating, visualizing, linking, analyzing, trading/sharing and anonymizing data.**
- University/Research organizations and SMEs are collecting aircraft data mostly “Under Request” while transport organizations and large enterprises are collecting aircraft data mostly “Continuously in Real Time”.
- University/Research organizations and SMEs are collecting airport data mostly “Under Request” while transport organizations and large enterprises are collecting airport data mostly “Continuously in Real Time”.
- University/Research organizations and SMEs are collecting aviation business data mostly “Under Request” while transport organizations and large enterprises are collecting aviation business data mostly “Continuously in Real Time”.
- All types of organizations are collecting city/region data mostly “Under Request”.
- University/Research organizations are collecting environmental data mostly “Under Request” while SMEs, transport organizations and large enterprises are collecting environmental data mostly “Continuously in Real Time”.
- Transport organizations are collecting health and epidemics data mostly “Monthly” and “Under Request”.

- University/Research organizations and SMEs are collecting web data mostly “Daily” while transport organizations and large enterprises are collecting web data mostly “Continuously in Real Time”.

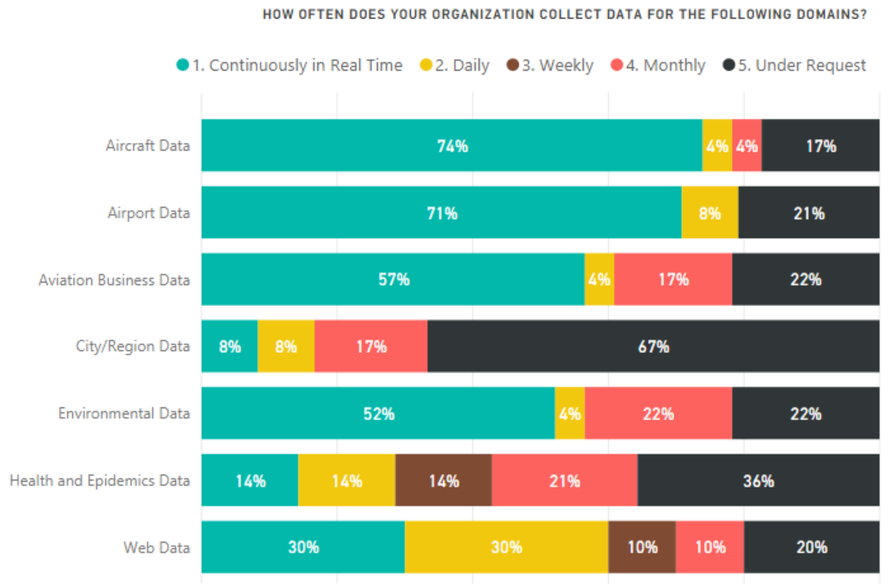


Figure 3-7: Data Velocity

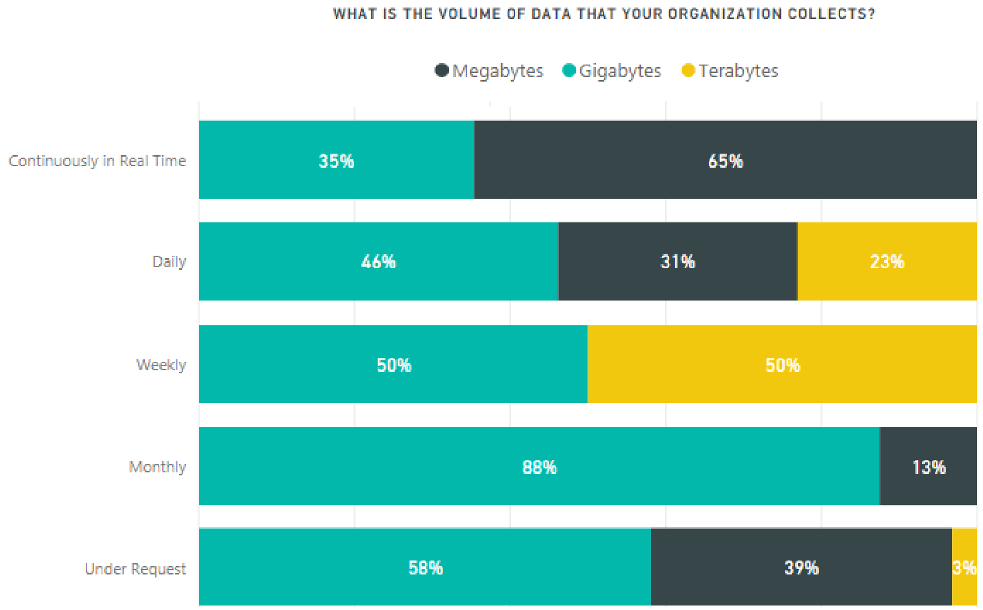


Figure 3-8: Data Volume

The aircraft and airport data are mostly collected from in-house sources (55% and 47% respectively) while environmental and health/epidemics data are mostly collected from other providers. The aviation business data, city/region data and web data are mostly collected either by in-house sources or other providers (Figure 3-9). In particular, **transport organizations are collecting aircraft, airport, aviation business, environmental and web data mostly from in-house sources while SMEs,**

large enterprises and university/research organizations are collecting most of their data (except web data) from other providers.

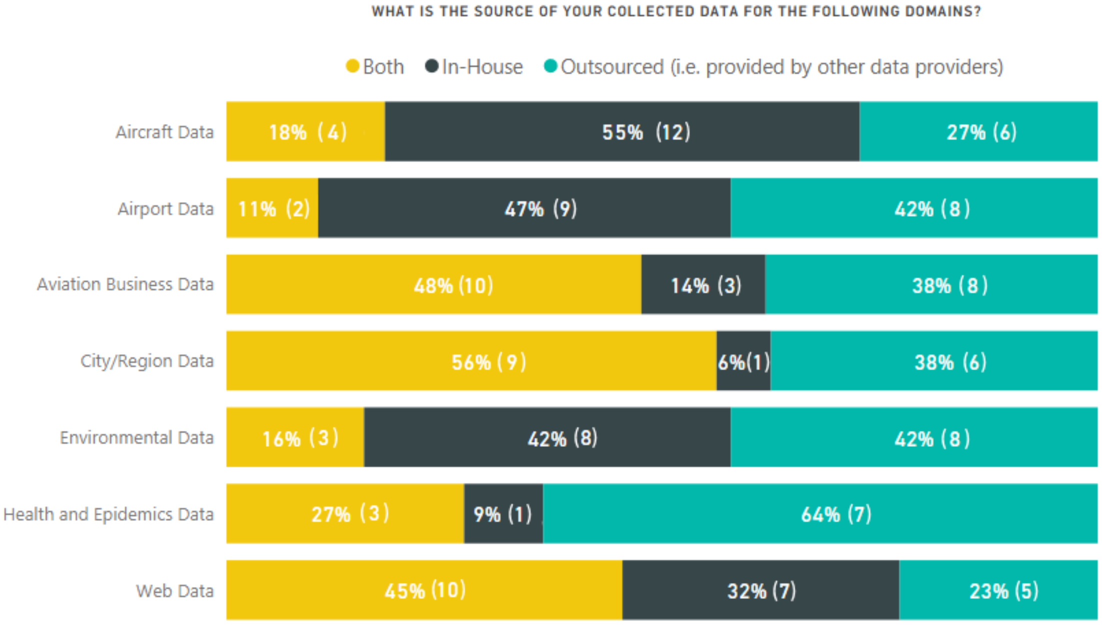


Figure 3-9: Data Sources

Aircraft data are collected mostly using either a “custom in-house mechanism” (31%) or “APIs” (34%) while airport, aviation business, environmental and web data are mostly collected using “APIs”. Furthermore, city/region data are collected mostly “via email” (33%) while health/epidemics data are collected mostly “via intermediate third parties” (30%) (Figure 3-10). After a further analysis, we observe that **organizations that are using “data marketplaces” and “APIs” find it easier to collect data while organizations that are using “custom in-house mechanisms” find it harder to collect data.**

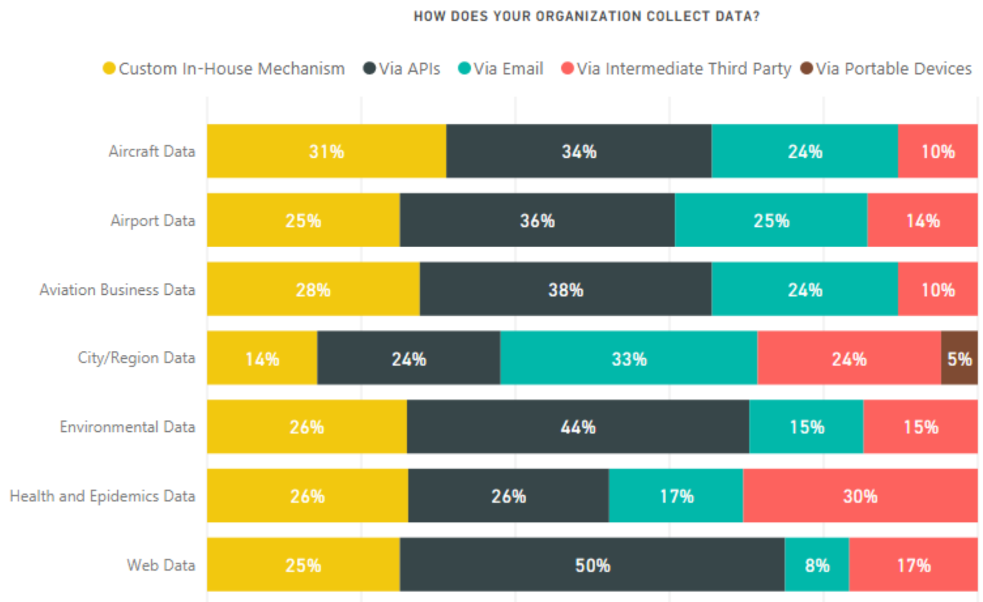


Figure 3-10: Mechanisms for Collecting Data

Most of the organizations are collecting data primarily for "Performance Measurement/Management Benefits" and "Economic Benefits" (Figure 3-11). In particular, the most important reason for SMEs and transport organizations to collect data is the "Economic Benefits" while large enterprises are more interested in "Performance Measurement/Management Information".

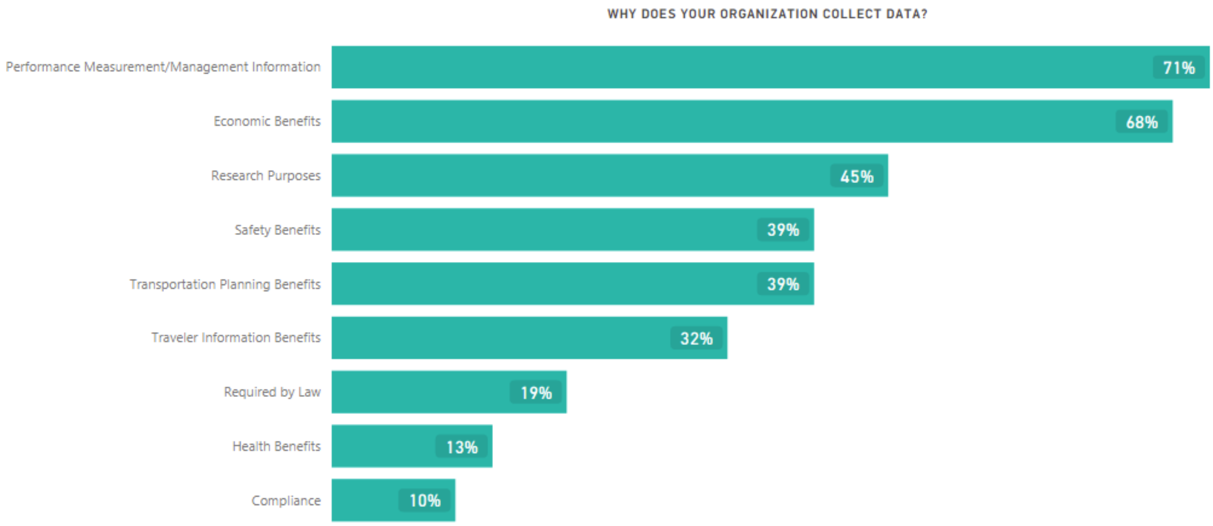


Figure 3-11: Reasons for Collecting Data

Figure 3-12 depicts the availability of a data linking mechanism for external and internal data. From this figure, we observe that the majority of participants are already using a data linking mechanism for external (33%) and internal data (52%). Interestingly, **organizations that have already in place data linking mechanisms still find the data linking process very difficult.** Furthermore, we observe the following regarding the availability of data linking mechanisms:

- Most of university/research organizations have already in place mechanisms for linking external and internal data.
- **Most of SMEs state that it would be valuable a mechanism for linking external and internal data.**
- **87% of transport organizations already have a mechanism for linking internal data but 50% of the transport organizations do not have a mechanism for linking external data.**

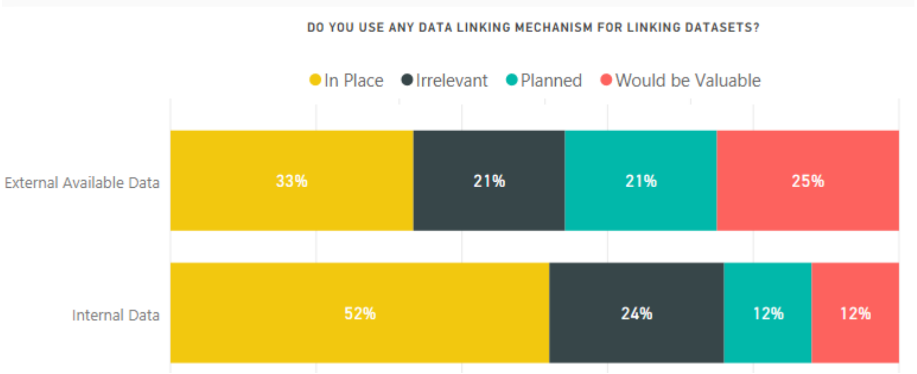


Figure 3-12: Availability of Data Linking Mechanism

The most common issue that makes it difficult for organizations to collect data is the "budget/cost constraints" (89%) (Figure 3-13). Particularly, university/research organizations are very much affected by "regulations constraints", while large enterprises and transport organizations are heavily affected by "trust/security issues" and "technical issues".

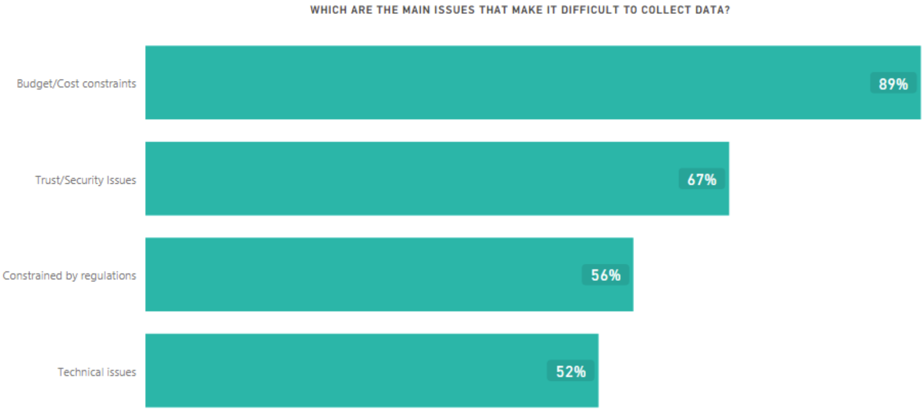


Figure 3-13: Main Issues of Collecting Data

The majority (69%) of the organizations already provide their data to other organizations. In particular, 100% of the transport organizations and 83% of the large enterprises are providing their data to others, while 55% of the SMEs and 80% of the university/research organizations are not providing their data. Furthermore, aircraft, airport and aviation business data are provided very often by organizations related to the aviation sector. On the other hand, city/region and health/epidemics data are much rarer (Figure 3-14).

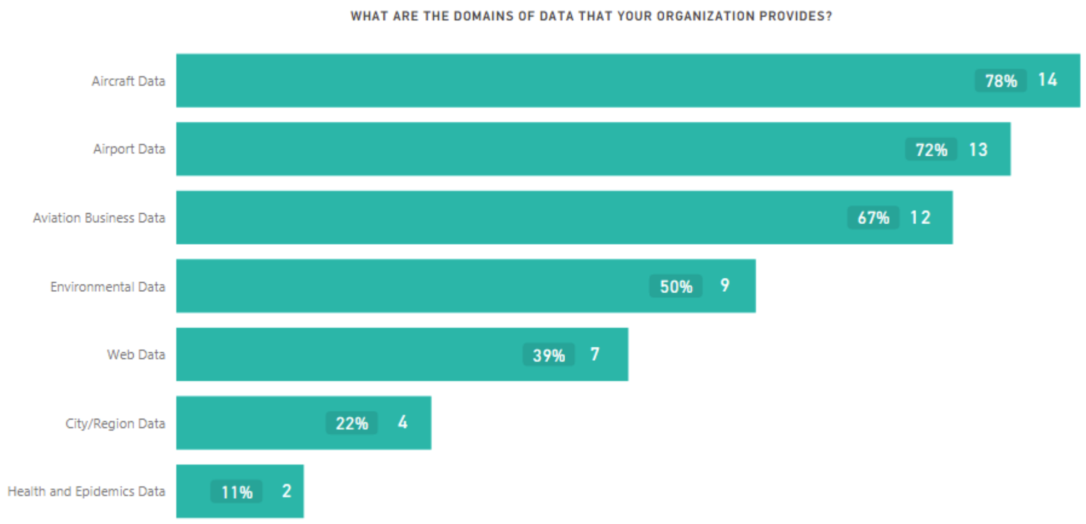


Figure 3-14: Domains of Data Provided by the Organizations

On the other hand, **31% of the organizations do not share their data due to the sensitivity of data (89%) and company policies (67%)** (Figure 3-15). However, they stated that if they could overcome those challenges, most of them (56%) would be interested to share/trade their data.

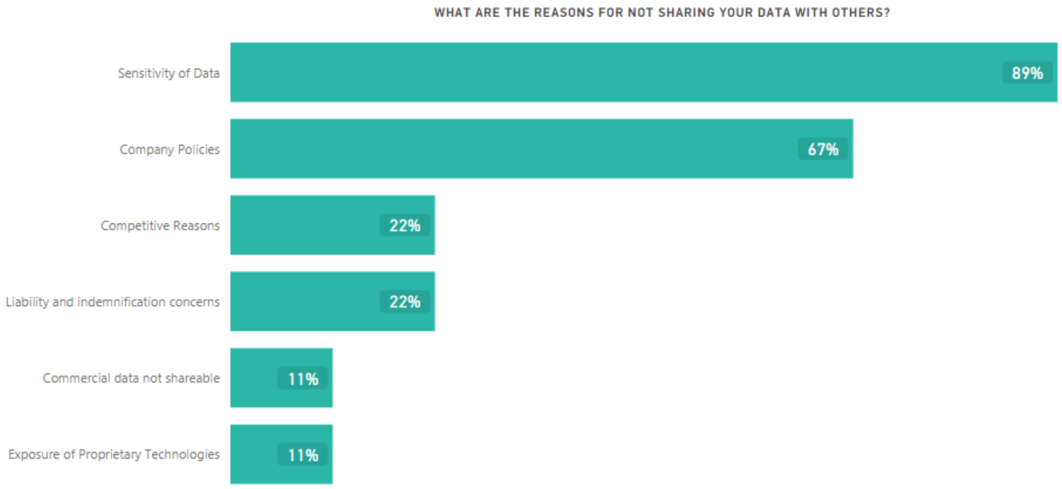


Figure 3-15: Reasons for not Sharing Data

Most organizations provide their data “upon bilateral agreements” (negotiated separately per case) (Figure 3-16). Specifically, **all the large enterprises and transport organizations are providing their data “upon bilateral agreements”. Furthermore, 41% of SMEs are providing their data “upon bilateral agreements” and another 29% are providing them as “Open Data”.**

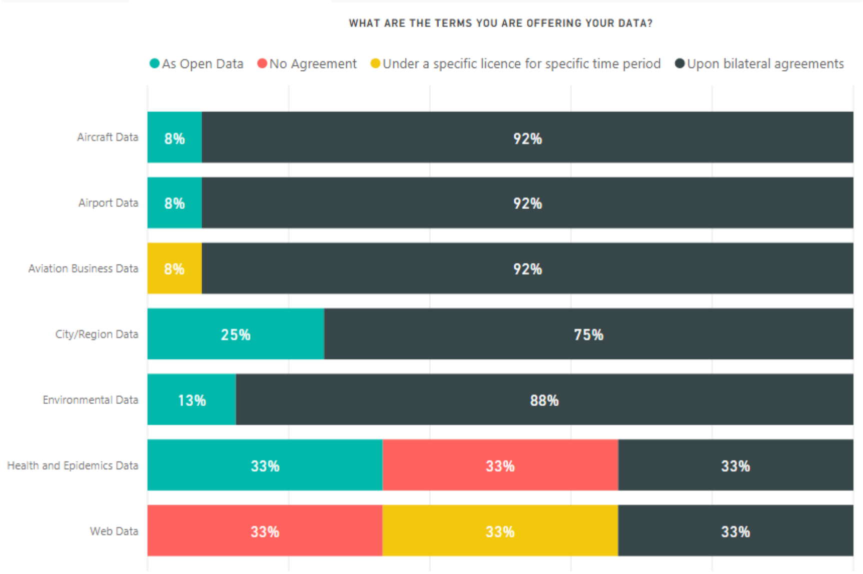


Figure 3-16: License of Provided Data

Figure 3-17 depicts the format of data provided in each domain. From this figure, we observe that **most data are in text format, while health/epidemics and web data are also containing images.**

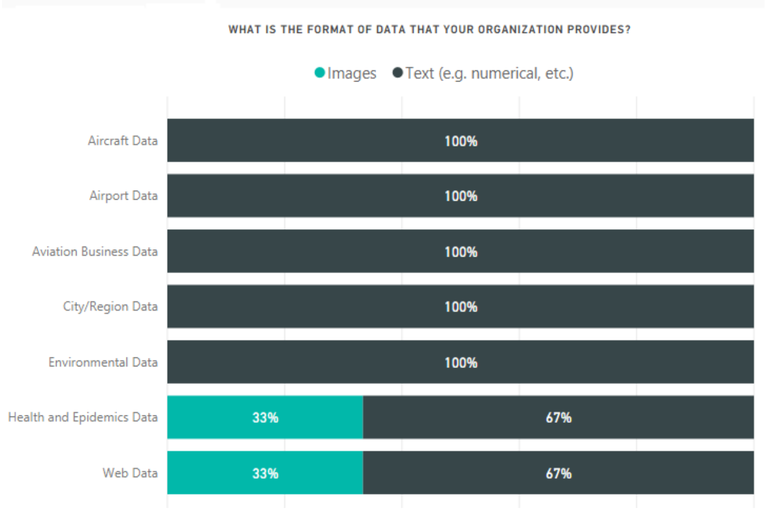


Figure 3-17: Format of Provided Data

To provide their data, most of organizations use APIs (36%) (Figure 3-18). In particular, university/research organizations and transport organizations provide their data mostly “via APIs” while large enterprises and SMEs provide their data mostly “via intermediate third parties (e.g. Data Exchange Platforms)”.

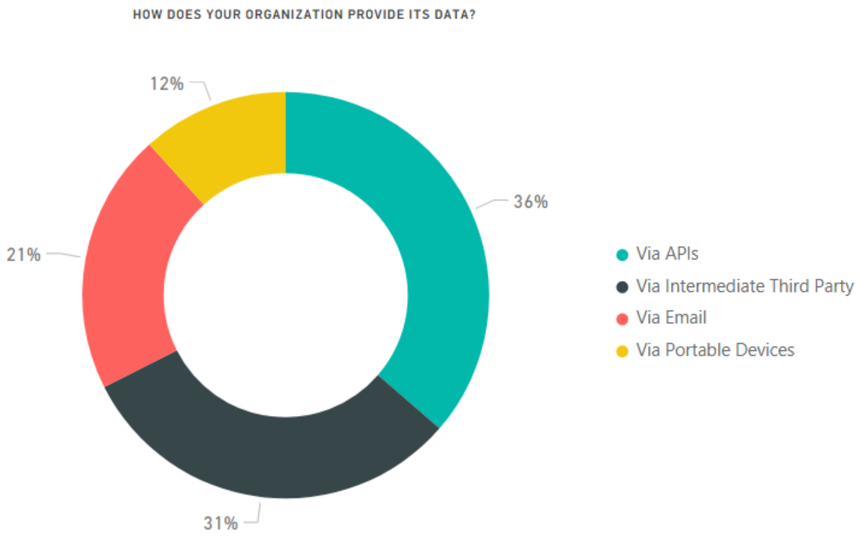


Figure 3-18: Mechanisms for Providing Data

To secure their data, 89% of the participants uses “Authentication”, 67% “Encryption”, 33% “Userspace-Level Resource Isolation” and 19% “Data Persisted on Separate Physical or Virtual Machines” (Figure 3-19). These numbers confirm the DZone [174] reports. After a further analysis, we observe that transport organizations are using mostly "data encryption" to secure their data while large enterprises and SMEs are using many different kind of protection mechanisms (e.g. "authentication", "data persisted on separate physical or virtual machines", "userspace-level resource isolation" and "encryption") to protect their data.

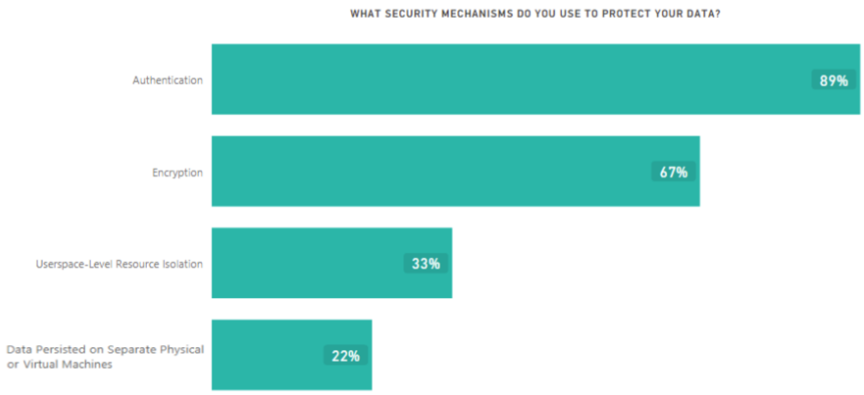


Figure 3-19: Security Mechanisms for Data

3.3.3 Data Analytics

The survey highlights that 50% of the participants state that they already have big data architectures/platforms for data analysis (Figure 3-20). In particular, all the university/research organizations have in place architectures and platforms for data analysis while many transport organizations, SMEs and large enterprises do not have in place architectures and platforms for data analysis.

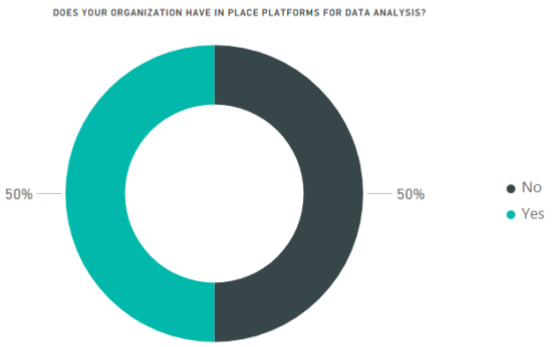


Figure 3-20: Usage of Big Data Analysis Platform

On the other hand, **50% of the participants state that they do not have architectures/platforms for big data analysis because of budget/cost constraints (100%) and lack of experience (86%)** (Figure 3-21).

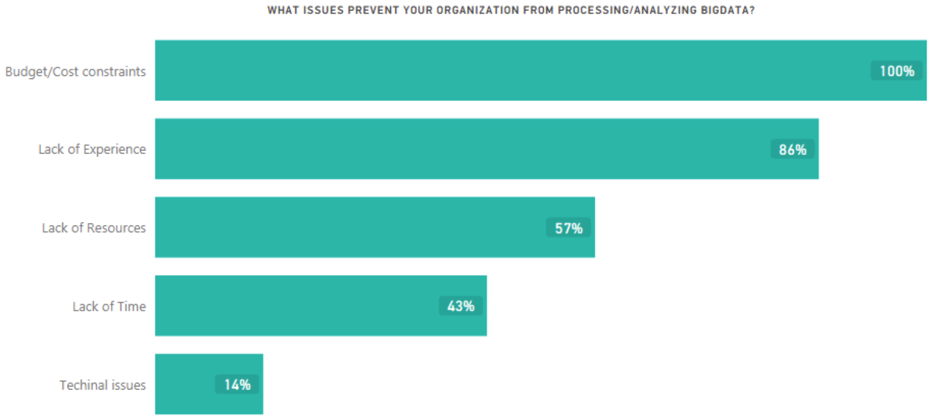


Figure 3-21: Issues Preventing Processing/Analyzing Big Data

The majority (60%) of the participants, state that they use data analytics “to improve operations” (Figure 3-22). In particular, most SMEs and large enterprises use data analytics “to improve operations” and “to improve performance” while all university/research organizations use data analytics for “research”.

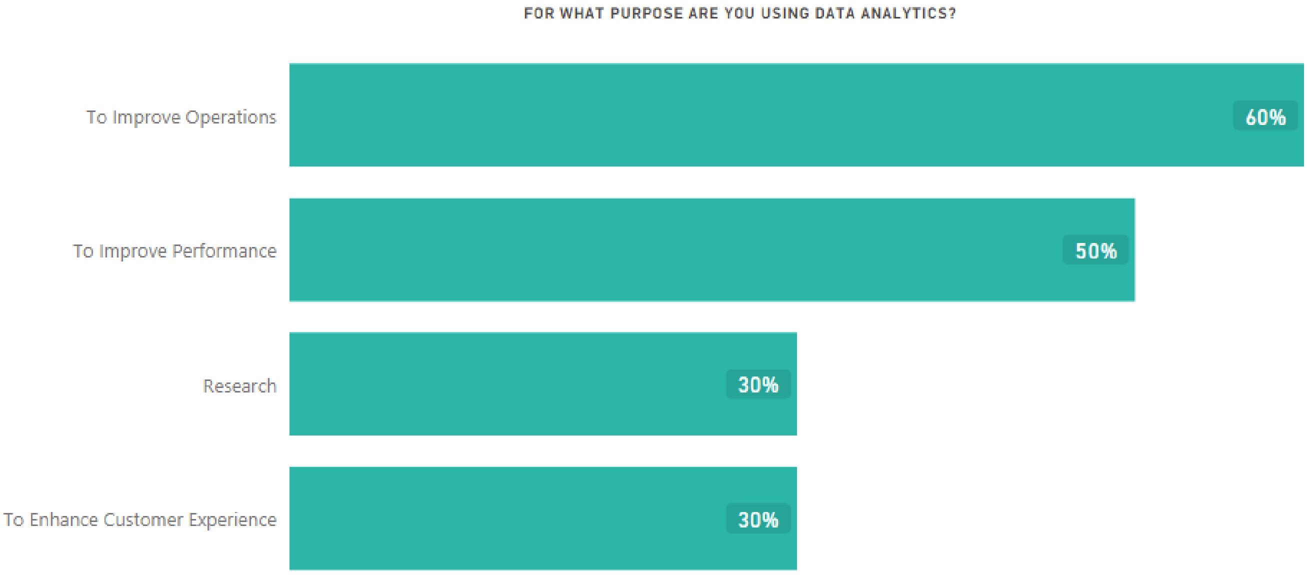


Figure 3-22: Purpose of Data Analytics

Figure 3-23 depicts the types of big data processing frameworks used by organizations. From this figure, we observe that in-house and open source big data processing frameworks are more commonly used than commercial frameworks (these numbers confirm the DZone's [170] report. Specifically, most of the large enterprises use in-house software for big data processing while most of the university/research organizations and SMEs use open source software for big data processing.

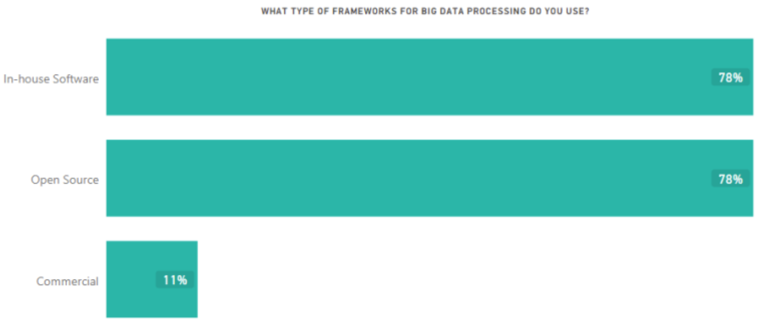


Figure 3-23: Type of Big Data Processing Frameworks

Apache Hadoop (67%), Apache Solr (33%) and Apache Spark (33%) are the most popular choices that are currently used for big data processing (Figure 3-24). These numbers confirm DZone's [170] reports.

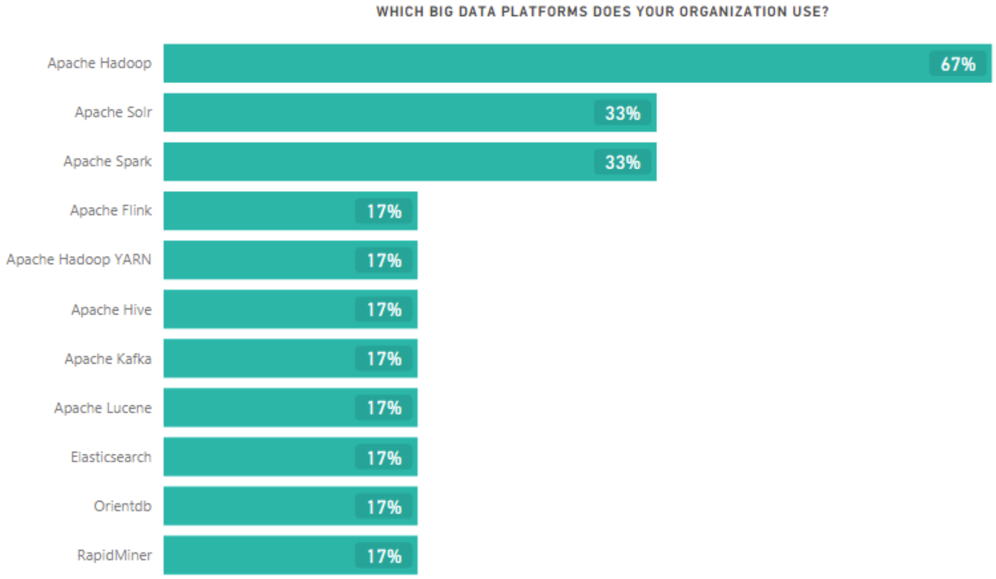


Figure 3-24: Popular Big Data Platforms

The majority of the participants (75%) use Python for data analytics while the second most popular choice (63%) is Java (Figure 3-25). Furthermore, C/C++ (50%) and MATLAB (50%) are the third most popular programming languages for data analytics while R (38%) is the fourth most popular. The popularity of Python is also confirmed by DZone’s [172] reports, however, the popularity of R is much lower.

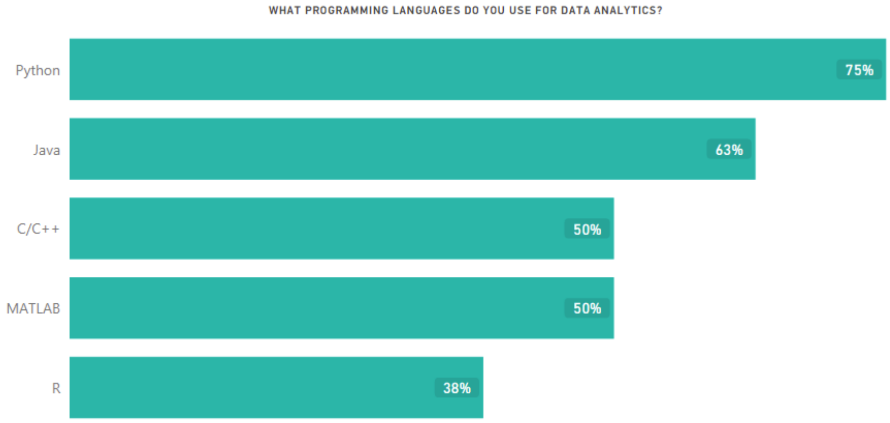


Figure 3-25: Popular Programming Language for Data Analytics

Most participants use TensorFlow (50%), Apache Spark MLlib (33%) and DeepLearning4j (33%) for data analytics (Figure 3-26). These numbers confirm DZone [170] reports.

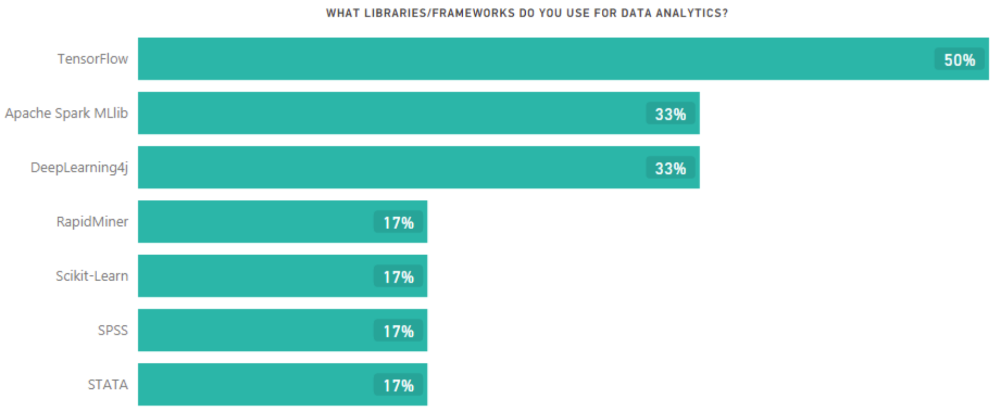


Figure 3-26: Popular Libraries/Frameworks for Data Analytics

The majority (57%) of participants do not use any tool for collaborative data analytics and visualization, while 43% of the participants use Jupyter Notebook (Figure 3-27). Particularly, most of SMEs and large enterprises do not use tools for collaborative data analytics and visualization, while most of the university/research organizations use Jupyter Notebook.

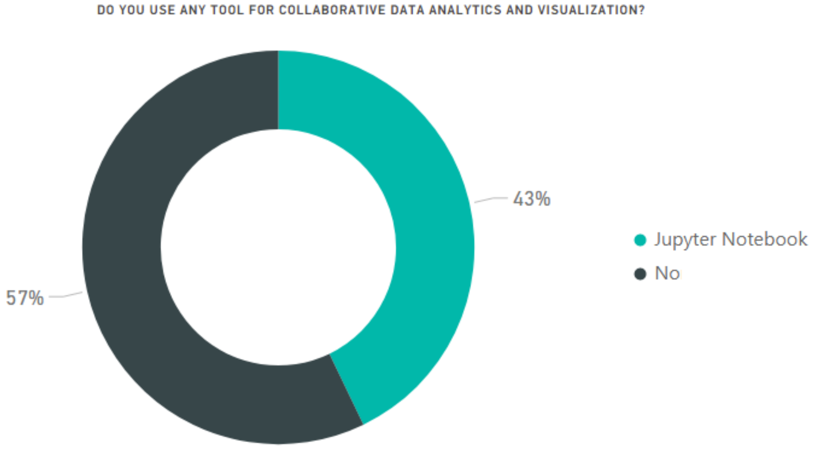


Figure 3-27: Usage of Collaborative Tools for Data Analytics and Visualization

Most participants use D3.js (75%), Chart.js (38%) and Tableau (25%) for data visualization (Figure 3-28). These numbers confirm DZone [170] reports.

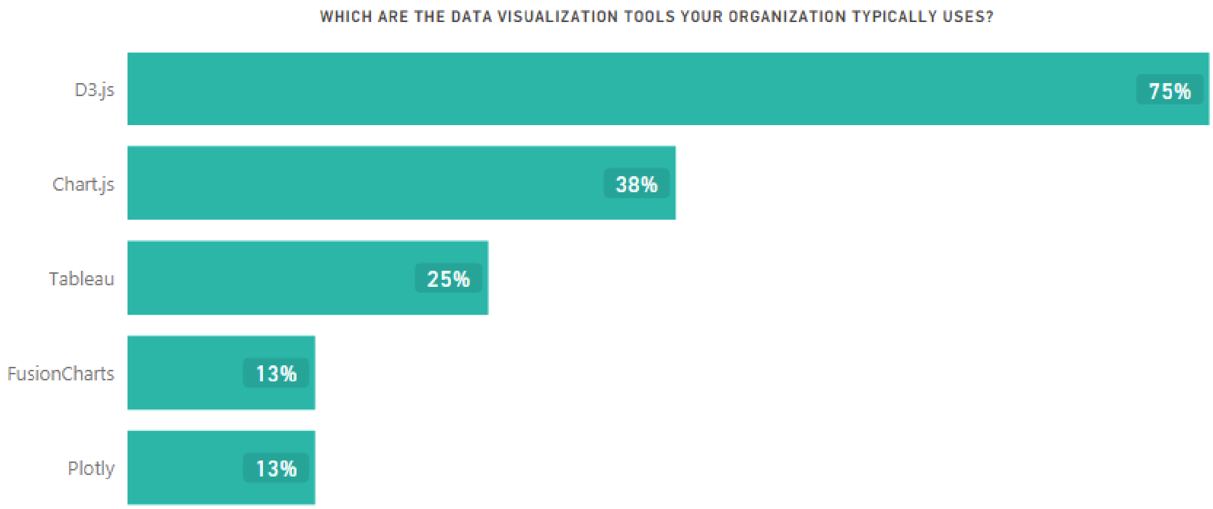


Figure 3-28: Popular Data Visualization Tools

3.3.4 ICARUS Platform

Regarding the functionalities of the ICARUS platform, **almost all of the participants are interested** in them (Figure 3-29). Particularly, the participants state the following for the functionalities of the ICARUS platform:

- **More than 85%** of the participants are interested in a platform that contains **a secure experimentation playground for experimenting with datasets** before purchasing them. In particular, all types of organizations are very interested in this functionality.
- **More than 82%** of the participants are interested in a platform that contains **a service that recommends similar datasets** based on the datasets currently explored. Particularly, transport organizations, university/research organizations and SMEs are very interested in this functionality.
- **More than 68%** of the participants are interested in a platform that contains a **data notification service** that permits any stakeholder to post requests for specific datasets. In particular, transport organizations and university/research organizations are interested in this functionality.
- **More than 78%** of the participants are interested in a platform that **guarantees specific agreements without intermediate third parties**. Particularly, transport organizations and university/research organizations are very interested in this functionality.
- **More than 82%** of the participants are interested in a platform that contains an **intuitive dashboard with interactive visualization capabilities**. In particular, transport organizations, university/research organizations and SMEs are very interested in this functionality.
- **More than 78%** of the participants are interested in a platform that contains **a semi-automated negotiation service** between data/service owners and prospective customers. In particular, transport organizations, university/research organizations and SMEs are interested in this functionality.

Therefore, it is clearly seen that the ICARUS stakeholders are **very interested** in **a marketplace for sharing data** which supports the previously mentioned functionalities.

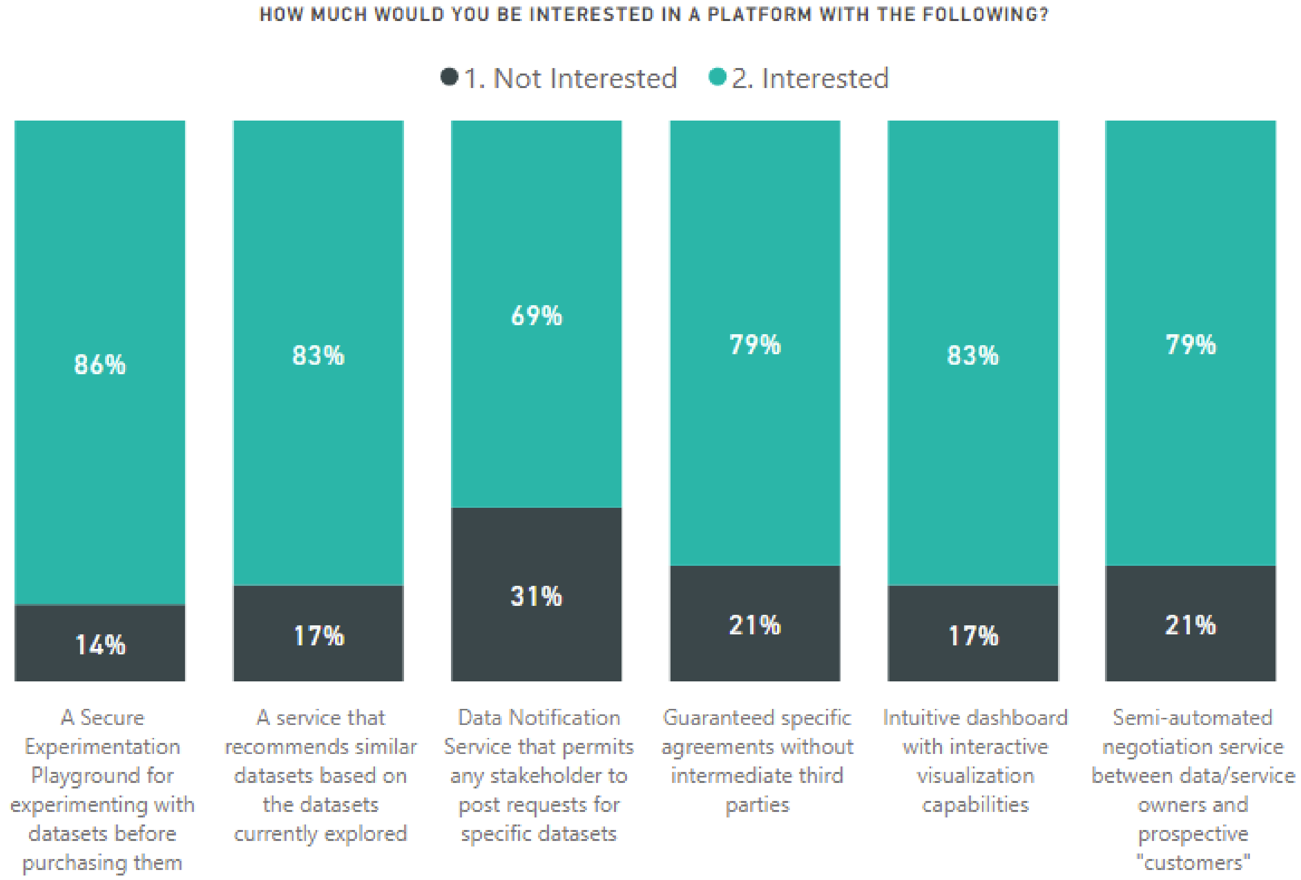


Figure 3-29: Interest in ICARUS Functionalities

Figure 3-30 depicts the interest of the participants regarding specific domains of data. From this figure, we observe that all data domains are popular, but the most popular domain of data is the airport data (93%). Furthermore, the environmental and the health/epidemics data are the least popular. After a further analysis, we observe the following:

- University/Research organizations are mostly interested in airport data and web data.
- SMEs are mostly interested in airport, aircraft and city/region data.
- Large enterprises are mostly interested in airport, aircraft and aviation business data.
- Transport organizations are mostly interested in airport, aircraft, aviation business and web data.

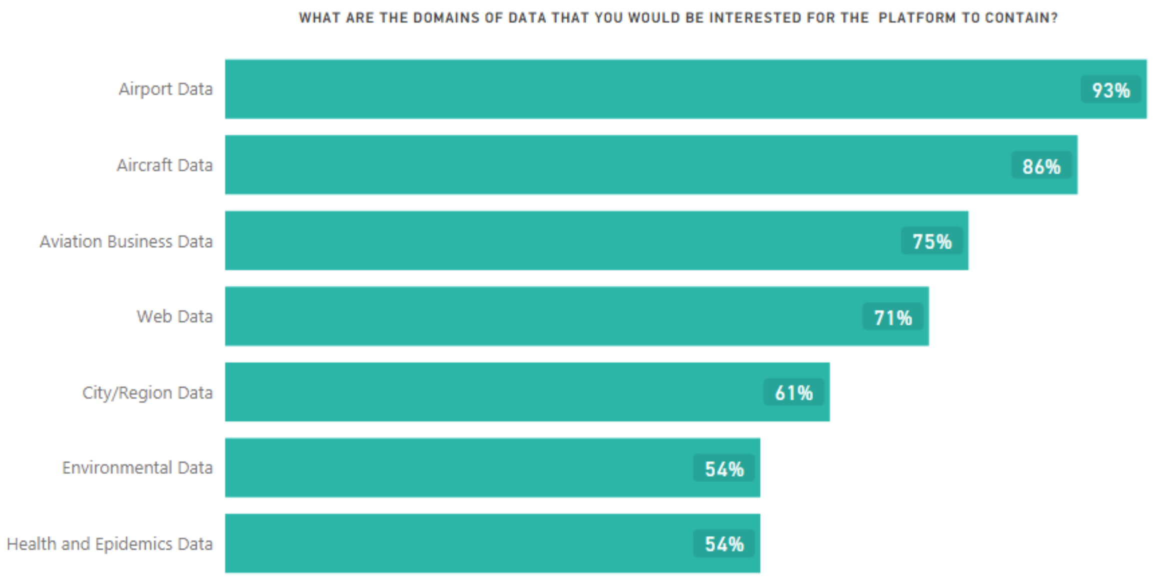


Figure 3-30: Interest in Specific Data Domains

Almost all participants are concerned about "privacy/confidentiality" (97%) and "security" (90%) issues when it comes to data sharing through an intermediary (Figure 3-31).

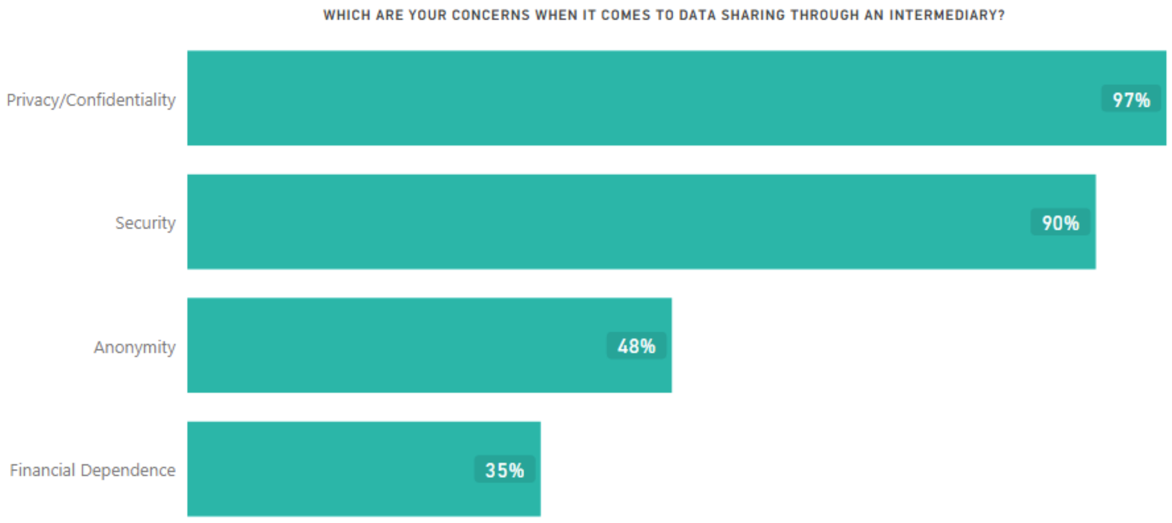


Figure 3-31: Concerns of Data Sharing Through an Intermediary

3.3.5 ICARUS Survey Key Findings

Table 3-3 summarizes the main key findings from the ICARUS survey.

Survey Sections	Key Findings
Respondents Profile	• The most difficult processes for organizations are the data anonymization and data linking. • 100% of the transport organizations state that it is not easy to collect data and 50% of the transport organizations find it difficult to visualize data. • 40% of the university/research organizations find it difficult to curate data while 100% of the transport organizations state that it is not easy to curate data.

	• 40% of the university/research organizations, 30% of the SMEs and 50% of the transport organizations find it difficult to anonymize their data.
Data collection	• Organizations that have already in place data linking mechanisms still find the data linking process very difficult. In particular, SMEs and transport organizations state that it would be valuable a mechanism for linking external available data. • Aviation related data that have high velocity and high volume, affect the difficulty of curating, visualizing, linking, analyzing, trading/sharing and anonymizing data. • Organizations that use "data marketplaces" and "APIs" find it easier to collect data while organizations that use "custom in-house mechanisms" find it harder. • Aircraft, airport and aviation business data are provided very often by organizations related to the aviation sector, while city/region and health/epidemics data are much rarer.
Data Analytics	• 50% of the participants state that they do not have architectures/platforms for big data analysis because of budget/cost constraints and lack of experience.
ICARUS platform	• 86% of the participants are interested in a platform that contains a secure experimentation playground for experimenting with datasets before purchasing them. • 83% of the participants are interested in a platform that contains a service that recommends similar datasets based on the datasets currently explored. • 69% of the participants are interested in a platform that contains a data notification service that permits any stakeholder to post requests for specific datasets. • 79% of the participants are interested in a platform that guarantees specific agreements without intermediate third parties. • 83% of the participants are interested in a platform that contains an intuitive dashboard with interactive visualization capabilities. • 79% of the participants are interested in a platform that contains a semi-automated negotiation service between data/service owners and prospective customers. • Almost all participants are concerned about "privacy/confidentiality" (97%) and "security" (90%) issues when it comes to data sharing through an intermediary.

Table 3-3: ICARUS Survey Key Findings

4 Aviation Datasets Collection, Protection, IPR and Brokerage

In this section, the aviation-related data assets that are owned by the ICARUS consortium or are available in open data repositories are documented in detail and appropriately classified in order to populate the 1st-2nd-3rd tiers of the ICARUS data value chain. As expected, such a documentation reflects the demonstrators' visibility at the beginning of the project and will be continuously updated in accordance with the project's advancements. In brief, the iterative approach that was followed during the data assets collection phase included the following steps: (a) Brainstorming session during the ICARUS kick-off meeting regarding the demonstrators' data availability and needs; (b) Preparatory data profiling by the demonstrators according to specific online templates, (c) Discussion on the preliminary data profiling (by the demonstrators and OAG) during the ICARUS Data Meeting that was held in Luton on March 13th, 2018, (d) Iterative updates on the profiling aspects for data available to the demonstrators (reported in sections 4.1.1-4.1.4) and to OAG (section 4.1.5), as well as for data needed by the demonstrators (presented in section 4.2); (e) Extensive search on different available data sources ranging from open data repositories to well-known aviation sources (as documented in section 4.3) in accordance with the preliminary demonstrators needs; (f) Assessment of the data assets by the demonstrators and cross-comparison of their different data needs. Finally, this section investigates the state-of-play with regard to the data protection and sharing aspects in order to extract key considerations for the next steps of the ICARUS project.

4.1 Aviation Datasets Collection from the ICARUS Consortium

In order to describe the aviation data assets in a homogeneous manner from the beginning of the ICARUS project, a common data profiling template has been adopted containing metadata to appropriate levels of detail in order to gain an initial understanding of the data assets that are to be handled in ICARUS. Such a multi-facet template practically features 6 core dimensions, namely:

- General Info, that contains a unique identifier (ID) for the data asset following the convention "DEM_no" (e.g. AIA_01), the title by which the data asset is formally known and a brief (free-text) description of the data asset.
- Data Asset Features elaborating on the expected scale of the data (Volume in terms of X GBs / records / transactions per hour / day / month), the different forms of the data (Variety at high level in terms of Structured / Unstructured / Semi-structured), the type of the data (e.g. Text / Image / Video / Audio / Other), the format in which the data are currently available (e.g. csv, xml, json, other), the speed at which data are generated or updated and become available for analysis (i.e. Real-time, Near Real-time, Batch), the availability of historical data along with their frequency (i.e. Hourly, Daily, Weekly, Monthly, Yearly, other) and their temporal and spatial coverage (in terms of time periods and locations which the data asset concerns), the language of the data asset, the relevant standards to which the data comply (i.e. exact ISO/IATA/... standards), the uncertainty and bias introduced in the data (Veracity in terms of identifying whether the data asset is raw as captured, pre-processed or processed as the output of an analysis), and the existing dependencies to other sources that have been linked to a data asset.

- Data Asset Availability in order to identify whether the data asset is owned or co-owned by the ICARUS consortium and / or is available from 3rd parties, and which are the mechanisms through which the data assets become accessible (e.g. through an API, as a downloadable file, as a database extract, etc.). In case a data asset is available from 3rd parties, the data asset provider is also named.
- Data Asset Rights that describes the privacy aspects whether a data asset is confidential (not to be shared), proprietary (to be shared with appropriate licensing) or public (available to all), the exact license that is applicable (e.g. CC Attribution-NonCommercial-ShareAlike (CC BY-NC-SA), or Case-by-Case Bilateral Agreement), the price at which a data asset is sold (per transaction, on a subscription or PAYG basis), and whether there is need for anonymization.

It needs to be noted that the profiling of the ICARUS data assets as documented in this section at a preliminary stage is expected to be maintained, monitored online and enriched throughout the project implementation. Such data assets are considered as the first assets that shall eventually populate the ICARUS platform (starting from its beta release).

4.1.1 Demonstrator 1 (AIA): Data Profiling

The AIA demonstrator has at its disposal six (6) data assets ranging from passengers to ground handling processes and are overall categorized as primary aviation data.

General Info		
ID	**Data Asset Title**	**Description**
AIA_01	Time stamps and status of ground handling processes	Aircraft turnaround related timestamps (e.g. On-Block, ATD, STA etc.).
AIA_02	Checked passengers per flight	Anonymous information regarding passengers that have scanned their boarding passes through automatic boarding pass control gates.
AIA_03	Expected passengers per flight	Provisional numbers of passengers per flight based on the declarations of the airlines.
AIA_04	Connecting passengers per flight	Information for transfer passengers per flight based on the declarations of the airlines.
AIA_05	Passengers who need assistance per flight	Number and category of PRM (Persons with Reduced Mobility) passengers per flight.
AIA_06	Gate open time	Gate related timestamps (e.g. boarding start time, final call) as the aircraft turnaround process is progressing.

Table 4-1: AIA Data Profiling – General Info

As indicated in Table 4-2, the AIA data assets range in terms of volume (from 500 to 50,000 records per day), while they are generally structured, referring to text data in English, and typically available in ASCII or JSON formats as raw data (without any kind of processing). There are certain data assets that are more static and become available at batch level (e.g. expected, connecting, PRM passengers per flight) whereas there are certain data assets to which real-time availability is critical for the airport-related stakeholders (e.g. ground handling processes, checked passengers, gate-related information).

There are daily historical records for all data assets that date back to 2001 when AIA started its operation or later, and refer to ATH (Athens International Airport, El. Venizelos). Whenever applicable, the data assets abide with the relevant IATA standards. It needs to be underlined that none of the data assets is already linked to other data sources.

General Info	Data Assets Features											
ID	Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Relevant Standards	Veracity	Dependency / Linking to Other Sources
AIA_01	8,000 records per day	Structured	Text	json	Real-time	Y	Daily	From 2001 to 2018	ATH	English	N.A.	Raw	N
AIA_02	50,000 records per day	Structured	Text	json	Real-time	Y	Daily	From 2001 to 2018	ATH	English	IATA	Raw	N
AIA_03	500 records per day	Structured	Text	ASCII	Batch	Y	Daily	From 2010 to 2018	ATH	English	N.A.	Raw	N
AIA_04	10,000 records per day	Structured	Text	ASCII	Batch	Y	Daily	From 2010 to 2018	ATH	English	IATA	Raw	N
AIA_05	500 records per day	Structured	Text	ASCII	Batch	Y	Daily	From 2008 to 2018	ATH	English	N.A.	Raw	N
AIA_06	1,000 records per day	Structured	Text	json	Real-time	Y	Daily	From 2001 to 2018	ATH	English	N.A.	Raw	N

Table 4-2: AIA Data Profiling – Data Assets Features

As defined in Table 4-3, the AIA data assets are typically owned by the airport (with the exception of AIA_02, AIA_03, AIA_04 that also involve airlines) and are accessible with different mechanisms: the core AIA-owned data assets are available via APIs while the data assets involving airlines are available as downloadable files or database extracts. With regard to the rights of the AIA data assets, they are characterized as proprietary data, for which bilateral agreements are reached on a case-by-case basis with the interested stakeholders and for a varying price depending on the purpose for which the data shall be used. Finally, with the exception of AIA_05 concerning the PRM passengers, there is no need for anonymization of the data assets.

General Info	Data Assets Availability				Data Assets Rights			
ID	**Data Asset Owned**	**Data Asset Available from 3rd Party**	**Data Asset Provider**	**Accessibility**	**Privacy**	**License**	**Pricing**	**Need for Anonymization**
AIA_01	Y	N	AIA	API	Proprietary	Case-by-Case Bilateral Agreement	Varied	N
AIA_02	Y	Y	Airlines	API	Proprietary	Case-by-Case Bilateral Agreement	Varied	N
AIA_03	N	Y	Airlines	Downloadable files	Proprietary	Case-by-Case Bilateral Agreement	Varied	N
AIA_04	N	Y	Airlines	Downloadable files	Proprietary	Case-by-Case Bilateral Agreement	Varied	N
AIA_05	Y	N	AIA	As database extract	Proprietary	Case-by-Case Bilateral Agreement	Varied	Y
AIA_06	Y	N	AIA	API	Proprietary	Case-by-Case Bilateral Agreement	Varied	N

Table 4-3: AIA Data Profiling – Data Assets Availability & Rights

4.1.2 Demonstrator 2 (PACE/TXT): Data Profiling

The PACE/TXT demonstrator has at its disposal two (2) data assets broadly concerning different types of cost for alternative routes in the same city and the overall aircraft costs, while being overall categorized as primary aviation data.

General Info		
ID	**Data Asset Title**	**Description**
PACE_01	Alternative routes comparison	A comparison of the alternative routes by fuel consumption, pollution, and airport fees for cities with more than one airport that shall be updated as a result of the demonstrator.
PACE_02	AC performance data	Aircraft cost (AC) performance data that are recorded with the PaceLab Mission Suite and could be used for AC comparison.

Table 4-4: PACE/TXT Data Profiling – General Info

As explained in Table 4-5, the PACE data assets are structured (as text data) and become available at batch level as ASCII files. There is no historical data availability of relevant data assets at the moment. They provide data for routes worldwide in English, and comply with proprietary standards developed by PACE/TXT. Such data represent the outcome of analysis conducted in the Pacelab Mission Suite and are thus characterized as processed while they are independent without being already linked to any other data sources.

	Data Asset Features												
ID	**Volume**	**Variety**	**Type**	**Format**	**Velocity**	**Historical Data Availability**	**Historical Data Frequency**	**Temporal Coverage**	**Spatial Coverage**	**Language**	**Relevant Standards**	**Veracity**	**Dependency / Linking to Other Sources**
PACE_01	N.A. (1 record per run)	Structured	Text	ASCII	Batch	N	N.A.	N.A.	World	English	Proprietary	Processed	N
PACE_02	20 MB	Structured	Text	ASCII	Batch	N	N.A.	N.A.	World	English	Proprietary	Processed	N

Table 4-5: PACE/TXT Data Profiling – Data Assets Features

As depicted in Table 4-6, both data assets are owned by PACE and are accessible at the moment either through a web interface or as downloadable files. Since they are proprietary data assets that are firmly associated with the PACE commercial offerings, there is a need for a case-by-case bilateral agreement for their disposal to 3rd parties.

	Data Asset Availability				Data Asset Rights			
ID	**Data Asset Owned**	**Data Asset Available from 3rd Party**	**Data Asset Provider**	**Accessibility**	**Privacy**	**License**	**Pricing**	**Need for Anonymization**
PACE_01	Y	N	PACE	Web Interface	Proprietary	Case-by-Case Bilateral Agreement	N.A.	N
PACE_02	Y	N	PACE	Downloadable files	Proprietary	Case-by-Case Bilateral Agreement	N.A.	N

Table 4-6: PACE/TXT Data Profiling – Data Assets Availability & Rights

4.1.3 Demonstrator 3 (ISI): Data Profiling

The ISI demonstrator has at its disposal three (3) data assets broadly concerning aviation-related or aviation-derived data in health like the population data, the virus data and the GLEAM simulation outputs.

General Info		
ID	**Data Asset Title**	**Description**
ISI_01	Population data	Data representing gridded estimates of the world population (http://sedac.ciesin.columbia.edu/data/collection/gpw-v4)
ISI_02	GLEAM Simulation output	Output data concerning the time evolution of GLEAM simulations: day by day number of individuals in each disease's compartment per census area, as long as new transitions between compartments and their cumulative number.
ISI_03	Virus & infections data	Data describing the number of influenza viruses detected, total number of influenza positive/negative viruses, ILI activity, etc., by different countries and influenza transmission zones (Influenza Laboratory Surveillance Information: http://apps.who.int/flumart/Default?ReportNo=12)

Table 4-7: ISI Data Profiling – General Info

As described in detail in Table 4-8, the ISI data assets significantly vary in terms of volume (from 1MB to 1GB per simulation) and are either structured (text) or semi-structured (as of other type). As the data are not real-time critical, each data asset becomes available at batch level in a certain format which is different depending on the case (e.g. HDF5, ASCII), but always in the English language. With the exception of ISI_02 that represents simulated data, there is significant availability of historical data at a worldwide level. In particular, ISI_01 that concerns population data and ISI_03 that refers to virus data are dated from 1995 (earliest data available for ISI_03) and 2020 (latest data available for ISI_01). Overall, the ISI data assets are released as processed data upon different types of pre-processing and analysis/simulations conducted, whereas they do not follow any specific international standards and are not already linked to any other data sources (although there are certain opportunities for data linking).

	Data Asset Features											
ID	**Volume**	**Variety**	**Type**	**Format**	**Velocity**	**Historical Data Availability**	**Historical Data Frequency**	**Temporal Coverage**	**Spatial Coverage**	**Language**	**Relevant Standards**	**Veracity**	**Dependency / Linking to Other Sources**
ISI_01	~ 250MB	Semi-structured	Text	ASCII	Batch	Y	Every five years	From 2000 to 2020	World	English	N.A.	Processed	N
ISI_02	~ 1GB per simulation	Semi-structured	Other	HDF5	Batch	NA (simulated data)	NA (simulated data)	NA (simulated data)	World	English	N.A.	Processed	N
ISI_03	~ 1MB	Structured	Text	CSV	Batch	Y	Yearly	From 1995 to 2018	World	English	N.A.	Processed	N

Table 4-8: ISI Data Profiling – Data Assets Features

As depicted in Table 4-9, the ISI data assets that are owned (ISI_02) will be available as downloadable files that are partly public and partly confidential. The exact license is dependent on bilateral agreements, but typically for free as ISI is a research institute. On the other hand, the rest of the ISI data assets are publicly distributed for free: ISI_01 under a Creative Commons Attribution 4.0 International License by Columbia University and ISI_3 by the World Health Organization (WHO). As the ISI identified data assets do not contain any personal data, there is no need for anonymization.

	Data Asset Availability				Data Asset Rights			
ID	**Data Asset Owned**	**Data Asset Available from 3rd Party**	**Data Asset Provider**	**Accessibility**	**Privacy**	**License**	**Pricing**	**Need for Anonymization**
ISI_01	N	Y	Columbia University	Downloadable files	Public	Creative Commons Attribution 4.0 International License	Free	N
ISI_02	Y	N	ISI	Downloadable files	Confidential/ Public	Case-by-case bilateral agreement	Free	N
ISI_03	N	Y	WHO	Downloadable files	Public	Public	Free	N

Table 4-9: ISI Data Profiling – Data Assets Availability & Rights

Table 4-10 also presents an indicative extract of the ISI data assets through the title, a brief description and the type of each attribute that they contain.

	Indicative Data Asset Extract
ID	**Indicative attributes**
ISI_01	Estimates of the world population: this dataset contains an array of floating point values corresponding to the estimated population for each cell of a grid, whose resolution is 15x15 arc minute, covering the Earth surface.
ISI_02	Time evolution of GLEAM simulations: The dataset contains arrays of floating point values representing the daily time series of the number of individuals per disease's compartment, and the number of new and cumulative transitions between compartments.
ISI_03	- Country, area or territory: string - WHO region (region defined by the World Health Organization): string - Influenza transmission zone: string - Year: integer - Week: integer - Start date: ISO 8601 date - End date: ISO 8601 date - Number of specimens: integer - Number of influenza A viruses detected by subtype: integer - Number of influenza B viruses detected by subtype: integer - Total number of influenza positive viruses: integer - Total number of influenza negative viruses: integer - ILI activity (activity of influenza-like illness): string

Table 4-10: ISI Data Profiling – Indicative Data Asset Extract

4.1.4 Demonstrator 4 (CELLOCK): Data Profiling

The CELLOCK demonstrator has at its disposal four (4) data assets broadly concerning core aviation data like passenger profiling data and extra-aviation data concerning retail and entertainment during a flight.

General Info		
ID	**Data Asset Title**	**Description**
CELLOCK_01	Retail and F&B in-flight sales	Onboard sales including food and beverages (F&B), as well as duty-free products.
CELLOCK_02	Number of Passengers	Number of Passengers in a flight collected/updated after a flight is completed.
CELLOCK_03	In-flight, IFE Passenger data	Data collected through the in-flight entertainment (IFE) system, such as phone type, operating system, age, gender, nationality.
CELLOCK_04	IFE Content data	Browsing history on the in-flight entertainment (IFE) system.

Table 4-11: CELLOCK Data Profiling – General Info

In terms of volume, the CELLOCK data assets range from 100 records per day (CELLOCK_02) to 8,000 records per day (CELLOCK_04) as depicted in Table 4-12. They are generally structured in JSON format and since they refer to data that become available after a flight lands, they are provided at batch level. Aggregated historical data at month level are available for flights within Europe. The language of all data assets is English. As with the rest of the ICARUS demonstrators, the data assets are not currently linked to other data sources.

	Data Asset Features												
ID	**Volume**	**Variety**	**Type**	**Format**	**Velocity**	**Historical Data Availability**	**Historical Data Frequency**	**Temporal Coverage**	**Spatial Coverage**	**Language**	**Relevant Standards**	**Veracity**	**Dependency / Linking to Other Sources**
CELLOCK_01	2,000 records/ day	Structured	Text	json	Batch	Y	Monthly	N.A.	Europe	English	N.A.	N.A.	N
CELLOCK_02	100 records / day	Structured	Text	json	Batch	Y	Monthly	N.A.	Europe	English	N.A.	N.A.	N
CELLOCK_03	4,000 records/ day	Structured	Text	json	Batch	Y	Monthly	N.A.	Europe	English	N.A.	N.A.	N
CELLOCK_04	8,000 records/ day	Structured	Text	json	Batch	Y	Monthly	N.A.	Europe	English	N.A.	N.A.	N

Table 4-12: CELLOCK Data Profiling – Data Assets Features

From Table 4-13, it can be easily noticed that the privacy of the available data assets is varying from proprietary and confidential to public. As the specific data assets are not available to other stakeholders at the moment, the licensing and pricing terms remain to be defined in due time. Since CELLOCK_03 and CELLOCK_04 though contain non-sensitive personal data (either aggregated or not), there is an explicit need for anonymization prior to their publication.

In addition, certain data assets are co-owned by CELLOCK and other data providers in the aviation data value chain (i.e. caterers) or owned by airlines. All data assets are accessible through APIs.

	Data Asset Availability				Data Asset Rights			
ID	**Data Asset Owned**	**Data Asset Available from 3rd Party**	**Data Asset Provider**	**Accessibility**	**Privacy**	**License**	**Pricing**	**Need for Anonymization**
CELLOCK_01	Y	Y	Caterer	API	Proprietary	TBD	TBD	N
CELLOCK_02	Y	Y	Caterer	API	Public	TBD	TBD	N
CELLOCK_03	N	Y	Airline	API	Confidential	TBD	TBD	Y
CELLOCK_04	N	Y	Airline	API	Confidential	TBD	TBD	Y

Table 4-13: CELLOCK Data Profiling – Data Assets Availability & Rights

4.1.5 OAG Aviation Data

OAG, a core data provider in ICARUS, has the world's largest network of air travel data, including the definitive schedules database of more than 900 airlines and over 4,000 airports, and the most extensive flight status information database in the market. At a glance, OAG handles more than 52 million records of flight status updates per year, processes 1,4 billion requests and continues to deliver in excess of 35 million dynamic flight status updates daily.

Table 4-14 provides a glimpse of different data assets that are created by OAG and currently become available through 3 different OAG products, namely:

- OAG Analytics in order to monitor airline frequency and capacity trends, identify new routes and services, understand passenger traffic flows and evaluate airline connection performance. It contains a number of analytics modules, such as schedules analyzer, connections analyzer, traffic analyzer, mapper, DOT analyzer, market intelligence, and Apex.
- OAG Flightview providing accurate, timely and real-time flight status information through flight data feeds and APIs, websites, historical flight status reports and digital displays to ensure the most relevant day-of-travel information is readily available.
- OAG Schedules delivering solutions to help airlines and airports and related services drive growth and performance.

The OAG data assets that are presented in the following tables (4-14 and 4-15) are compiled based on data provided by different stakeholders in the aviation data value chain, mainly airlines and airports.

ID	Data Asset Title	Description
OAG_01	Schedules	Planned carrier schedules looking forward (up to almost 1 year ahead depending on the airline).
OAG_02	Flight Status	Live feeds of flight status information for an individual flight or all flights on an airline to/from specific airports. The scope of flights can be worldwide, by countries, by airport(s) or by airline(s). In addition, historical data detailed information on scheduled and actual departure and arrival times, delay and cancellation information (e.g. irregular operations or events), terminal, gate, baggage claim, aircraft equipment and codeshare information, is available.
OAG_03	Flight Tracking	Live positional data as well as continually updated delay status information on inbound and departing flights at hundreds of airports, for US market only.
OAG_04	Carrier File	Airline names, codes and decodes, domicile country.
OAG_05	Locations File	City, port, names codes and decodes, longitude and latitude.
OAG_06	MCT (Minimum Connection Times)	Details on Minimum Connection Times (MCT) as directly supplied (and updated daily) by airlines for connection generation.
OAG_07	DST (Daylight Saving time)	Details regarding seasonal time changes.
OAG_08	Country File	Country names, codes and decodes.
OAG_09	Connections	Pre-built connections data file combing the Schedules data (OAG_01) and MCT data (OAG_06), including airline-specific exceptions. Missed connections are also available at a gateway airport for scheduled flights and potential flights (phantom flights).
OAG_10	OTP (On Time	Analysis carried out using Flight Status – Historical data providing

ID	Data Asset Title	Description
	Performance)	flight performance times to schedule by airline, airport or specific flights over time.

Table 4-14: OAG Data Profiling – General Info

As presented in Table 4-15, the OAG data assets are typically characterized as structured, cleansed data, containing text in English in csv and txt formats. All data assets owned by OAG are available as downloadable files with the exception of OAG_2 (Flight Status Information) that is also available via APIs and XML Web Services. With regard to the volume of the OAG data assets, it presents significant variations, from 24 million rows per day (OAG_01) to 120,000 rows per day (OAG_06). Most of the OAG data assets are provided at batch level (OAG_03 – OAG_10) with the exception of OAG_01 and OAG_02 for which there is a need for real-time and near real-time provisioning as well.

There is a plethora of historical data over the years for all data assets at different frequency (aggregated or not, on a per minute, per week, per month and per flight basis), with the schedules (OAG_01) being available since 1969, certain data assets (OAG_04 – OAG_10) since 1996 and the first trails for the rest of the data assets set in 2012. Such historical and current data from OAG typically have a worldwide coverage with the exception of OAG_03 in which flight tracking is currently available only for the US.

Since all data assets contain aggregated information from many data providers in the aviation data value chain, OAG undertakes their cleansing, curation and processing prior to their storage, thereby they are not provided as raw data, but as processed data only.

	Data Asset Features												
ID	Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Relevant Standards	Veracity	Dependency / Linking to Other Sources
OAG_01	Batch up to 4.5GB, 24m rows per day	Structured	Text	.csv,.txt	Real-time, Batch	Y	Weekly	1969 - current day (various formats)	Early Data North America, then Worldwide	English	N.A.	Processed	None
OAG_02	Approx. 4 deliveries per minute containing on average 350 transactions	Structured	Text	.csv,.txt, .xml	Real-time, Near Real-time, Batch	Y	Varies - up to by minute	2012-	Worldwide	English	N.A.	Processed	None
OAG_03	Updates every 3 minutes	Structured	Text	.csv,.txt	Batch	Y	Final status of each flight (op and non/op)	2012-	US	English	N.A.	Processed	None
OAG_04	1000 rows	Structured	Text	.csv,.txt	Batch	Y	Monthly	1996-	Worldwide	English	N.A.	Processed	None
OAG_05	12000 rows	Structured	Text	.csv,.txt	Batch	Y	Monthly	1996-	Worldwide	English	N.A.	Processed	None
OAG_06	120000 rows	Structured	Text	.csv,.txt	Batch	Y	Monthly	1996-	Worldwide	English	N.A.	Processed	None
OAG_07	500 rows	Structured	Text	.csv,.txt	Batch	Y	Monthly	1996-	Worldwide	English	N.A.	Processed	None
OAG_08	300 rows	Structured	Text	.csv,.txt	Batch	Y	Monthly	1996-	Worldwide	English	N.A.	Processed	None
OAG_09	Varies	Structured	Text	.csv,.txt	Batch	Y	Monthly	1996-	Worldwide	English	N.A.	Processed	None
OAG_10	Varies	Structured	Text	.csv,.txt	Batch	Y	Monthly	1996-	Worldwide	English	N.A.	Processed	None

Table 4-15: OAG Data Profiling – Data Assets Features

With regard to the data asset rights, OAG follows a specific data agreement template. The data licenses are usually valid for a predefined time period (e.g. 5 years) and specify in detail flexible terms for the frequency, the type, and the fields of the data purchased. It needs to be noted that OAG is bound not to sell its data assets to airlines. Finally, none of the OAG data assets is currently linked to any external data sources.

	Data Asset Availability				Data Asset Rights			
ID	Data Asset Owned	Data Asset Available from 3rd Party	Data Asset Provider	Accessibility	Privacy	License	Pricing	Need for Anonymization
OAG_01	Y	Y	Compiled by data collected by airlines	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_02	Y	Y	Compiled by data collected by airlines	API, Web Service, Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_03	Y	Y	Compiled by data collected by airlines	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_04	Y	Y	Compiled by data collected by airlines	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_05	Y	Y	Compiled by data collected by airlines	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_06	Y	Y	Compiled by data collected by airlines	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_07	Y	N	OAG	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_08	Y	N	OAG	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_09	Y	N	OAG	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N
OAG_10	Y	N	OAG	Downloadable files	Proprietary	OAG Data Contract per Client	Depending on terms	N

Table 4-16: OAG Data Profiling – Data Assets Availability & Rights

4.2 Initial Data Needs of the ICARUS Demonstrators

Although the ICARUS demonstrators' scenarios have not yet been consolidated (in WP5) at such an early phase of the project implementation, the demonstrators already have identified a number of data assets they would like to have access to, link to other data sources and perform analytics. Such data assets have been identified in detail following a similar data profiling template as in Section 4.1.

As it can be easily noted from Table 4-17, the data assets requested by the demonstrators shall be investigated in detail by the consortium (especially OAG) to help identify appropriate data sources, whenever such data are not already available by OAG (as it is the case for AIA_DR_36 and AIA_DR_37, for example) or directly provided by airlines and ground handlers. The open data requests on their behalf are mostly related to weather data and environmental data (for which data sources are identified in section 4.3).

A data asset, namely AIA_DR_20, is discretely deleted in order to indicate that it is no longer needed by AIA and that the evolution of the data assets is naturally already happening, yet the ICARUS consortium keeps track of all changes in the data assets availability and requirements.

ID	Data Asset Title	Description	Potential Data Asset Provider[2]
AIA_DR_01	Passengers on board (how many boarded already, how many remain)	Details on the boarding time process and the use of gate.	Airline/Ground Handler/OAG
AIA_DR_02	Cabin cleaning (start-end times)	Aircraft interior cleaning during turn-round process	Airline/Ground Handler/OAG
AIA_DR_03	Catering (Start-end time)	Catering loading/unloading, during turn-round process	Airline/Ground Handler/OAG
AIA_DR_04	Unloading bags - cargo - total pieces	Loading-unloading process times during turn-round process	Airline/Ground Handler/OAG
AIA_DR_05	Refueling operation - refuel track	Start and end time of refueling or defueling operation	Airline/Ground Handler/OAG
AIA_DR_06	APU-ASU	Ground service equipment	Airline/Ground Handler/OAG
AIA_DR_07	Pushback tractor (type, power)	Arrival time at the stand, aircraft type that can serve	Airline/Ground Handler/OAG
AIA_DR_08	Lavatory service	Start and end time during turn-round process	Airline/Ground Handler/OAG
AIA_DR_09	Loading bags - cargo	Loading-unloading process times during turn-round process	Airline/Ground Handler/OAG
AIA_DR_10	Wheelchairs	Total number and type of wheelchairs expected for the specific flight	Airline/Ground Handler/OAG
AIA_DR_11	Deicing info	Expected start time, chemicals, place, de-anti icing	Airline/Ground Handler/OAG
AIA_DR_12	Flight update messages	Information provided by the ATC during flight	Airline/Ground Handler/OAG
AIA_DR_13	ACARS data	Information provided by the Aircraft Communications, Addressing and Reporting System during flight	Airline/Ground Handler/OAG

[2] Including assistance in Locating Appropriate Data Provider

ID	Data Asset Title	Description	Potential Data Asset Provider[2]
AIA_DR_14	Actual commence of ground handling time - Actual end	Ramp handling activities during turn-round process (baggage loading/unloading, cargo loading/unloading, aircraft cleaning, fueling-defueling, catering)	Airline/Ground Handler/OAG
AIA_DR_15	Actual commence of deicing time	Expected start time, chemicals, place, de-anti icing	Airline/Ground Handler/OAG
AIA_DR_16	Actual ready time	Pushback attached on the aircraft	Airline/Ground Handler/OAG
AIA_DR_17	Minimum turn-around time agreed between AO/??	Minimum times defined by the airlines with full passenger and de-load movement include the time to start engines	Airline/OAG
AIA_DR_18	Airlines schedule-planning	Schedules planning per airline	Airline/OAG
AIA_DR_19	Aircraft movement data	Aircraft movement data in the airport	Airline/Ground Handler/OAG
~~**AIA_DR_20**~~	~~WX conditions at flight levels above 18000 feet~~	~~Weather conditions above FL180 can be provided only by aircraft weather radar or ATC~~	
AIA_DR_21	Engine type of Aircraft Operator fleet	During engine test time, test bed usage. ICAO Annex 16 Environmental Protection, Volume II — Aircraft Engine Emissions	Airline/OAG
AIA_DR_22	Aircraft noise level	Reason of request for better structure and separation from ANSP, during approach and take-off phase.	Airline/OAG
AIA_DR_23	Aircraft Dimensions	Aircraft specifications (wingspan, length, weight, height etc.)	Airline/OAG
AIA_DR_24	Aircraft Operator fleet seating capacity	Aircraft seating capacity	Airline/OAG
AIA_DR_25	Aircraft Operator delay codes (internal)	IATA delay codes	Airline/OAG
AIA_DR_26	Aircraft Operator fleet MTOW	Aircraft specification	Airline/OAG
AIA_DR_27	No of passengers using CIP lounges	Identify profile of passengers	Airline/Ground Handler/OAG
AIA_DR_28	No of passengers entitled fast lane access	Volume of users, operational effectiveness	Airline/Ground Handler/OAG
AIA_DR_29	No of online check-in passengers who have drop-off luggage	Operational effectiveness at check-in	Airline/Ground Handler/OAG
AIA_DR_30	Scheduled check-in opening/closing time	As per airline internal procedures at each airport	Airline/Ground Handler/OAG
AIA_DR_31	Scheduled gate closing time	Operational effectiveness in use of gates	Airline/Ground Handler/OAG
AIA_DR_32	No of passengers visiting lost&found booths	Airport infrastructure effectiveness	Airline/Ground Handler/OAG
AIA_DR_33	Web-CUSS-mobile-airport check-in	Number of passengers checked in Web-CUSS-mobile-airport for operational effectiveness at check-in	Airline/Ground Handler/OAG
AIA_DR_34	Expected passenger loads (provisional)	Expected passenger loads (provisional) for operational effectiveness	Airline/Ground Handler/OAG
AIA_DR_35	Connecting passengers per flight	Connecting passengers per flight for operational effectiveness	Airline/Ground Handler/OAG
AIA_DR_36	Route transfer flights	Identify all connecting routes from/to each airport	Airline/Ground Handler/OAG
AIA_DR_37	Route trends	Categorization and trends for all routes	Airline/OAG
AIA_DR_38	Duty-free shopping analytics	Volume - type of products purchased, per flight, per nationality etc.	Concessionaires/OAG?

ID	Data Asset Title	Description	Potential Data Asset Provider[2]
AIA_DR_39	Non duty-free shopping data	Volume - type of products purchased, per flight, per nationality etc.	Concessionaires/OAG?
AIA_DR_40	Car parking data	Data regarding the car parking service that is not pre-booked online, but on the spot.	Concessionaires/OAG?
AIA_DR_41	Passenger Profiles	Nationality, Gender, Age, Frequent Flyer data per passenger	Airline/Ground Handler/OAG
PACE_DR_01	Airport Data	Airport specifications including but not limited to name, IATA, ICAO, Elevation, Latitude, Longitude, Type of Airfield, APU Time, Taxi time, Runway PCN, TORA, TODA, ASDA, LDA, Obstacles	Jeppesen
PACE_DR_02	Airfield Weather Data	Statistical airport temperatures	Boeing
PACE_DR_03	En-route Weather Data	Statistical en-route temperatures and winds	Boeing
PACE_DR_04	Typical Airport Departure Paths	Paths through the city/area for departure, Stars & Sids	Jeppesen, OAG
PACE_DR_05	Typical Airport Arrival Paths	Paths through the city/area for landing, Stars & Sids	Jeppesen, OAG
PACE_DR_06	Typical Airport Taxi In/Out Times	Taxi time, fuel burn on ground,	OAG
PACE_DR_07	Statistical operational costs	To compare the costs between different airports, taxi times, power costs, etc.	OAG
ISI_DR_01	Passenger stratification	Number of passengers per trip (booking from source airport to destination airport) aggregated by age and gender.	Booking company. Amadeus (?).
ISI_DR_02	Length of stay	Estimate of the time spent at destination by travelers (distribution of). This can be inferred using return tickets as a proxy.	OAG? (booking info about return tickets can be a good starting point)
ISI_DR_03	Recurring travelers	Estimate distribution of travelers doing multiple trips to the same destination (frequency/duration/...)	Booking companies (?)
ISI_DR_04	Historical passenger data	Number of passengers per trip (source to destination) over multiple years, allowing to infer seasonality patterns on the various routes	OAG
ISI_DR_05	Travelers' wealth indicators	Information about the home address of passengers could be used as a proxy for wealth status, which has been shown to be an important factor estimating the incidence of some infectious diseases	Booking companies (?)
CELLOCK_DR_01	Weather data	Weather data on each destination	Open data
CELLOCK_DR_02	Passenger demographics (pre-flight)	Passengers demographics based on booking engines	Airlines, Booking companies
CELLOCK_DR_03	Aircraft flight routes	Flight Schedule	OAG
CELLOCK_DR_04	Airport retail data	Sales from airport duty free shops	Airport, Duty free shops
CELLOCK_DR_05	Flight Delays	Flight delays / cancelations	Airports, Airlines, OAG

Table 4-17: ICARUS Demonstrators Data Needs Profiling – General Info

In Table 4-18, the desired features of the data assets are reflected in detail. As expected, there are different expectations per demonstrator for real-time data (mainly in the cases of AIA and CELLOCK) and batch data (mostly for PACE/TXT and ISI) yet the historical data availability is a prerequisite in all cases. Such historical data are to be available for at least 1 recent year or for minimum 2 years, with the exception of ISI_DR_04 that requests 5-10 years in order to get an appropriate mass of data for analytics. The expected volume varies per data asset, but it provides a rough estimation of the scale of the data that the demonstrators need to get their hands on. In terms of variety, most data assets should be structured, semi-structured or in any form. The preferable formats are naturally machine-readable formats like json and xml. Finally, the veracity of the data tends to weight in towards pre-processed data at the moment.

ID	Expected Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Veracity
AIA_DR_01	Daily, weekly or monthly aggregations per trip. Each aggregation about 200000 records(?).	Structured	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Global	English (but not relevant)	Pre-processed
AIA_DR_02	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_03	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_04	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_05	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_06	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed

ID	Expected Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Veracity
AIA_DR_07	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_08	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_09	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_10	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_11	1000 records/messages per day during adverse weather condition period	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_12	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_13	100.000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_14	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_15	1000 records/messages per day during adverse weather condition period	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed

ID	Expected Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Veracity
AIA_DR_16	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_17	Reference Data 1000 records per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Yearly		Any	English	Pre-processed
AIA_DR_18	200 records/messages per day	Any	Text	json/xml	Batch (previous day)	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_19	2000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
~~AIA_DR_20~~	~~N.A.~~	~~Any~~	~~Text~~	~~json/xml~~	~~Any~~	~~Y~~	~~Any~~	~~Any~~	~~Any~~	~~English~~	~~Pre-processed~~
AIA_DR_21	Reference Data 1000 records per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Yearly	Minimum two last years	Any	English	Pre-processed
AIA_DR_22	Reference Data 1000 records per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Yearly	Minimum two last years	Any	English	Pre-processed
AIA_DR_23	Reference Data 1000 records per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Yearly	Minimum two last years	Any	English	Pre-processed
AIA_DR_24	Reference Data 1000 records per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Seasonal (every 6 months)	Minimum two last years	Any	English	Pre-processed

ID	Expected Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Veracity
AIA_DR_25	100 records per airline approx 2000 per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Seasonal (every 6 months)	Minimum two last years	Any	English	Pre-processed
AIA_DR_26	Reference Data 1000 records per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Seasonal (every 6 months)	Minimum two last years	Any	English	Pre-processed
AIA_DR_27	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_28	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_29	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_30	1000 records/messages per day	Any	Text	json/xml	Seasonal (every 6 months)	Y	Seasonal (every 6 months)	Minimum two last years	Any	English	Pre-processed
AIA_DR_31	1000 records/messages per day	Any	Text	json/xml	Seasonal (every 6 months)	Y	Seasonal (every 6 months)	Minimum two last years	Any	English	Pre-processed
AIA_DR_32	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_33	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_34	1000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	at least for one year (recent)	Any	English	Pre-processed
AIA_DR_35	1000 records/message	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed

ID	Expected Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Veracity
	s per day										
AIA_DR_36	1000 records/messages per day	Any	Text	json/xml	Seasonal (every 6 months)	Y	Seasonal (every 6 months)	Minimum two last years	Any	English	Pre-processed
AIA_DR_37	500 records/messages per season	Any	Text	json/xml	Seasonal (every 6 months)	Y	Seasonal (every 6 months)	Minimum two last years	Any	English	Pre-processed
AIA_DR_38	10000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_39	10000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_40	N.A.	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
AIA_DR_41	50000 records/messages per day	Any	Text	json/xml	Real time	Y	Daily, weekly or monthly	At least for one year (recent)	Any	English	Pre-processed
PACE_DR_01	N.A.	Structured	Text	xml	Batch	N	N.A.	Any	Global	English	Any
PACE_DR_02	N.A.	Structured	Text	other	Batch	Y	Monthly/yearly	Any	Global	English	Any
PACE_DR_03	N.A.	Structured	Text	other	Batch	Y	Monthly/yearly	Any	Global	English	Any
PACE_DR_04	N.A.	Structured	Text	other	Batch	Y	Monthly/yearly	Any	Global	English	Any
PACE_DR_05	N.A.	Structured	Text	other	Batch	Y	Monthly/yearly	Any	Global	English	Any

ID	Expected Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Veracity
PACE_DR_06	N.A.	Structured	Text	other	Batch	Y	Monthly/yearly	Any	Global	English	Any
PACE_DR_07	N.A.	Structured	Text	other	Batch	Y	Monthly/yearly	Any	Global	English	Any
ISI_DR_01	Daily, weekly or monthly aggregations per trip. Each aggregation about 200000 records(?).	Structured	Text	Any	Batch	Y	Daily, weekly or monthly	at least for one year (recent)	Global	English (but not relevant)	Pre-processed
ISI_DR_02	Depends on the format.	Structured	Text	Any	Batch	Y	Depends on format. Daily granularity.	at least for one year (recent)	Global	English (but not relevant)	Pre-processed
ISI_DR_03	Depends on the format.	Structured	Text	Any	Batch	Y	Yearly aggregation	at least for one year (recent)	Global	English (but not relevant)	Pre-processed
ISI_DR_04	Weekly or monthly aggregations per trip. Each aggregation about 200000 records(?).	Structured	Text	Any	Batch	Y	Weekly or monthly	5 to 10 years	Global	English (but not relevant)	Pre-processed
ISI_DR_05	Depends on the aggregation level and format.	Structured	Text	Any	Batch	Y	Monthly or finer.	at least for one year (recent)	Global	English (but not relevant)	Pre-processed
CELLOCK_DR_01	N.A.	Structured	Text	json	Real time	Y	Any	Any	Any	English	Any
CELLOCK_DR_02	N.A.	Structured	Text	json	Real time	Y	Any	Any	Any	English	Any
CELLOCK_DR_03	N.A.	Structured	Text	json	Real time	Y	Any	Any	Any	English	Any
CELLOCK_DR_04	N.A.	Structured	Text	json	Real time	Y	Any	Any	Any	English	Any

ID	Expected Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Veracity
CELLOCK_DR_05	N.A.	Structured	Text	json	Real time	Y	Any	Any	Any	English	Any

Table 4-18: ICARUS Demonstrators Data Needs Profiling – Data Assets Desired Features

It is acknowledged by the demonstrators that the data that are needed are typically proprietary, so they would be eventually willing to pay a subscription fee or follow a PAYG pricing mode in the future in order to get access to the assets, preferably via APIs. The preliminary intentions for analysis of the data assets, apart from liking to other available data sources, is to perform pattern and trend analysis. It is worth noting that anonymization concerns are not generally raised since the demonstrators expect pre-processed data, e.g. aggregated distributions. However, if raw data, from which to derive the distributions, are taken into consideration, anonymization would be then required.

4.3 Aviation Datasets Collection beyond the ICARUS Consortium

4.3.1 Aviation Data Sources

An extensive search was performed on March 16th, 2018 to identify data sources and collect data sources that are related to aviation. The criteria that were used for the search include: (a) appearance of one of the keywords “aviation data”, “aerospace data”, or “air traffic data”, and (b) focus on the pan-European coverage (instead of country-specific data sources or worldwide data sources). It needs to be noted that additional aviation data sources include the Civil Aviation Authorities per country, yet they are not considered in the following table as the pan-European data sources were prioritized for this release.

The data sources that were identified are mostly collected and released by EUROCONTROL and range from the European AIS Database and the Network Operations Portal to broader aviation statistics published by Eurostat. Additional data sources that are popular in the aviation domain, particularly for the flight deck, include the Jeppesen[3] Standard Terminal Arrival Route (STAR) and Standard Instrument Departure (SID), Departure, and Arrival charts on the flight deck, as well as Landing and Take-off Minimums on Approach and Airport chart, yet they are not analyzed in the following tables as they are proprietary reports for which the detailed features are not publicly available.

[3] http://ww1.jeppesen.com/company/publications/publications.jsp

ID	Data Source	Description	Datasets Available
ADS_01	Eurostat Air Transport Statistics[4]	A collection of Air Transport Statistics is based on Regulation (EC) No 437/2003 of the European Parliament and of the Council of 27 February 2003, on statistical returns in respect of the carriage of passengers, freight and mail by air as well as the subsequent implementing Commission Regulations 1358/2003, 546/2005 and 158/2007. Data are supplied by all Member States + EFTA countries (NO, CH and IS). Some CC countries are also participating in this data collection (TR, MK, MN and SB).	• Air transport infrastructure • Air transport equipment • Air transport-Enterprises, economic performances and employment • Air transport measurement - passengers • Air transport measurement - freight and mail • Air transport measurement - traffic data by airports, aircrafts and airlines • Air transport - regional statistics
ADS_02	European AIS (Aeronautical Information Service) Database (EAD)[5]	A centralized reference database of quality-assured aeronautical information and, simultaneously, a fully integrated, state-of-the-art AIS solution. It offers instant access to the most up-to-date digital aeronautical information from the European Civil Aviation Conference (ECAC) area, NOTAM (Notices to Airmen), Pre-flight Information Bulletins (PIBs) from around the world.	• International Notice to Airmen (NOTAM) Operations (INO): Original NOTAM, SNOWTAM and ASHTAM • Static Data Operations (SDO): Full set of aeronautical information data published in AIP, i.e. Aerodrome information, including Procedures and Obstacles; En-route information such as Airspaces, Routes, Navaids and Waypoints; General information such as Organization, Authority and Units. • Published Aeronautical Information Publication (AIP) Management System (PAMS): AIP, Amendments, Supplements, AIC and Charts.
ADS_03	Eurocontrol Network Operations Portal (NOP)[6]	A centralized portal to monitor the real-time status of traffic, airspace and air traffic flow and capacity management (ATFCM) measures and to collaboratively plan pan-European operations from the strategic to the tactical phases, thus optimizing the use of available ATM capacity.	• Air Traffic Flow and Capacity Management (ATFCM) Network Situation Data: High level indicators at Network level on the real time status of: ○ Traffic ○ Delays ○ Delay causes ○ Slot windows compliance ○ Suspended flights • Daily Eurocontrol Network Weather Assessment
ADS_04	ICAO Engine Emissions Databank[7]	Information on exhaust emissions of production aircraft engines, measured according to the procedures in ICAO Annex 16, Volume II, as provided by the engine manufacturers, who are solely responsible for its accuracy. The databank	• ICAO Engine Emissions. Updated yearly. Last version: 11/2017

[4] http://ec.europa.eu/eurostat/web/transport/data/database
[5] http://www.eurocontrol.int/articles/european-ais-database-ead
[6] https://www.public.nm.eurocontrol.int/PUBPORTAL/gateway/spec/index.html
[7] https://www.easa.europa.eu/easa-and-you/environment/icao-aircraft-engine-emissions-databank

ID	Data Source	Description	Datasets Available
		covers engine types whose emissions are regulated, namely turbojet and turbofan engines with a static thrust greater than 26.7 kilonewtons.	
ADS_05	European Aviation Environmental Report[8]	Information and data collected by EASA, EEA and EUROCONTROL to evaluate the environmental performance of the European aviation sector.	Aggregated data appearing in figures: • Total flights 2005-2014 • Total flights 2005-2035 • Daily flight distribution • Connectivity • Fleet age • STAPES Lden • Full-flight CO2 emissions • Full-flight NOx emissions • Combined indicators • STAPES Lnight • Full-flight HC, CO, PM emissions • Air traffic summary • Noise summary • IMPACT emissions summary • Certified aircraft noise level • Certified helicopter noise levels • Certified engine NOx emissions • Average NOx margin to CAEP6 limit

Table 4-19: Aviation Data Sources Profiling

Based on the information that is publicly available, most of the aviation data sources are structured, provide historical data with different frequency but always covering Europe in terms of spatial coverage, as evidenced in Table 4-20. The data are available in different formats ranging from typical csv and xml to sdmx and aixm. With the exception of ADS_04, the relevant data sources are processed and provisioned at batch and near-real time levels. None of the sources is dependent on other data sources or already linked to relevant data sources.

[8] https://www.easa.europa.eu/eaer/downloads

General Info	Data Assets Features												
ID	Volume	Variety	Type	Format	Velocity	Historical Data Availability	Historical Data Frequency	Temporal Coverage	Spatial Coverage	Language	Relevant Standards	Veracity	Dependency / Linking to Other Sources
ADS_01	N.A.	Structured	Text	csv, sdmx, pdf, spss, excel, tsv, etc.	Batch	Y	Yearly	From 2007 to 2016	Europe	English	N.A.	Processed	N
ADS_02	N.A.	Semi-structured	Text	AIXM, ARINC424, xml, etc.	Batch, Real-time	Y	Y	N.A.	Europe	English	AICM / AIXM, SARPS, OPADD	Pre-processed, Processed	N
ADS_03	N.A.	Structured	Text	N.A.	Near real-time (refreshed every 10 minutes)	Y	Daily	N.A.	Europe	English	N.A.	Processed	N
ADS_04	571 records in total	Structured	Text	csv	Batch	N (only record of changes)	-	N.A.	Worldwide	English	N.A.	Raw	N
ADS_05	855KB in total	Structured	Text	csv	Batch	Y	Every few years	Sporadic from 2005-2014	Europe	English	N.A.	Processed	N

Table 4-20: Aviation Data Sources Profiling – Data Assets Features

The aviation data assets described in Table 4-21 are typically collected by multiple aviation stakeholders, pre-processed and aggregated by EUROCONTROL and EASA. In their large majority, they are accessible as downloadable files although there are certain cases where the data are available via web services and APIs (e.g. ADS_02 and ADS_01, respectively) or through a web browser to directly navigate to the data. Such data assets are usually public and free without evidence of any applicable licenses with the exception of ADS_02 and ADS_03 that require specific agreements and have either service charges and royalty frees (ADS_02) or contribution fees (ADS_03).

General Info	Data Assets Availability				Data Assets Rights			
ID	Data Asset Owned	Data Asset Available from 3rd Party	Data Asset Provider	Accessibility	Privacy	License	Pricing	Need for Anonymization
ADS_01	N	Y	Eurostat	Downloadable files, Via web browser, SDMX Web Services	Public, Open Data	N.A.	Free	N
ADS_02	N	Y	EUROCONTROL (certified under the Single European Sky legislation)	EAD BASIC access: Via web browser EAD Pro access: APIs, Software package for local installation, Downloadable files	Confidential	EAD Agreement for Data Users	Service charges and royalty fees	N
ADS_03	N	Y	EUROCONTROL	Via web browser	Public	NM Agreement for NOP Protected	Free / contribution fees	N
ADS_04	N	Y	ICAO, hosted by EASA	Downloadable file	Public, Open Data	N.A.	Free	N
ADS_05	N	Y	EASA	Downloadable file	Public, Open Data	N.A.	Free	N

Table 4-21: Aviation Data Sources Profiling – Data Assets Availability & Rights

An additional aviation data source that shall be further investigated at a later stage of the ICARUS project includes the ADSBexchange[9] which is a community-driven approach bringing together ADS-B/Mode S/MLAT feeders from around the world, to provide one of the world's largest sources of unfiltered flight data. What is notable is the difference of ADS-B Exchange from typical flight tracking sites as all data sent in from the community are, in turn, made available back to the community through various archives and APIs.

[9] http://www.ADSBexchange.com

4.3.2 Open Data Repositories & Linked Open Data Cloud

Over the past decade, the open data ecosystem not only reached its peak of expectations, but can be now pragmatically placed on its plateau of productivity. Despite the criticism that open data received over the years, numerous active and growing open data initiatives actually reinforce the European economy, with a plethora of data sources ranging from open governmental data to statistics and from environmental and transport data to health and well-being data.

An extensive search was performed on March 20th, 2018 to identify open data sources that are somehow related to: (a) the aviation data value chain, and (b) the initial data needs that were expressed by the ICARUS demonstrators in section 4.2. As in section 4.3.1, the search criteria that were used include: (a) appearance of one of the keywords "aviation data", "aerospace data", "air traffic data", "environment", "fuel emissions", "health", "influenza" and (b) focus on the pan-European coverage (instead of country-specific data sources or worldwide data sources).

Table 4-22 presents the results of the initial search on popular open data repositories and open data sources that were considered as highly relevant. It needs to be highlighted that the fragmentation of the aviation data value chain and the lack of data sharing was once again confirmed by the findings of this search activity. The number of aviation-related datasets in comparison to environment or health is practically negligible while the data assets that are relevant to the demonstrators' needs are restricted to open weather data and open air quality data.

With the proliferation of linked data technologies, the Linked Open Data (LOD) Cloud[10] was also considered as an additional data source, yet the search performed did not yield any results relevant to ICARUS at the moment (e.g. the airport data from http://airports.dataincubator.org/ that would be indeed relevant are not maintained!).

ID	Open Data Repository	Total Number of Datasets / Volume	Number of Related Datasets	Brief Profile of Relevant Indicative Datasets
ODR_01	EU Open Data Portal[11]	12,229	Aviation: >147 Environment: 1,743 Health: 735	• Reports for SESAR Solutions - Multi-Sector Planning; SESAR Solutions - User Preferred Routing; SESAR Solutions - Sector Team Operations - En-route Air Traffic Organizer. Provider: SESAR. Format: pdf
ODR_02	European Data Portal[12]	821,503	Aviation: 21 Environment: 3,828 Health: 288	• No datasets very relevant to ICARUS at the moment
ODR_03	Eurostat[13]	>4,600	Aviation: 399 Environment: 2,329 Health: 3,143	• See data asset ADS_01 in section 4.3.1
ODR_04	OECD Data[14]	672	Aviation: - Environment: 110 Health: 49	• No datasets very relevant to ICARUS at the moment
ODR_05	World Bank Data[15]	21,434	Aviation: 38 Environment: 1,957 Health: 3,048	• Air transport, passengers carried. Formats: csv, excel, tabbed txt. Historical data available: Y. Privacy: Public. License: CC BY-

[10] http://lod-cloud.net/
[11] https://data.europa.eu/euodp/data/
[12] https://www.europeandataportal.eu/
[13] http://ec.europa.eu/eurostat/web/transport/data/database
[14] https://data.oecd.org/
[15] https://data.worldbank.org/

ID	Open Data Repository	Total Number of Datasets / Volume	Number of Related Datasets	Brief Profile of Relevant Indicative Datasets
				4.0. • Population Estimates and Projections from 1960 to 2050 for 217 economies. Historical data available: Y. Privacy: Public. License: CC-BY 4.0.
ODR_06	World Health Organization[16]	Health-related statistics for >1000 indicators for 194 Member States	Health-related only	• FluNet virological data for tracking the movement of viruses globally and interpreting the epidemiological data. Formats: csv, xml, pdf, mhtml, excel. Historical data available: Y. Temporal Coverage: 1996-Today. Spatial Coverage: Worldwide. Relevant Standards: SDMX-HD. Privacy: Public.
ODR_07	OpenWeatherMap[17]	N.A.	N.A.	• Current weather API for weather data for any location including over 200,000 cities, frequently updated based on global models and data from more than 40,000 weather stations. Velocity: Real-time. Format: json, xml, html. Accessibility: API. Relevant Standards: ISO 3166 country codes. Privacy: Public. License: Free/Paid • Historical Data API, providing city historical weather data for 37,000+ cities. Velocity: Batch. Historical Data Availability: Y. Temporal Coverage: From 5 years previous to 1-month previous depending on the type of account. Privacy: Proprietary. • History Bulk. Velocity: Batch. Accessibility: API. Spatial Coverage: Over 37,000+ cities. Temporal Coverage: 5 years. Pricing: $10 per city. • Weather map layers, incl. precipitation, clouds, pressure, temperature, wind. Velocity: Real-time. Accessibility: as layers in Direct Tiles, OpenLayers, Leaflet, and Google Maps. Privacy: Public. Licensing: Mixed. • Air Pollution (Beta), with main indexes of CO, O3, NO2 and SO2. Velocity: Real-time, batch. Accessibility: API. Temporal Coverage: 11/2015-today. Privacy: Public. Licensing: Mixed.
ODR_08	Copernicus[18]	N.A. – Collecting ~8 PBs/year	N.A.	• Atmosphere Monitoring Service, i.e. Air quality and atmospheric composition, Emissions and surface fluxes. Formats: NetCDF, GRIB2. Historical Data Available: Y. Veracity: Pre-processed & Processed. Accessibility: Downloadable files, API. Privacy: Public.
ODR_09	OpenAQ (Air Quality Data)[19]	196,092,604 air quality	Aggregations of PM2.5, PM10,	Variety: Structured, Text. Format: csv, json. Velocity: Real-time, Batch. Historical Data

[16] http://apps.who.int/gho/data/node.home
[17] https://openweathermap.org/
[18] http://www.copernicus.eu/main/overview

ID	Open Data Repository	Total Number of Datasets / Volume	Number of Related Datasets	Brief Profile of Relevant Indicative Datasets
		measurements from 8,440 locations in 65 countries	ozone (O3), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and black carbon (BC) data	Availability: Y. Historical Data Frequency: Daily. Temporal Coverage: 2016-2018. Spatial Coverage: Worldwide. Language: English. Veracity: Pre-processed. Accessibility: API (with certain time limits for API calls), Downloadable files, PostgeSQL Database snapshot. Privacy: Public. License: Creative Commons Attribution 4.0 Generic License. Pricing: Free. Need for Anonymization: N

Table 4-22: Open Data Repositories and Sources

4.3.3 Other Relevant Data Sources

Additional initiatives that may provide data sources of interest to ICARUS include EC funded projects under the Horizon 2020 programme, particularly in the BDV-PPP (Big Data Value Public-Private Partnership) and the Transport Work Programme. As many of the projects have only started its activities, the synergies around common data assets of interest (e.g. with BigMedilytics on health data) shall be investigated in due time (e.g. in face-to-face meetings during the BDVA events). The following table thus focuses on H2020 projects that started in 2017 and whose results are not only clearly described online, but are also aligned with the initial data needs in ICARUS as elaborated in section 4.2.

ID	EC Project	Description	Relation to ICARUS
RDS_01	TransformingTransport ("Big Data Value in Mobility and Logistics") Open Data Portal[20] - H2020-ICT-15-2016	To provide the community working on transport data across the different transport domains identified for TT (Smart Highways, Sustainable Connected Vehicles, Rail, Ports, Airports, Integrated Urban Mobility and Dynamic Supply Networks) with open datasets that they can reuse for their own purposes, as well as links and metadata to existing datasets that cannot be published under an open data license, but where ad-hoc agreements may be established between the data producers (identified as part of such metadata) and the potential data re-users.	• 1 data asset "Weather forecasts" under the airports group described in detail (e.g. Volume: 1 GB, Velocity: Daily Batch, Historical Data Availability: N, Temporal Coverage: 2017-Today, Spatial Coverage: Worldwide, Relevant Standards: GRIB2) *Important note*: As AIA also participates to TT, potential data collaborations among ICARUS and TT are to be further investigated.
RDS_02	SafeClouds ("Data-driven research addressing aviation safety intelligence")[21] – H2020-MG-3.1-2016	SafeClouds proposes a big data-driven approach to achieve a deeper understanding of the dynamics of the system, where risks are pro-actively identified and mitigated. This will be achieved through a user-requirements driven approach for data mining in aviation safety, as well as novel data structures and safety knowledge representation. Two concrete	*Important note*: No details are publicly provided for data assets that are to be provisionally shared between the aviation stakeholders. Further investigation is needed for potential data synergies among ICARUS and SafeClouds.

[19] https://openaq.org
[20] http://data.transformingtransport.eu/dataset
[21] http://innaxis.org/safeclouds

ID	EC Project	Description	Relation to ICARUS
		implementation scenarios will be considered: airline operations perspective (SafeOps) and runway safety perspective (SafeRunway).	
RDS_03	EW-Shopp ("Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey")[22] - H2020-ICT-14-2016	EW-Shopp aims at supporting companies operating in the fragmented European ecosystem of the eCommerce, Retail and Marketing industries to increase their efficiency and competitiveness by leveraging deep customer insights that are too challenging for them to obtain today.	• Retail-related data assets ranging from Consumer Data (Purchase Intent and History) to Market Data (Sales) as identified in the EW-Shopp Data Management Plan (D2.1).

Table 4-23: Other Relevant Data Sources

4.4 ICARUS Data Preliminary Assessment

4.4.1 Data Positioning to ICARUS Data Tiers

With the data assets that have been identified in the previous sections (4.1-4.3) only initiating the data collection activities, ICARUS shall provide tangible access to one of the largest gold mines of data as aviation is often characterized with data generated by the ICARUS data value chain being classified into three core tiers:

- Data Tier 1: Primary Aviation Data consists of aircraft sensor data, scheduled route plans, airport traffic, fuel emissions, passenger data that pile up in heaps of data in every flight. Typical data providers include airports, airlines, and aircraft manufacturers.
- Data Tier 2: Extra-Aviation Data features data collected by airport services providers, and aviation-related service providers. Such data concern passengers' profiles which are complemented by Linked Open Data (indicatively weather, environment) and other historical data.
- Data Tier 3: Aviation-derived & Aviation-combined Data contains data and knowledge from other sectors (like Health, Tourism, Public Sector) that can be combined with aviation data from tiers 1 and 2 to produce new derived data and create new knowledge that would be impossible to deduce otherwise.

Figure 4-1 depicts the positioning of the different data assets that either belong to the ICARUS consortium (demonstrators and OAG) or are available from 3rd party sources, on the ICARUS data value chain. As it can be easily noticed, the data that are already available to the consortium enlist 25 data assets that mostly belong to the Data Tier 1, while the data that can be collected from other open data sources include 16 data sources/assets that are distributed along the 3 data tiers.

[22] http://data.transformingtransport.eu/dataset

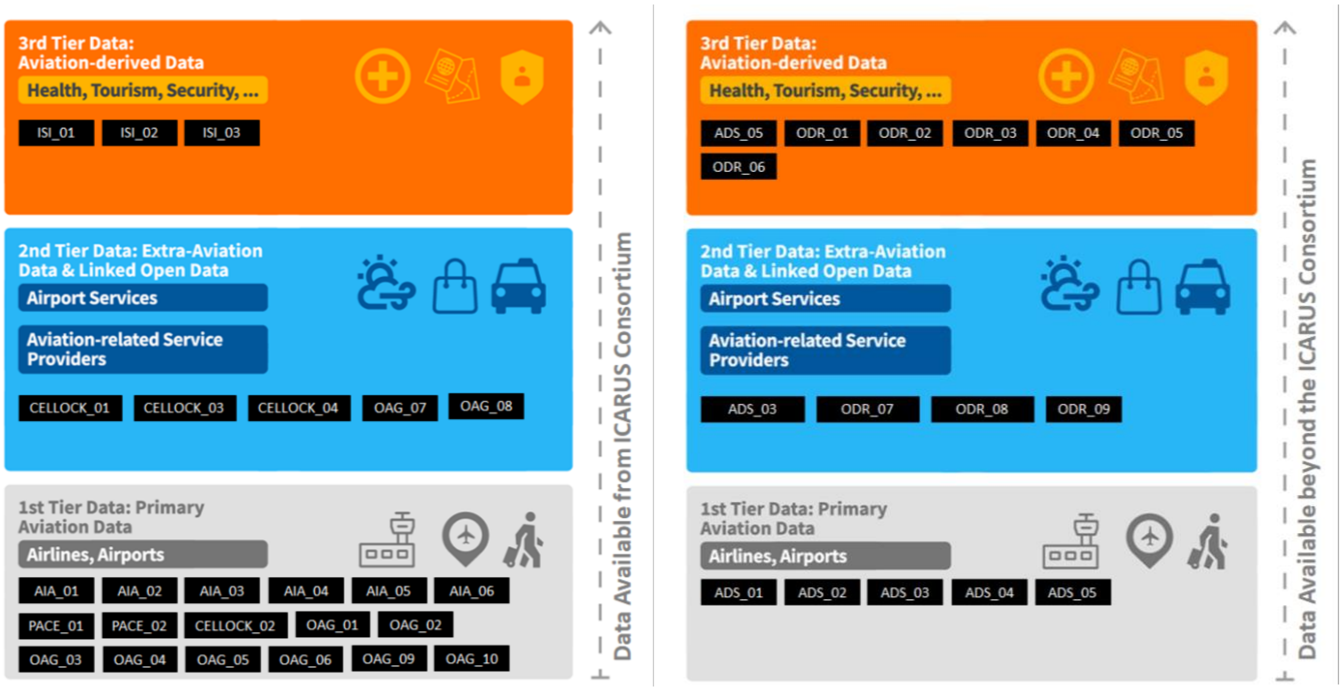


Figure 4-1: ICARUS Data Assets Available classified in the tiers of the aviation data value chain

With regard to the ICARUS demonstrators needs, Figure 4-2 visualizes the preliminary needs for different data assets that the demonstrators have identified at the beginning of the project.

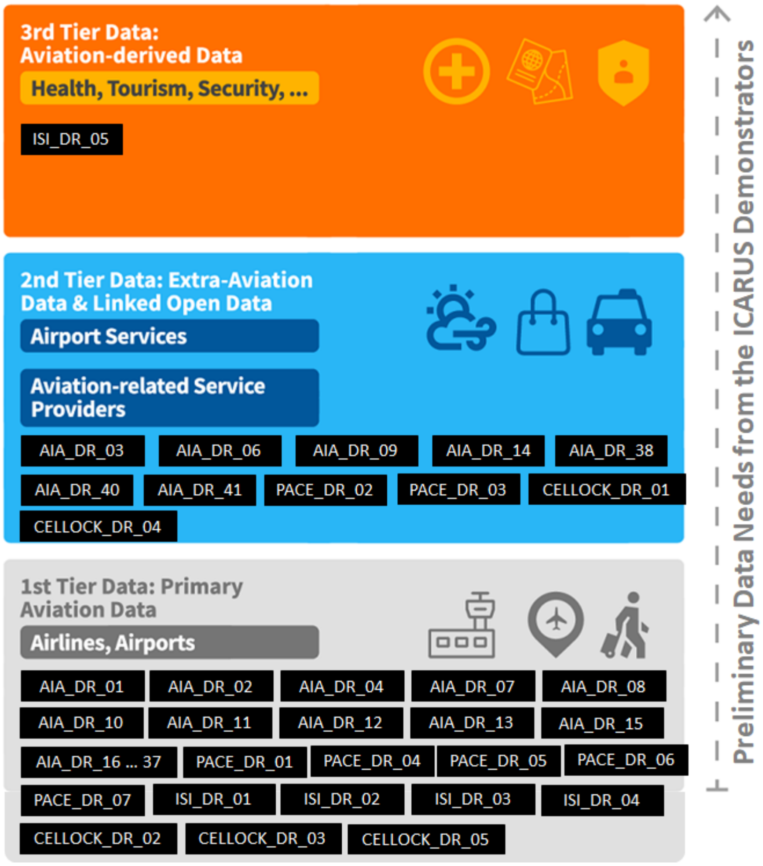


Figure 4-2: ICARUS Data Assets needed as classified in the tiers of the aviation data value chain

As it can be easily noticed, the majority of the data assets that are needed by the demonstrators fall within the 1st and 2nd tier of the value chain. They typically originate from airlines or service providers like ground handling companies and duty-free companies, which ICARUS shall strive to bring on board to share their data assets in ICARUS in a win-win manner.

4.4.2 ICARUS Demonstrators Data High-level Evaluation

In order to gain some high-level insights on the data quality of the data that the demonstrators currently have at their disposal (as documented in sections 4.1.1, 4.1.2, 4.1.3, 4.1.4), a high-level evaluation was performed by the demonstrators as presented in Table 4-24. The criteria that were assessed include:

- Accuracy as a measure of correctness and precision, e.g. whether the dataset is error-free, ranked from 1 (Low) to 5 (High).
- Completeness as the degree to which a data asset is sufficient in scope and depth, also ranked from 1 (Low) to 5 (High).
- Timeliness defining for how long a data asset remains up-to-date.

According to the demonstrators, the data assets they have at their disposal and shall make available through ICARUS are broadly of high quality and completeness while they are expected to remain useful for some years to indefinitely.

ID	Data Asset Title	Accuracy	Completeness	Timeliness
AIA_01	Time stamps and status of ground handling processes	5	5	Indefinitely
AIA_02	Checked passengers per flight	5	5	Indefinitely
AIA_03	Expected passengers per flight	5	5	Indefinitely
AIA_04	Connecting passengers per flight	5	5	Indefinitely
AIA_05	Passengers who need assistance per flight	5	5	Indefinitely
AIA_06	Gate open time	5	5	Indefinitely
PACE_01	Alternative routes comparison	5	5	Indefinitely
PACE_02	AC performance data	5	5	Indefinitely
ISI_01	Population data	4	4	A few years
ISI_02	GLEAM Simulation output	N.A.	4	N.A.
ISI_03	Virus & infections data	4	4	A few years
CELLOCK_01	Retail and F&B in-flight sales	1	1	N.A.
CELLOCK_02	Number of Passengers	1	1	N.A.
CELLOCK_03	In-flight, IFE Passenger data	3	2	N.A.
CELLOCK_04	IFE Content data	3	2	N.A.

Table 4-24: ICARUS Demonstrators Data Assets Available Assessment

4.4.3 Initial Data Needs Prioritization in ICARUS

In order to prioritize the preliminary data needs expressed by the demonstrators, an initial ranking was performed by the demonstrators regarding the relevance of a data asset for the demonstrator and the importance (criticality) for the implementation of the draft scenarios that the demonstrator partners have identified. Both criteria have been assessed from 1 (Low) to 5 (High).

As depicted in Figure 4-3, there are significant variations on the way the different demonstrators have assessed the preliminary data sources they have identified, a fact that can be – to an extent - attributed to their different mentalities and perspectives, yet deserves further discussion with the demonstrators. A number of data assets (17, in total) is considered as of high importance and of high relevance by the AIA and PACE/TXT demonstrators, with 20 more data assets as highlighted by AIA and ISI following behind. A number of data assets by AIA and CELLOCK (14 in total) are considered of medium importance and relevance. Finally, 1 data asset is considered as very relevant by ISI but of medium criticality (ISI_DR_04) while 7 data assets are considered of medium relevance and relatively low importance by CELLOCK, ISI and PACE/TXT.

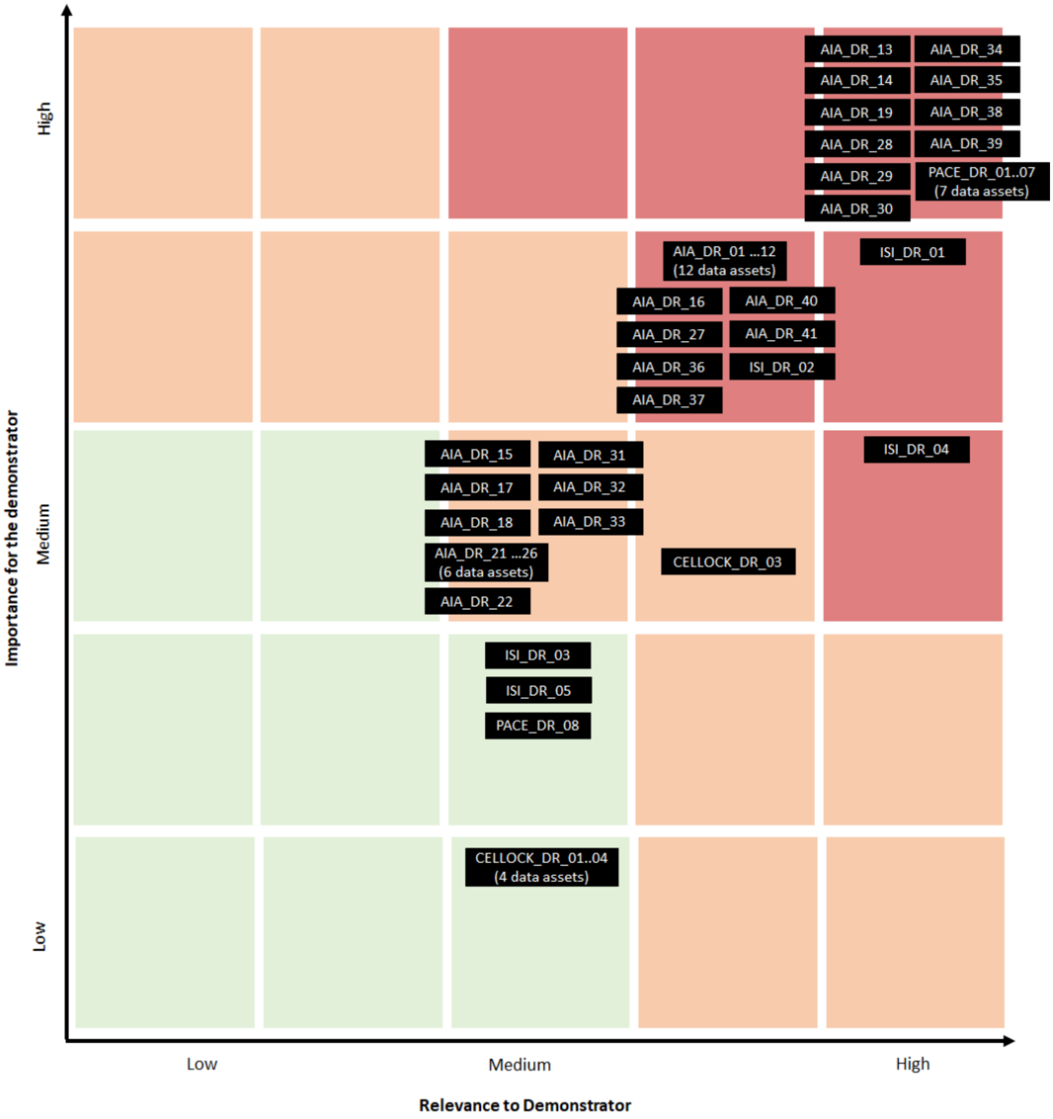


Figure 4-3: Initial assessment of the ICARUS Data Assets needed

In addition, the data assets of each demonstrator as documented in sections 4.1.1-4.1.4 were studied and cross-checked by the rest of the demonstrators in order to identify common data needs and requirements and eventually facilitate the prioritization of the data assets. As noted in Figure 4-4, a number of data assets needed from airlines and identified by AIA (e.g. AIA_DR_21, 22, 26) are also very relevant for PACE. Two data assets defined by AIA are also relevant to different degrees by ISI (AIA_DR_35 characterized as low relevance and AIA_DR_41 as high relevance). A data asset proposed by PACE (PACE_DR_02) is of low relevance to ISI. PACE and CELLOCK have common need for 2 data assets, namely CELLOCK_DR_03 and CELLOCK_DR_05) while ISI would be also highly interested in CELLOCK_DR_02 put forward as a data asset needed by CELLOCK. Finally, it is worth underlying that similar needs for data assets regarding passenger demographics (CELLOCK_DR_02 and AIA_DR_35) and airport retail data (CELLOCK_DR_04 and AIA_DR_38) have been identified by CELLOCK and AIA.

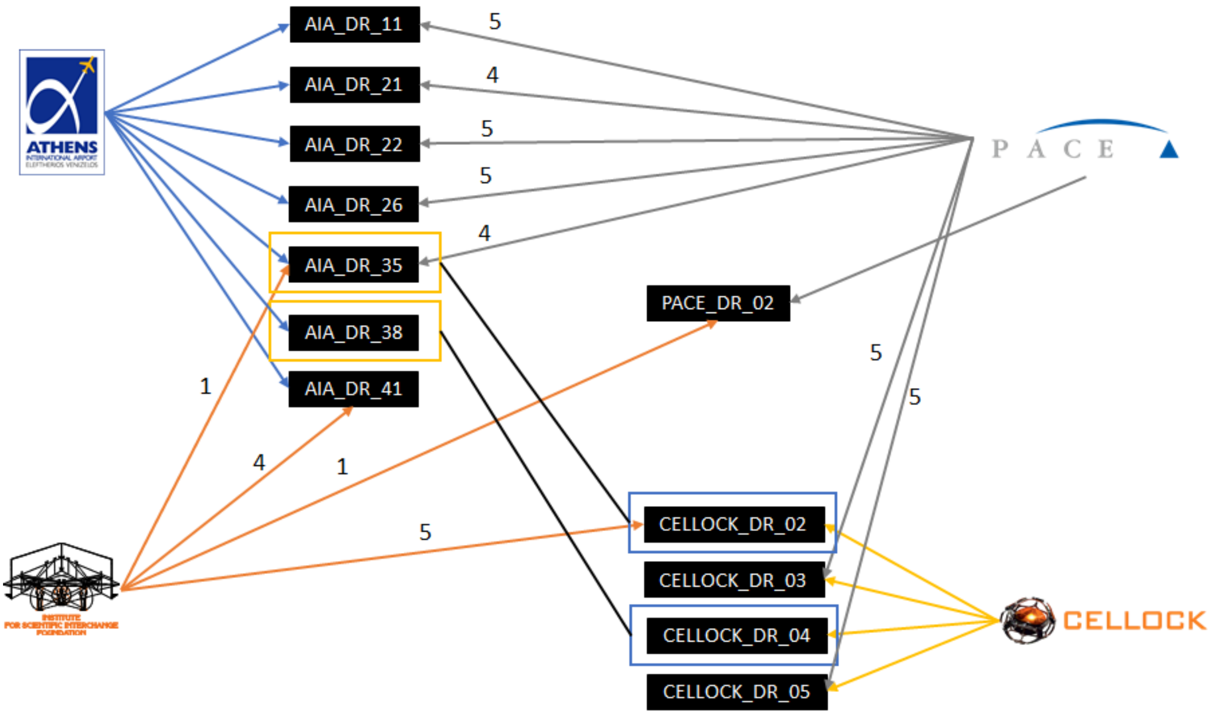


Figure 4-4: Interrelations among the ICARUS demonstrators regarding the data assets needed (note: the numbers on the arrow signify relevance for other demonstrators, the boxes on data assets depict similarity)

4.5 Aviation Data Protection and Sharing

4.5.1 Data IPR and Licensing State-of-Play

In a competitive industry like aviation, data are typically considered as an asset to be safeguarded rather than a commodity to be shared. Lack of trust, uncertainty about data ownership and data access and fear of competition are listed as the core business, legal and cultural concerns that hinder adoption of a data sharing mentality[23]. In this context, the importance of safeguarding the data IPR through appropriate licensing that forges a balance inside the traditional “all rights reserved” setting that the copyright law inherently creates becomes more and more instrumental.

In order to grant permissions over the data use to potential consumers, licenses and waivers can generally be put into use in the following ways:

- A license is a legal instrument for the data provider to authorize and permit a data consumer to utilize the data in a manner that would otherwise infringe on the rights held. As only the data provider as the rights holder (or someone with a proven right or license to act on their behalf) can grant a license, it is imperative that the intellectual property rights (IPR) pertaining to the data are appropriately established before any licensing takes place.

[23] http://www.datalandscape.eu/data-driven-stories/what-limits-data-sharing-europe

- A waiver is a legal instrument for data providers to give up their rights, so that infringement becomes a non-issue. Again, only the entity that holds the content rights (or someone with a proven right or license to act on their behalf) can waive them.

From a data perspective, data licenses effectively expand or restrict what a data consumer is allowed to do with the data and grant permissions based on the idea that certain terms have to be met. Although the precise details vary, three conditions regarding attribution, copyleft, and non-commerciality are commonly found in licenses: An attribution requirement means that the data provider must be given due credit for the data asset when it is distributed, shared, visualized, or analyzed to derive a new data asset; A copyleft requirement means that any new results derived from the licensed data must be released under the same license, and only that license; The intent of a non-commercial license is to prevent the consumer from exploiting the data-driven results commercially, yet a dual-licensing regime often applies in such cases with alternative licenses envisaged to allow commercial uses upon payment to the data provider.

The data licenses are often classified along three categories:

- *Prepared licenses* when the legal department of an organization has already drafted a license "template" that is applicable to any data exchange with minor adaptations to the terms according to the certain circumstances of the data and / or the data consumer.
- *Bespoke licenses* which are prepared in a custom, bilateral way by the legal departments of the organizations involved when there is a significant commercial value associated with the data or the providers need to elaborate on their responsibilities and the responsibilities of the consumers with respect to the data reuse.
- *Standard licenses* including the most commonly used licenses, as prepared by international organizations accredited for addressing IPR, as depicted in the following table.

In general, datasets are particularly prone to attribution stacking, where a derivative work must acknowledge all contributors to each work from which it is derived, no matter how distantly. If a dataset is at the end of a long chain of derivations, the list of credits might become too lengthy and unmanageable, while such an issue amplifies if different sets of contributors have to be credited in a different way. In addition, the selection of a copyleft license hinders the integration of the licensed data with other data released under a different copyleft license as the derived dataset is not able to satisfy different license terms at the same time, even if they are relatively compatible [175]. Non-commercial licenses may also have broader repercussions due to the ambiguity of what constitutes a commercial use: compensation for the derivative results (data derived or intelligence reports) or sales of the derivative results even if there is no financial benefit for the data consumer.

ID	License	Description	Types	Features[24]	Considerations for ICARUS
DL_01	Creative Commons[25]	The Creative Commons licenses give the creators of creative works (ranging from music, images and video, to data) finer-grained control over how they may be used than simply declaring them public domain or reserving all rights. Apart from the legal text, the licenses provide clear concise summaries and a canonical URL for use in HTML, RDF and other code.	Attribution (CC BY)	BY: Yes - SA: No	**+** Appropriate for very simple, factual datasets **-** Only version 4 applicable for data (since it explicitly includes sui generis database rights that are in force in the European Union unless the licensor specifically reserves them) **-** Datasets and databases considered as a whole, creating difficulty in certain complex cases such as collections of variously copyrighted works **-** No distinction use of data as part of a new collection/database from use of data to generate intelligence and visualize results **-** Attribution stacking **-** Attention to: (a) the NC condition to be only used with dual licensing, (b) the SA condition as it reduces interoperability, (c) the ND condition as it severely restricts reuse
			Attribution Share Alike (CC BY-SA)	BY: Yes - SA: Yes	
			Attribution No Derivatives (CC BY-ND);	BY: Yes - SA: No	
			Attribution Non-Commercial (CC BY-NC)	BY: Yes - SA: No	
			Attribution Non-Commercial Share Alike (CC BY-NC-SA)	BY: Yes - SA: Yes	
			Attribution Non-Commercial No Derivatives (CC BY-NC-ND)	BY: Yes - SA: No	
			Creative Commons Zero (CC0) dedicating works to the public domain	BY: No - SA: No - All rights waived	**+** Data to be used by anyone or for any purpose **+** Simplification of integration with other data **-** Lack of control over how data are reused **-** Lack of protection against unfair competition
			Creative Commons Public Domain Mark (CC PDM)	BY: No - SA: No - To assert that a work is already in the public domain	
DL_02	Community Data License Agreement (CDLA)[26]	The CDLA license agreements enable sharing data openly, embodying best practices learnt over decades of sharing source code in the Linux Foundation.	CDLA-Sharing 1.0 (encouraging contributions of data to the community)	BY: Yes - SA: Yes	+ Appropriate for datasets as a whole as well as their individual contents + Distinction between data and “results” obtained by processing or analyzing that data - Not explicit in the CDLA-Sharing whether the data are royalty-free - Agnostic with regard to data privacy
			CDLA-Permissive 1.0 (not requiring any additional sharing of data)	BY: Yes - SA: No	

[24] BY = Requires Attribution
SA = Require Share-Alike

[25] https://creativecommons.org/

[26] https://cdla.io/

ID	License	Description	Types	Features[24]	Considerations for ICARUS
DL_03	Open Data Commons[27]	Maintained by the Open Knowledge Foundation, the Open Data Commons licenses present striking similarities to the Creative Commons licenses, yet they are designed specifically for databases.	Open Data Commons Attribution License (ODC-BY)	BY: Yes - SA: No	**+** Appropriate for databases and for generating non-data products **+** Distinction between licensing of the database and licensing of the data **-** Attribution stacking **-** Attention to: (a) the copyleft condition as it reduces interoperability, and (b) the Digital Rights Management (DRM) clause as it may put off some data consumers.
			Open Data Commons Open Database License (ODC-ODbL)	BY: Yes - SA: Yes	
			Open Data Commons Public Domain Dedication and License (PDDL)	BY: No - SA: No - All rights waived	**+** Data to be used by anyone or for any purpose **+** Simplification of integration with other data **-** Lack of control over how data are reused **-** Lack of protection against unfair competition
			Open Data Commons Database Contents License (ODC-DbCL)	BY: No - SA: No - Copyright waiver for the contents of the database without affecting the rights of database itself	

Table 4-25: High-level Comparison of Standard Data Licenses

[27] http://opendatacommons.org/

As discussed in Table 4-25, the CDLA license appears as more appropriate for big data and generally large-scale streaming datasets that are constantly changing, to support machine learning or artificial intelligence systems, due to the clarity it provides to data providers and consumers regarding their ability to curate, use, and share data (in a similar way to how open source software is developed).

In order to translate such licenses into a standard format that can be exchanged between information systems and queried, a Rights Expression Language (REL) needs to be put into use. Typical RELs today include:

- Creative Commons Rights Expression Language (CC REL)[28], the standard recommended by Creative Commons (CC) for machine-readable expression of copyright licensing terms and related information. CC REL metadata, as encoded using RDFa or XMP, may be embedded in a variety of filetypes. Although CC REL is specified in an abstract syntax-free way, as an extensible set of properties to be associated with a licensed document, it does not seem to support the expression of the fine-grained constraints and actions usually required by data licenses.
- Open Digital Rights Language (ODRL)[29], providing flexible and interoperable mechanisms to support transparent and innovative use of digital content in publishing, distribution, and consumption of digital media. The ODRL specification features the core model, the XML and the JSON encodings, the ontology, the information model, and the vocabulary.

Despite the availability of such RELs, though, none has gained widespread adoption and implementation so far.

Lately, with the rise of open source blockchain implementations like Hyperledger Fabric and Etherium, as well as of blockchain-based data marketplaces (e.g. datum, Repux, IOTA Data Marketplace)[30], it becomes more and more crucial to be able to rapidly reach data agreements with the same data assets being associated with multiple data contracts. Over the last years, a number of interesting approaches regarding data contracts that follow a similar mentality to service contracts have emerged. Truong et al (2012) [176] provide an analysis of current data contracts in order to identify relevant data contract properties and methods for data-as-a-service and propose an abstract data contract model for developing data contracts in order to facilitate the right selection and utilization of data assets in data marketplaces. The data contract terms upon which their proposed model is built includes: (a) Data rights in terms of Derivation, Collection, Reproduction, Attribution, Noncommercial use, (b) Quality of data in terms of Accuracy, Completeness, Up-to-dateness, (c) Compliance (e.g. for privacy), (d) Pricing model (explicitly defining the cost, the time and the number of transactions), and (e) Control and responsibility. Cao et al (2016) [177] present techniques for managing data contracts based on different cost models, quality of data, and data rights while supporting Obligation-free contracts (a type of data contract which does not require involving parties to have any obligation to conform to terms and conditions specified in the contract); User-centric contracts (focusing on requirements of a service that the service provider has to deliver to users); Provider-centric contracts (with requirements on data rights and regulatory compliance of the data that users have to follow) and Customizable contracts (that allow users to modify any of the above contract models). Vu et al (2012) [178] also introduce a general linked model to cover all basic information of data-as-a-service, as well as integrating existing work in describing quality of data, data and service contracts, data dependency, and Quality of Service (QoS). The model for describing pricing

[28] https://wiki.creativecommons.org/wiki/CC_REL
[29] https://www.w3.org/community/odrl/
[30] Note: the state-of-play on data brokerage will be extensively analyzed in WP2.

information on data-as-a-service that has been thus designed covers different payment plans, including payment on access (API call), payment on resource consumption, payment on data type and data size, and payment on plan (fixed payment in a period). Finally, in the AEGIS H2020 project, different concerns and metadata concur to define the AEGIS Data Policy Framework[31], namely: (a) Data Assets Rights (DAR) (including Permissions, Requirements and Prohibitions), (b) Quality of Data Assets (QoDA) encapsulating Accuracy, Completeness, Consistency, Credibility, and Timeliness, (c) Pricing Model consisting of: Price Scheme, Cost, Coverage, Exclusivity of use, Duration of use, Duration of offline retention, and Maximum Use, (d) Policy Terms in terms of Liability, Privacy Compliance, Online Availability Guarantees, Versioning & updates, Applicable Law.

In summary, such data licenses need to go beyond the typical copyright licenses and approaches that have been defined so far in order to effectively address the diverse and contextual constraints on data from different data providers and consumers in domain-specific value chain approaches like in ICARUS.

4.5.2 Legal and Regulatory Legislation applicable to the ICARUS Data Tiers

The data to be collected, shared and analyzed in ICARUS along the three data tiers need to abide to the pertaining legal and regulatory legislation that can be summarized on the following regulations and directives at EU level:

(A) European Civil Aviation Handbook: Part I. Regulations and Directives[32] that is an informative document, yet it contains the Regulations, Directives, Decisions, Case Law and International Agreements of the European aviation law. A selection of Regulations and Directives that are considered as most relevant to ICARUS taking into account the data assets documented in the previous sections concern the following aspects:

- **Airports**: Allocation of slots [Council Regulation (EEC) No 95/93]; Access to ground handling market [Council Directive 96/67/EC]
- **Air traffic management**: Framework for the creation of the Single European Sky (SES) [Regulation (EC) No 549/2004]; Provision of air navigation services in the SES [Regulation (EC) No 550/2004]; Organization and use of the airspace in the SES [Regulation (EC) No 551/2004]; Interoperability of the European ATM network [Regulation (EC) No 552/2004]; Requirements for the provision air navigation services [Commission Regulation (EC) No 2096/2005]; Rules for the flexible use of airspace [Commission Regulation (EC) No 2150/2005]; Air traffic controller licence [Directive 2006/23/EC]; Airspace classification and access of flights [Commission Regulation (EC) No 730/2006]; Automatic systems for the exchange of flight data [Commission Regulation (EC) No 1032/2006]; Procedures for flight plans in the preflight phase for the SES [Commission Regulation (EC) No 1033/2006]; Common charging scheme for air navigation services [Commission Regulation (EC) No 1794/2006]; Establishment of a joint undertaking to develop SESAR [Council Regulation (EC) No 219/2007]; Requirement for the application of a flight message transfer protocol [Commission Regulation (EC) No 633/2007]; Safety oversight in air traffic management [Commission Regulation (EC) No 1315/2007]

[31] AEGIS D2.1 Semantic Representations and Data Policy and Business Mediator Conventions. Available online at: http://www.aegis-bigdata.eu/
[32] https://ec.europa.eu/transport/modes/air/internal_market/handbook/part1_en

- **Air transport and market issue** including the Internal Market: Licensing of air carriers [Regulation (EC) No 1008/2008]; Access for Community air carriers to intra-Community routes [Regulation (EC) No 1008/2008]; Fares and rates for air services [Regulation (EC) No 1008/2008]; Insurance requirements for air carriers and aircraft operators [Regulation (EC) No 785/2004]; Code of conduct for computerized reservation systems [Regulation (EC) No 80/2009]; Statistical returns [Regulation (EC) No 437/2003]; Implementing rules of statistical returns regulation [Commission Regulation (EC) No 1358/2003]
- **Passenger rights**: Air carrier liability [Regulation (EC) No 889/2002]; Denied boarding cancellation or long delay of flight [Regulation (EC) No 261/2004]; Right of disabled persons [Regulation (EC) No 1107/2006]
- **Safety**: Collection and exchange of information on the safety of aircraft [Commission Regulation (EC) No 768/2006]; Harmonization [Regulation (EC) No 216/2008]; Investigation of civil aviation accidents and incidents [Council Directive 94/56/EC]; Common rules - EASA establishment [Regulation (EC) No 216/2008]; Occurrence reporting in civil aviation [Directive 2003/42/EC]; Rules for the airworthiness [Commission Regulation (EC) No 1702/2003]; Continuing airworthiness [Commission Regulation (EC) No 2042/2003]; Safety of third country aircraft using Community airport [Regulation (EC) No 216/2008]; Board of appeal of the EASA [Regulation (EC) No 216/2008]; Community list of air carrier subject to an operating ban [Regulation (EC) No 2111/2005]; Implementing rules for the banned air carrier list [Commission Regulation (EC) No 473/2006]; Fees and charges levied by the EASA [Commission Regulation (EC) No 593/2007]; EASA working methods for standardization [Commission Regulation (EC) No 736/2006]; Common rules - EASA establishment [Regulation (EC) No 216/2008]; List of banned air carrier [Commission Regulation (EC) No 1043/2007]
- **Security**: Common basic standards on aviation security [Commission Regulation (EC) No 820/2008]; National civil aviation security quality control programmes [Commission Regulation (EC) No 1217/2003]; Security restricted areas at airports [Commission Regulation (EC) No 1138/2004]; Procedures for conducting inspections in the civil aviation security [Commission Regulation (EC) No 1486/2003]; Civil aviation security [Regulation (EC) No 300/2008]
- **Environmental protection**. For Noise emission: Limitation of noise [Council Directive 89/629/EEC]; Operation of aeroplanes covered by Part II, Chap.3 Vol.1 of Annex 16 [Directive 2006/93/EC]; Introduction of noise-related restrictions [Directive 2002/30/EC]; Management of environmental noise [Directive 2002/49/EC]. For Gas emission: Greenhouse gas emission allowance trading within the Community [Directive 2008/101/EC]

(B) European General Data Protection Regulation (GDPR)[33] [Regulation (EU) 2016/679 — protection of natural persons with regard to the processing of personal data and the free movement of such data[34]] that repeals the "Data Protection Directive" 95/46/EC and was designed to harmonize data privacy laws across Europe, to protect all EU citizens from privacy and data breaches and to reshape the way organizations across the region

[33] https://www.eugdpr.org/

[34] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679

approach data privacy. GDPR is now applicable in any processing of personal data by controllers and processors in the EU, regardless of whether the processing takes place in the EU or not.

(C) EU Database Directive [Directive 96/9/EC on the 'legal protection of databases'[35]] which provides for two types of protection for databases: (i) databases can be protected, when original, under copyright law, and (ii) databases for which a substantial investment has been made can benefit from the "sui generis" protection. Owners of protected databases can prevent reproduction, communication, extraction or re-use of their database content on the basis of the protection granted by this directive. The directive also guarantees rights to the users, including the provision of specific exceptions in the fields of teaching, scientific research, public security or for private purposes. The EC launched a consultation to better understand how the Database Directive is used, to evaluate its impact on users and to identify possible needs of adjustment and performs an ongoing ex-post evaluation of the Directive[36].

Additional legislation that is under preparation at the moment includes: the proposal for Proposal for a Regulation on a framework for the **free flow of non-personal data in the European Union**[37], and the Proposal for a Regulation on **promoting fairness and transparency for business users of online intermediation services**[38].

4.5.3 Key Considerations for ICARUS Data Sharing

Upon obtaining an initial understanding of the data assets that ICARUS shall get involved within its 3-tiered data value chain and studying the state-of-the art concerning data protection and sharing, a number of key considerations have emerged and are posed as a set of intriguing, high-level questions for the project continuation:

- How to specify multiple yet thorough data licenses for the same data asset that protect the IP interests of the stakeholders in the aviation data value chain in an immutable manner and under trusted and fair conditions?
- How to ensure flexibility in data sharing for prosumers while defining detailed terms in clear data licenses?
- Which are the particular intentions, conditions and requirements for data sharing and brokerage in the broader aviation ecosystem?
- In light of the underlying bilateral agreements for sharing most of the demonstrators' data and the OAG data assets, what types of data licenses should be defined in ICARUS?
- How to track any possible infringement of a data license (i.e. any use outside the policy terms)?
- How to proactively resolve any IPR / data licences incompatibility issues that may hinder data integration and analytics?
- What is the appropriate balance between the terms that shall be eventually written in a blockchain and the metadata to be stored in the ICARUS platform?

[35] http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:31996L0009

[36] https://ec.europa.eu/digital-single-market/en/news/summary-report-public-consultation-legal-protection-databases

[37] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2017%3A495%3AFIN

[38] https://ec.europa.eu/digital-single-market/en/news/regulation-promoting-fairness-and-transparency-business-users-online-intermediation-services

- How to handle compensation / payments to external data providers during the project duration (with ICARUS only simulating a virtual currency approach during the project and putting in place monetization services after the end of the project)?
- How to ensure (near real-time) data collection, availability and analytics in light of the need for data replication from data marketplaces, open data portals and 3rd parties in accordance with the terms of the data licences that are in force?

5 Conclusion

This final section of this document summarizes the content presented in the Deliverable 1.1, which was the outcome of a research upon industrial data collected during the initial project activities. In the state-of-play analysis phase, modern state-of-the-art software (libraries / frameworks / platforms) and their main aspects were presented, as well as EU projects that are relevant to this project. In the requirements analysis phase, a logical process has been followed in order to identify the ICARUS stakeholders and target audience, understand the current state and derive the initial user needs. The steps of this process involved active contribution by all partners and the results of this analysis provide the pillars on which the technical and research work that will follow, will be based.

The first step of this process was to elaborate on the state-of-the-art key technology axes relevant to this project. In Section 2, a reference guide to the specific technologies that are embraced by the communities targeted by ICARUS are provided. More precisely, various technologies that belong to three main topics relevant to Big Data were presented. These topics are: Data Collection which involves Data Anonymization, Data Quality, Semantic Enrichment and Annotation; Data Processing and Management which refers to Data Curation, Data Linking, Data Storage and Query Processing; Data Analytics regarding Machine Learning, Deep Learning and Data Visualization. Additionally, various EU projects relevant to ICARUS were presented. However, this deliverable does not provide an in-depth analysis of the state-of-the-art research methods and methodologies, which will be described in WP2.

The next step was to identify the main stakeholders and the target audience. Section 3 of this document depicts the full image of the ones that the final result of ICARUS platform aims at. In addition, the current state of the industry was analyzed through the key findings of related industry studies, with the purpose of deriving the market gaps that this project will contribute to. Another contribution was the development and analysis of the preliminary stakeholders' requirements questionnaire. The questionnaire targeted potential users for ICARUS and the analysis of the responses produced results that were in accordance to all major industry surveys of the field.

The analysis of the responses contributed in highlighting the main obstacles and difficulties that stakeholders are currently facing. More specifically, the most difficult processes for organizations are the data anonymization and data linking, while their main concerns for data sharing are privacy/confidentiality and security. Moreover, organizations that use data marketplaces and APIs find it easier to collect data, while organizations that use custom in-house mechanisms find it harder. Additionally, almost 50% of the respondents do not have in place mechanisms for big data analysis, due to budget/cost constraints and lack of experience. What is more important though, is that the majority of the respondents are interested in a data marketplace platform that offers functionalities such as secure experimentation playground for experimenting with datasets before purchasing them, a service that recommends similar datasets with the ones currently explored and a dashboard with interactive visualization capabilities.

The final outcome of the document was the introduction of a set of data sources which will eventually feed the ICARUS data value chain (described in Section 4). In addition, an investigation on data IRP policies that may be integrated to the system was conducted, aiming to contribute to the design and implementation of a regulatory data sharing framework for data protection.

In the forthcoming steps, the outcomes of D1.1 will feed the Deliverable D1.2 in order to define the ICARUS methodology and value chain definition and formulate the platform’s MVP. The tasks of this Deliverable will be constantly monitored and will be reported in Deliverable D1.3 (“Updated ICARUS Methodology and MVP”), as they remain active until the 15th month of the project. The results of this deliverable will be used not only for the WP1, but also for other WPs: in WP2, to define the main data management, transformation, intelligence extraction and sharing methods that will be supported by the ICARUS platform; in WP3, to help the design of the architecture and of the core features of the ICARUS platform; in WP7, as input to the market analysis to be conducted.

6 References

[1] BDV SRIA, “European Big Data Value.” [Online]. Available: http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf. [Accessed: 29-Mar-2018].

[2] “Anonymization and the Future of Data Science.” [Online]. Available: https://www.kdnuggets.com/2017/04/anonymization-future-data-science.html. [Accessed: 07-Mar-2018].

[3] R. Matsunaga, I. Ricarte, T. Basso, and R. Moraes, “Towards an Ontology-Based Definition of Data Anonymization Policy for Cloud Computing and Big Data,” in *Dependable Systems and Networks Workshop (DSN-W), 2017 47th Annual IEEE/IFIP International Conference on*, 2017, pp. 75–82.

[4] P. Goswami and S. Madan, “Privacy preserving data publishing and data anonymization approaches: A review,” in *Computing, Communication and Automation (ICCCA), 2017 International Conference on*, 2017, pp. 139–142.

[5] L. Sweeney, “k-anonymity: A model for protecting privacy,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[6] C. Dwork, “Differential privacy: A survey of results,” in *International Conference on Theory and Applications of Models of Computation*, 2008, pp. 1–19.

[7] J. Shao and H. Ong, “Semantic attack on anonymised transactions,” in *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIII*, Springer, 2016, pp. 75–99.

[8] A. Chakravorty, C. Rong, K. R. Jayaram, and S. Tao, “Scalable, Efficient Anonymization with INCOGNITO-Framework & Algorithm,” in *Big Data (BigData Congress), 2017 IEEE International Congress on*, 2017, pp. 39–48.

[9] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, “COAT: COnstraint-based anonymization of transactions,” *Knowl. Inf. Syst.*, vol. 28, no. 2, pp. 251–282, 2011.

[10] S.-L. Wang, Z.-Z. Tsai, T.-P. Hong, I.-H. Ting, and Y.-C. Tsai, “Anonymizing Multiple K-anonymous Shortest Paths For Social Graphs,” in *Innovations in Bio-inspired Computing and Applications (IBICA), 2011 Second International Conference on*, 2011, pp. 195–198.

[11] J. Voisin, C. Guyeux, and J. M. Bahi, “The metadata anonymization toolkit,” *arXiv Prepr. arXiv1212.3648*, 2012.

[12] F. M. C. Dias, “Multilingual Automated Text Anonymization,” *Inst. Super. Técnico Lisboa*, 2016.

[13] J. Gardner and L. Xiong, “An integrated framework for de-identifying unstructured medical data,” *Data Knowl. Eng.*, vol. 68, no. 12, pp. 1441–1451, 2009.

[14] C. Cumby and R. Ghani, “Inference control to protect sensitive information in text documents,” in *ACM SIGKDD workshop on intelligence and security informatics*, 2010, p. 5.

[15] S. Ribaric and N. Pavesic, “An overview of face de-identification in still images and videos,” in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, 2015, vol. 4, pp. 1–6.

[16] “ARX.” [Online]. Available: http://arx.deidentifier.org. [Accessed: 07-Mar-2018].

[17] “UTD Anonymization Toolbox.” [Online]. Available: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/.

[Accessed: 07-Mar-2018].

[18] "sdcMicro." [Online]. Available: https://cran.r-project.org/web/packages/sdcMicro/index.html. [Accessed: 07-Mar-2018].

[19] "Partition_for_Transaction." [Online]. Available: https://github.com/qiyuangong/Partition_for_Transaction%0A. [Accessed: 07-Mar-2018].

[20] "Apriori_based_Anonymization." [Online]. Available: https://github.com/qiyuangong/Apriori_based_Anonymization. [Accessed: 07-Mar-2018].

[21] "GraphAnon." [Online]. Available: https://github.com/sean-chester/graphAnon. [Accessed: 07-Mar-2018].

[22] "Mat." [Online]. Available: https://mat.boum.org/. [Accessed: 07-Mar-2018].

[23] "NLM-scrubber." [Online]. Available: https://scrubber.nlm.nih.gov/. [Accessed: 07-Mar-2018].

[24] "MIST." [Online]. Available: http://mist-deid.sourceforge.net. [Accessed: 07-Mar-2018].

[25] "Anonymizer." [Online]. Available: http://www.eyedea.cz/image-data-anonymization/. [Accessed: 07-Mar-2018].

[26] "Facepixelizer." [Online]. Available: https://www.facepixelizer.com/. [Accessed: 07-Mar-2018].

[27] A. Maydanchik, *Data quality assessment*. Technics publications, 2007.

[28] W. Dai, I. Wardlaw, Y. Cui, K. Mehdi, Y. Li, and J. Long, "Data profiling technology of data governance regarding big data: Review and rethinking," in *Information Technology: New Generations*, Springer, 2016, pp. 439–450.

[29] "KDnuggets." [Online]. Available: https://www.kdnuggets.com/2017/05/must-know-common-data-quality-issues-big-data.html. [Accessed: 29-Mar-2018].

[30] "Griffin." [Online]. Available: http://griffin.incubator.apache.org/. [Accessed: 07-Mar-2018].

[31] "OpenRefine." [Online]. Available: http://openrefine.org/. [Accessed: 07-Mar-2018].

[32] "DataQuality." [Online]. Available: http://www.agilelab.it/data-quality-for-big-data/. [Accessed: 07-Mar-2018].

[33] "Open Studio for Data Quality." [Online]. Available: https://www.talend.com/products/talend-open-studio/data-quality-open-studio/. [Accessed: 07-Mar-2018].

[34] C. Diamantini and N. Boudjlida, "About semantic enrichment of strategic data models as part of enterprise models," in *International Conference on Business Process Management*, 2006, pp. 348–359.

[35] M. Clarke and P. Harley, "How smart is your content? Using semantic enrichment to improve your user experience and your bottom line," *Sci. Ed.*, vol. 37, no. 2, pp. 40–44, 2014.

[36] R. M. Keller, "Ontologies for aviation data management," in *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*, 2016, pp. 1–9.

[37] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *arXiv Prepr. arXiv1511.08308*, 2015.

[38] "SpaCy." [Online]. Available: https://spacy.io/. [Accessed: 07-Mar-2018].

[39] "GATE." [Online]. Available: https://gate.ac.uk/. [Accessed: 07-Mar-2018].

[40] "OpenNLP." [Online]. Available: http://opennlp.apache.org/index.html. [Accessed: 07-Mar-2018].

[41] "Stanford CoreNLP." [Online]. Available: https://stanfordnlp.github.io/CoreNLP/index.html. [Accessed: 07-Mar-2018].

[42] "Cogcomp-NER." [Online]. Available: https://github.com/CogComp/cogcomp-nlp/tree/master/ner. [Accessed: 07-Mar-2018].

[43] "OpeNER." [Online]. Available: http://www.opener-project.eu/. [Accessed: 29-Mar-2018].

[44] "DBpedia-spotlight." [Online]. Available: http://www.dbpedia-spotlight.org/. [Accessed: 07-Mar-2018].

[45] "Gerbil." [Online]. Available: http://www.aksw.org/Projects/GERBIL.html. [Accessed: 07-Mar-2018].

[46] M. L. Brodie and J. T. Liu, "The power and limits of relational technology in the age of information ecosystems," in *Keynote at On The Move Federated Conferences*, 2010.

[47] M. Pennock, "Digital Curation: A life-cycle approach to managing and preserving usable digital information," *Libr. Arch. January*, 2007.

[48] P. B. Heidorn, C. L. Palmer, M. H. Cragin, and L. C. Smith, "Data curation education and biological information specialists," 2007.

[49] "Datacleaner." [Online]. Available: https://datacleaner.github.io/. [Accessed: 07-Mar-2018].

[50] "Trifacta Wrangler." [Online]. Available: https://www.trifacta.com/products/wrangler/. [Accessed: 07-Mar-2018].

[51] "Talend." [Online]. Available: https://www.talend.com/. [Accessed: 07-Mar-2018].

[52] "WinPure." [Online]. Available: http://www.winpure.com/. [Accessed: 07-Mar-2018].

[53] "Factual/Drake." [Online]. Available: https://www.factual.com/blog/introducing-drake-a-kind-of-make-for-data. [Accessed: 07-Mar-2018].

[54] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001.

[55] P. Hitzler and K. Janowicz, "Linked Data, Big Data, and the 4th Paradigm.," *Semant. Web*, vol. 4, no. 3, pp. 233–235, 2013.

[56] S. Rong, X. Niu, E. W. Xiang, H. Wang, Q. Yang, and Y. Yu, "A machine learning approach for instance matching based on similarity metrics," in *International Semantic Web Conference*, 2012, pp. 460–475.

[57] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synth. Lect. Semant. web theory Technol.*, vol. 1, no. 1, pp. 1–136, 2011.

[58] A. Ferraram, A. Nikolov, and F. Scharffe, "Data linking for the semantic web," *Semant. Web Ontol. Knowl. Base Enabled Tools, Serv. Appl.*, vol. 169, p. 326, 2013.

[59] "Apache Marmotta LDClient." [Online]. Available: http://marmotta.apache.org/ldclient/. [Accessed: 07-Mar-2018].

[60] "Virtuoso Sponger." [Online]. Available: https://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger. [Accessed: 07-Mar-

2018].

[61] "morph-RDB." [Online]. Available: https://github.com/oeg-upm/morph-rdb. [Accessed: 07-Mar-2018].

[62] "LODRefine." [Online]. Available: https://github.com/sparkica/LODRefine. [Accessed: 07-Mar-2018].

[63] "Silk." [Online]. Available: http://silkframework.org/. [Accessed: 07-Mar-2018].

[64] "LIMES." [Online]. Available: http://aksw.org/Projects/LIMES.html. [Accessed: 07-Mar-2018].

[65] "CSV2RDF4lod." [Online]. Available: https://github.com/timrdf/csv2rdf4lod-automation/wiki. [Accessed: 07-Mar-2018].

[66] "Any23." [Online]. Available: https://any23.apache.org. [Accessed: 07-Mar-2018].

[67] "D2RServer." [Online]. Available: http://d2rq.org/d2r-server. [Accessed: 07-Mar-2018].

[68] "Sparqlify." [Online]. Available: http://aksw.org/Projects/Sparqlify.html. [Accessed: 07-Mar-2018].

[69] "LinDA." [Online]. Available: http://linda-project.eu/tools/. [Accessed: 07-Mar-2018].

[70] N. Marz and J. Warren, *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2015.

[71] V. Turner, J. F. Gantz, D. Reinsel, and S. Minton, "The digital universe of opportunities: Rich data and the increasing value of the internet of things," *IDC Anal. Futur.*, p. 5, 2014.

[72] "Oracle." [Online]. Available: https://www.oracle.com/database/index.html. [Accessed: 07-Mar-2018].

[73] "MySQL." [Online]. Available: https://www.mysql.com/. [Accessed: 07-Mar-2018].

[74] "Microsoft SQL Server." [Online]. Available: https://www.microsoft.com/en-us/sql-server/sql-server-2017. [Accessed: 07-Mar-2018].

[75] "PostgreSQL." [Online]. Available: https://www.postgresql.org/. [Accessed: 07-Mar-2018].

[76] "MongoDB." [Online]. Available: https://www.mongodb.com/. [Accessed: 07-Mar-2018].

[77] "HBase." [Online]. Available: https://hbase.apache.org/. [Accessed: 07-Mar-2018].

[78] "Cassandra." [Online]. Available: http://cassandra.apache.org/. [Accessed: 07-Mar-2018].

[79] "Blazegraph." [Online]. Available: https://www.blazegraph.com/. [Accessed: 07-Mar-2018].

[80] "Neo4j." [Online]. Available: https://neo4j.com/. [Accessed: 29-Mar-2018].

[81] CallidusCloud, "OrientDB." [Online]. Available: https://orientdb.com/. [Accessed: 29-Mar-2018].

[82] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*, 2010, pp. 1–10.

[83] "Apache Hadoop." [Online]. Available: https://hadoop.apache.org/. [Accessed: 07-Mar-2018].

[84] P. Venkatesh and S. Nirmala, "NewSQL-The New Way to Handle Big Data," *Blog. http//www. opensourceforu. com/2012/01/newsql-handle-big-data/. Accessed Nov*, vol. 18, p. 2014, 2012.

[85] "VoltDB." [Online]. Available: https://www.voltdb.com/. [Accessed: 07-Mar-2018].

[86] "Hive." [Online]. Available: https://hive.apache.org/. [Accessed: 07-Mar-2018].

[87] "Apache Pig." [Online]. Available: https://pig.apache.org/. [Accessed: 07-Mar-2018].

[88] "Impala." [Online]. Available: https://impala.apache.org/. [Accessed: 07-Mar-2018].

[89] "Apache Accumulo." [Online]. Available: https://accumulo.apache.org/. [Accessed: 07-Mar-2018].

[90] "Apache Spark." [Online]. Available: https://spark.apache.org/. [Accessed: 07-Mar-2018].

[91] "ElasticSearch." [Online]. Available: https://www.elastic.co/. [Accessed: 07-Mar-2018].

[92] "SparqlMap." [Online]. Available: https://github.com/tomatophantastico/sparqlmap/. [Accessed: 07-Mar-2018].

[93] "FedX." .

[94] "Virtuoso." [Online]. Available: https://virtuoso.openlinksw.com/. [Accessed: 07-Mar-2018].

[95] "Apache Solr." [Online]. Available: http://lucene.apache.org/solr/. [Accessed: 07-Mar-2018].

[96] "SANSA." [Online]. Available: http://sansa-stack.net/. [Accessed: 29-Mar-2018].

[97] Apache Software Foundation, "Apache Flink." [Online]. Available: https://flink.apache.org/. [Accessed: 29-Mar-2018].

[98] P. Russom, "Big data analytics," *TDWI best Pract. report, fourth Quart.*, vol. 19, no. 4, pp. 1–34, 2011.

[99] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in {P}ython," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[100] T. Abeel, Y. Van de Peer, and Y. Saeys, "Java-ml: A machine learning library," *J. Mach. Learn. Res.*, vol. 10, no. Apr, pp. 931–934, 2009.

[101] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The {WEKA} data mining software: an update," *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009.

[102] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.," 2009.

[103] Intel, "Intel® Data Analytics Acceleration Library." [Online]. Available: https://software.intel.com/intel-daal. [Accessed: 07-Mar-2018].

[104] X. Meng *et al.*, "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1235–1241, 2016.

[105] H2O.ai, "package_or_module_title." .

[106] Apache Software Foundation, "Apache Mahout." [Online]. Available: http://mahout.apache.org//. [Accessed: 07-Mar-2018].

[107] Apache Software Foundation, "Apache PredictionIO." [Online]. Available: https://predictionio.apache.org/. [Accessed: 26-Mar-2018].

[108] IBM, "IBM Watson Data Platform." [Online]. Available: https://www.ibm.com/analytics/us/en/watson-data-platform/. [Accessed: 07-Mar-2018].

[109] RapidMiner, “RapidMiner.” [Online]. Available: https://rapidminer.com/. [Accessed: 07-Mar-2018].

[110] IBM, “IBM SPSS Statistics.” [Online]. Available: https://www.ibm.com/products/spss-statistics. [Accessed: 07-Mar-2018].

[111] StataCorp, “Stata.” [Online]. Available: https://www.stata.com/. [Accessed: 07-Mar-2018].

[112] J. Nickolls, I. Buck, M. Garland, and K. Skadron, “Scalable parallel programming with CUDA,” in *ACM SIGGRAPH 2008 classes*, 2008, p. 16.

[113] “Keras.” [Online]. Available: https://keras.io/. [Accessed: 07-Mar-2018].

[114] M. Abadi *et al.*, “TensorFlow: A System for Large-Scale Machine Learning.,” in *OSDI*, 2016, vol. 16, pp. 265–283.

[115] Microsoft, “Microsoft Cognitive Toolkit.” [Online]. Available: https://www.microsoft.com/en-us/cognitive-toolkit/. [Accessed: 07-Mar-2018].

[116] R. Al-Rfou *et al.*, “Theano: A {Python} framework for fast computation of mathematical expressions,” *arXiv e-prints*, vol. abs/1605.0, May 2016.

[117] T. Chen *et al.*, “Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems,” *arXiv Prepr. arXiv1512.01274*, 2015.

[118] “Gluon.” [Online]. Available: https://github.com/gluon-api/gluon-api/. [Accessed: 07-Mar-2018].

[119] R. Collobert, K. Kavukcuoglu, and C. Farabet, “Torch7: A matlab-like environment for machine learning,” in *BigLearn, NIPS workshop*, 2011, no. EPFL-CONF-192376.

[120] A. Paszke *et al.*, “Automatic differentiation in PyTorch,” 2017.

[121] Y. Jia *et al.*, “Caffe: Convolutional Architecture for Fast Feature Embedding,” *arXiv Prepr. arXiv1408.5093*, 2014.

[122] Facebook, “Caffe2.” [Online]. Available: https://caffe2.ai/. [Accessed: 07-Mar-2018].

[123] “BigDL.” [Online]. Available: https://bigdl-project.github.io/0.4.0/. [Accessed: 07-Mar-2018].

[124] Intel, “Intel Math Kernel Library.” [Online]. Available: https://software.intel.com/en-us/mkl. [Accessed: 07-Mar-2018].

[125] “Deeplearning4j.” [Online]. Available: https://deeplearning4j.org/. [Accessed: 07-Mar-2018].

[126] “PaddlePaddle.” [Online]. Available: http://paddlepaddle.org/. [Accessed: 07-Mar-2018].

[127] U. M. Fayyad, A. Wierse, and G. G. Grinstein, *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2002.

[128] “Apache Zeppelin.” [Online]. Available: https://zeppelin.apache.org/. [Accessed: 30-Mar-2018].

[129] “Beaker.” [Online]. Available: http://beakernotebook.com/. [Accessed: 30-Mar-2018].

[130] “Jupyter.” [Online]. Available: http://jupyter.org/. [Accessed: 30-Mar-2018].

[131] B. Shneiderman, “The eyes have it: A task by data type taxonomy for information visualizations,” in *The Craft of Information Visualization*, Elsevier, 2003, pp. 364–371.

[132] “15-most-common-types-of-data-visualisation.”

[133] D. E. Kee, L. Salowitz, and R. Chang, “Comparing interactive web-based visualization rendering techniques,” *Tufts Univ. Medford, MA*, 2012.

[134] X. Gong, Y. Jin, Y. Cui, and T. Yang, “Web visualization of distributed network measurement system based on HTML5,” in *Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on*, 2012, vol. 2, pp. 519–523.

[135] “amCharts.” [Online]. Available: https://github.com/amcharts. [Accessed: 07-Mar-2018].

[136] “AnyChart.” [Online]. Available: https://github.com/AnyChart. [Accessed: 07-Mar-2018].

[137] “CanvasJS.” [Online]. Available: https://canvasjs.com. [Accessed: 07-Mar-2018].

[138] “Chart.js.” [Online]. Available: https://github.com/chartjs. [Accessed: 07-Mar-2018].

[139] “Chartist.js.” [Online]. Available: https://github.com/gionkunz/chartist-js. [Accessed: 07-Mar-2018].

[140] “Cytoscape.js.” [Online]. Available: https://github.com/cytoscape/cytoscape.js. [Accessed: 07-Mar-2018].

[141] “D3.js.” [Online]. Available: https://github.com/d3. [Accessed: 07-Mar-2018].

[142] “Datawrapper.” [Online]. Available: https://github.com/datawrapper. [Accessed: 07-Mar-2018].

[143] “dc-js.” [Online]. Available: https://github.com/dc-js. [Accessed: 07-Mar-2018].

[144] “Dygraphs.” [Online]. Available: https://github.com/danvk/dygraphs. [Accessed: 07-Mar-2018].

[145] “Echarts.” [Online]. Available: https://github.com/ecomfe/echarts. [Accessed: 07-Mar-2018].

[146] “FusionCharts.” [Online]. Available: https://www.fusioncharts.com. [Accessed: 07-Mar-2018].

[147] “Google Charts.” [Online]. Available: https://developers.google.com/chart. [Accessed: 07-Mar-2018].

[148] “Highcharts.” [Online]. Available: https://www.highcharts.com. [Accessed: 07-Mar-2018].

[149] “Kibana.” [Online]. Available: https://github.com/elastic/kibana. [Accessed: 07-Mar-2018].

[150] “Knowage.” [Online]. Available: https://www.knowage-suite.com. [Accessed: 07-Mar-2018].

[151] “Leaflet.” [Online]. Available: https://github.com/Leaflet. [Accessed: 07-Mar-2018].

[152] “Mandola Dashboard.” [Online]. Available: http://mandola.grid.ucy.ac.cy/. [Accessed: 07-Mar-2018].

[153] “Metabase.” [Online]. Available: https://github.com/metabase. [Accessed: 07-Mar-2018].

[154] “OpenLayers.” [Online]. Available: https://github.com/openlayers. [Accessed: 07-Mar-2018].

[155] “plotly.js.” [Online]. Available: https://github.com/plotly/plotly.js. [Accessed: 07-Mar-2018].

[156] “RAWGraphs.” [Online]. Available: https://github.com/densitydesign/raw. [Accessed: 07-Mar-2018].

[157] “Sigma.js.” [Online]. Available: https://github.com/jacomyal/sigma.js. [Accessed: 07-Mar-2018].

[158] “Tableau.” [Online]. Available: https://www.tableau.com. [Accessed: 07-Mar-2018].

[159] "ZingChart." [Online]. Available: https://www.zingchart.com/. [Accessed: 07-Mar-2018].

[160] "BigDataOcean." [Online]. Available: http://www.bigdataocean.eu. [Accessed: 07-Mar-2018].

[161] "AEGIS." [Online]. Available: http://www.aegis-bigdata.eu. [Accessed: 07-Mar-2018].

[162] "TOREADOR." [Online]. Available: http://www.toreador-project.eu/. [Accessed: 07-Mar-2018].

[163] "UNICORN." [Online]. Available: http://www.unicorn-project.eu. [Accessed: 07-Mar-2018].

[164] "My Health - My Data." [Online]. Available: http://www.myhealthmydata.eu/. [Accessed: 30-Mar-2018].

[165] "SPECIAL." [Online]. Available: https://www.specialprivacy.eu/. [Accessed: 30-Mar-2018].

[166] "proDataMarket." [Online]. Available: https://prodatamarket.eu. [Accessed: 07-Mar-2018].

[167] FlightGlobal, "The Big Data Landscape." [Online]. Available: https://www.flightglobal.com/news/articles/insight-from-flightglobal-the-big-data-landscape-446681/. [Accessed: 23-Mar-2018].

[168] International Air Transport Association (IATA), "FUTURE OF THE AIRLINE INDUSTRY 2035." [Online]. Available: https://www.iata.org/policy/Documents/iata-future-airline-industry.pdf. [Accessed: 11-Mar-2018].

[169] DZone, "Artificial Intelligence: Machine Learning and Predictive Analytics." [Online]. Available: https://dzone.com/guides/artificial-intelligence-machine-learning-and-predi. [Accessed: 11-Mar-2018].

[170] DZone, "Big Data: Data Science and Advanced Analytics." [Online]. Available: https://dzone.com/guides/big-data-data-science-and-advanced-analytics. [Accessed: 11-Mar-2018].

[171] A. R. G. Harteveldt, Henry H., "The Future of Airline Distribution, 2016 - 2021." [Online]. Available: https://www.iata.org/whatwedo/airline-distribution/ndc/Documents/ndc-future-airline-distribution-report.pdf. [Accessed: 11-Mar-2018].

[172] DZone, "Big Data: Stream Processing, Statistics, and Scalability." [Online]. Available: https://dzone.com/guides/big-data-stream-processing-statistics-and-scalabil. [Accessed: 11-Mar-2018].

[173] MRO, "MRO BIG DATA – A LION OR A LAMB?" [Online]. Available: http://www.oliverwyman.com/content/dam/oliver-wyman/global/en/2016/apr/NYC-MKT9202-001MRO-Survey-2016_web.pdf. [Accessed: 11-Mar-2018].

[174] DZone, "Databases: Speed, Scale, and Security." [Online]. Available: https://dzone.com/guides/databases-speed-scale-and-security. [Accessed: 11-Mar-2018].

[175] S. Villata and F. Gandon, "Licenses compatibility and composition in the web of data," in *Third International Workshop on Consuming Linked Data (COLD2012)*, 2012.

[176] H.-L. Truong, M. Comerio, F. De Paoli, G. R. Gangadharan, and S. Dustdar, "Data contracts for cloud-based data marketplaces," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 4, pp. 280–295, 2012.

[177] T.-D. Cao, T.-V. Pham, Q.-H. Vu, H.-L. Truong, D.-H. Le, and S. Dustdar, "MARSA: A marketplace for realtime human sensing data," *ACM Trans. Internet Technol.*, vol. 16, no. 3, p. 16, 2016.

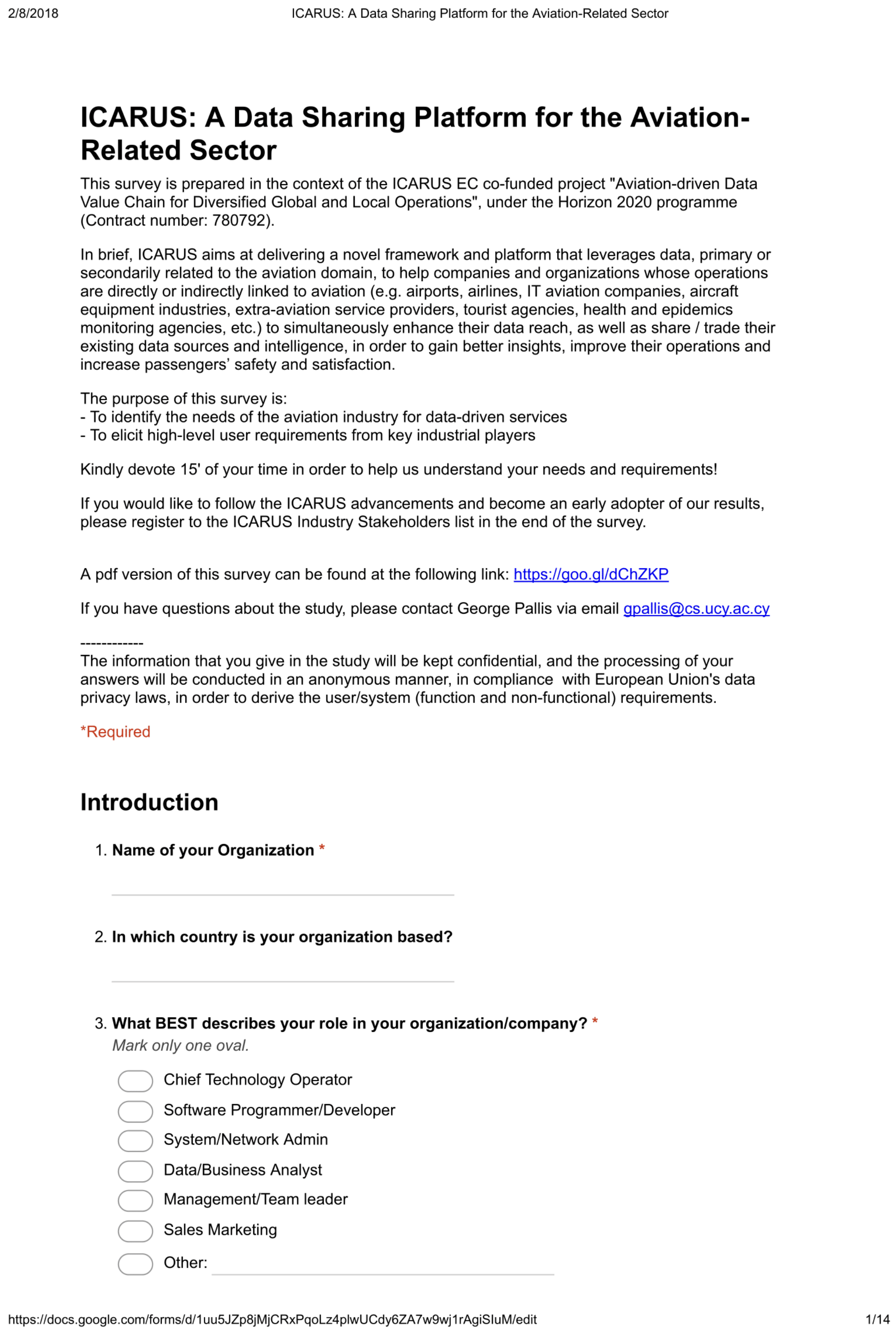
[178] Q. H. Vu, T.-V. Pham, H.-L. Truong, S. Dustdar, and R. Asal, "Demods: A description model for data-as-a-service," in *Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International*

Conference on, 2012, pp. 605–612.

7 Annex

7.1 Disseminated Questionnaire

In what follows is in printable format the ICARUS questionnaire. The online version of the questionnaire is accessible via the following link: https://goo.gl/forms/7Sn0JLieK4Rd0OGv1



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

ICARUS: A Data Sharing Platform for the Aviation-Related Sector

This survey is prepared in the context of the ICARUS EC co-funded project "Aviation-driven Data Value Chain for Diversified Global and Local Operations", under the Horizon 2020 programme (Contract number: 780792).

In brief, ICARUS aims at delivering a novel framework and platform that leverages data, primary or secondarily related to the aviation domain, to help companies and organizations whose operations are directly or indirectly linked to aviation (e.g. airports, airlines, IT aviation companies, aircraft equipment industries, extra-aviation service providers, tourist agencies, health and epidemics monitoring agencies, etc.) to simultaneously enhance their data reach, as well as share / trade their existing data sources and intelligence, in order to gain better insights, improve their operations and increase passengers' safety and satisfaction.

The purpose of this survey is:
- To identify the needs of the aviation industry for data-driven services
- To elicit high-level user requirements from key industrial players

Kindly devote 15' of your time in order to help us understand your needs and requirements!

If you would like to follow the ICARUS advancements and become an early adopter of our results, please register to the ICARUS Industry Stakeholders list in the end of the survey.

A pdf version of this survey can be found at the following link: https://goo.gl/dChZKP

If you have questions about the study, please contact George Pallis via email gpallis@cs.ucy.ac.cy

----
The information that you give in the study will be kept confidential, and the processing of your answers will be conducted in an anonymous manner, in compliance with European Union's data privacy laws, in order to derive the user/system (function and non-functional) requirements.

*Required

Introduction

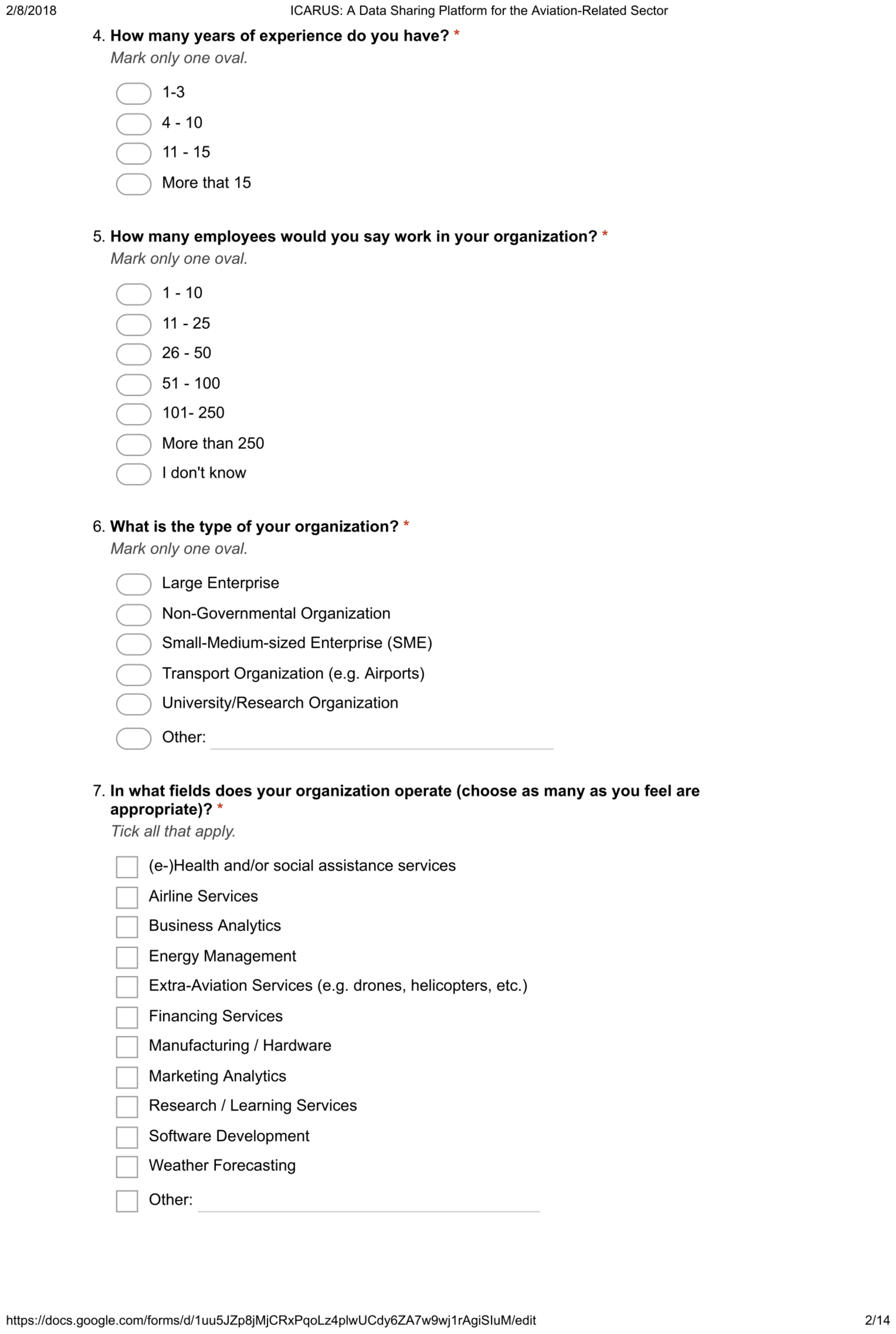
1. **Name of your Organization** *

2. **In which country is your organization based?**

3. **What BEST describes your role in your organization/company?** *
Mark only one oval.

○ Chief Technology Operator
○ Software Programmer/Developer
○ System/Network Admin
○ Data/Business Analyst
○ Management/Team leader
○ Sales Marketing
○ Other:

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSluM/edit 1/14



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

4. **How many years of experience do you have?** *

Mark only one oval.

○ 1-3
○ 4 - 10
○ 11 - 15
○ More that 15

5. **How many employees would you say work in your organization?** *

Mark only one oval.

○ 1 - 10
○ 11 - 25
○ 26 - 50
○ 51 - 100
○ 101- 250
○ More than 250
○ I don't know

6. **What is the type of your organization?** *

Mark only one oval.

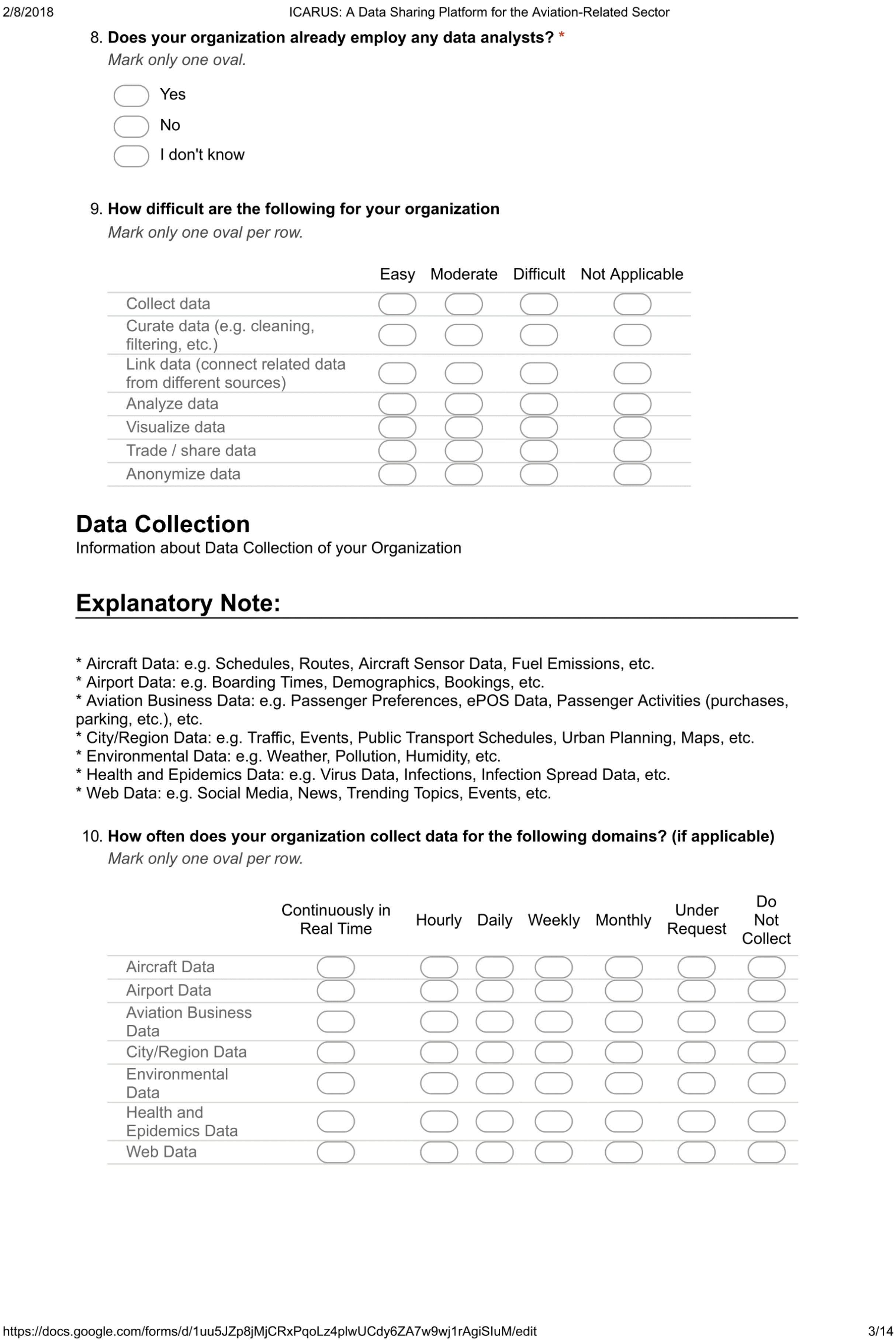
○ Large Enterprise
○ Non-Governmental Organization
○ Small-Medium-sized Enterprise (SME)
○ Transport Organization (e.g. Airports)
○ University/Research Organization
○ Other: ____________________

7. **In what fields does your organization operate (choose as many as you feel are appropriate)?** *

Tick all that apply.

□ (e-)Health and/or social assistance services
□ Airline Services
□ Business Analytics
□ Energy Management
□ Extra-Aviation Services (e.g. drones, helicopters, etc.)
□ Financing Services
□ Manufacturing / Hardware
□ Marketing Analytics
□ Research / Learning Services
□ Software Development
□ Weather Forecasting
□ Other: ____________________

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSluM/edit 2/14



8. **Does your organization already employ any data analysts?** *

Mark only one oval.

○ Yes
○ No
○ I don't know

9. **How difficult are the following for your organization**

Mark only one oval per row.

	Easy	Moderate	Difficult	Not Applicable
Collect data	○	○	○	○
Curate data (e.g. cleaning, filtering, etc.)	○	○	○	○
Link data (connect related data from different sources)	○	○	○	○
Analyze data	○	○	○	○
Visualize data	○	○	○	○
Trade / share data	○	○	○	○
Anonymize data	○	○	○	○

Data Collection

Information about Data Collection of your Organization

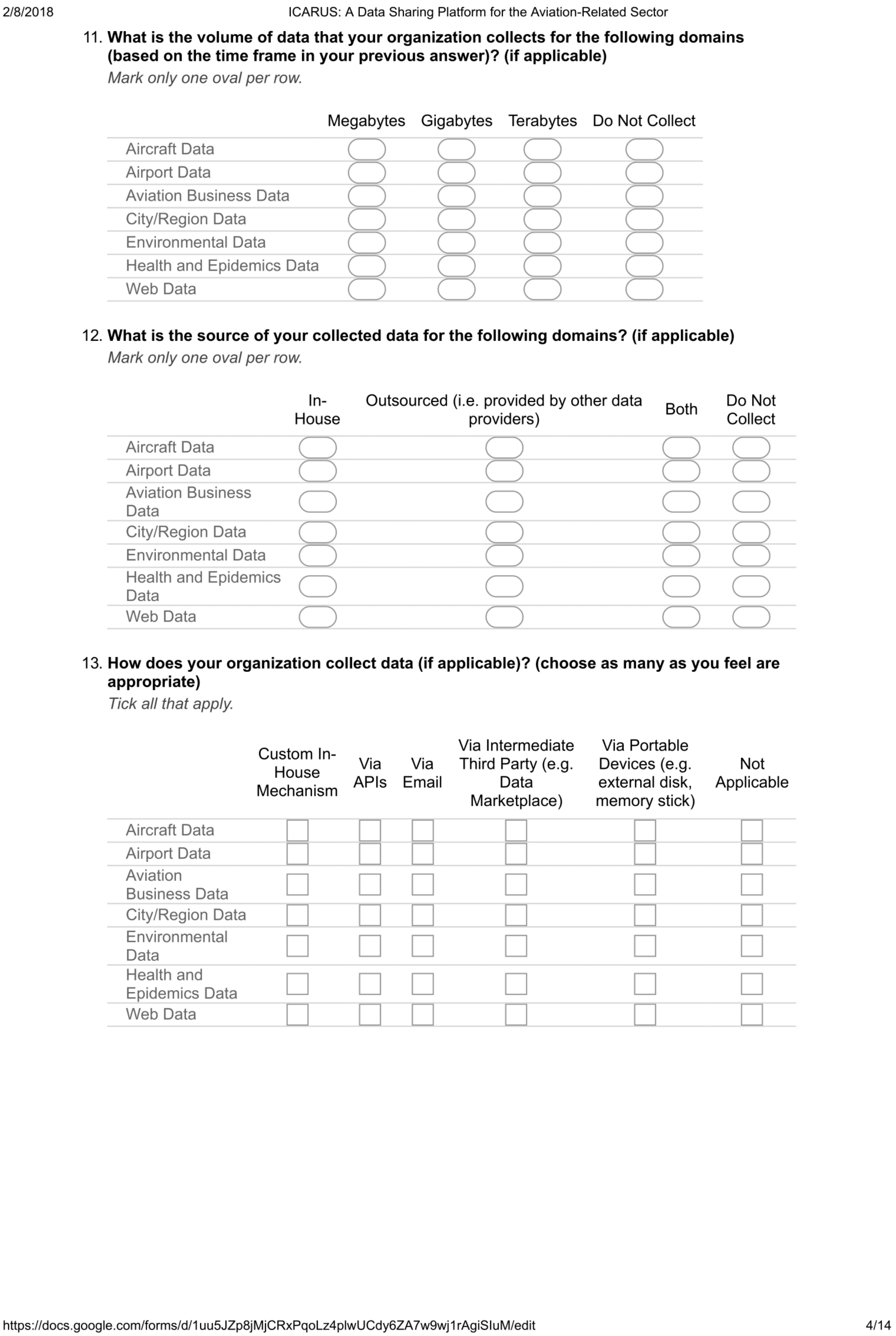
Explanatory Note:

* Aircraft Data: e.g. Schedules, Routes, Aircraft Sensor Data, Fuel Emissions, etc.
* Airport Data: e.g. Boarding Times, Demographics, Bookings, etc.
* Aviation Business Data: e.g. Passenger Preferences, ePOS Data, Passenger Activities (purchases, parking, etc.), etc.
* City/Region Data: e.g. Traffic, Events, Public Transport Schedules, Urban Planning, Maps, etc.
* Environmental Data: e.g. Weather, Pollution, Humidity, etc.
* Health and Epidemics Data: e.g. Virus Data, Infections, Infection Spread Data, etc.
* Web Data: e.g. Social Media, News, Trending Topics, Events, etc.

10. **How often does your organization collect data for the following domains? (if applicable)**

Mark only one oval per row.

	Continuously in Real Time	Hourly	Daily	Weekly	Monthly	Under Request	Do Not Collect
Aircraft Data	○	○	○	○	○	○	○
Airport Data	○	○	○	○	○	○	○
Aviation Business Data	○	○	○	○	○	○	○
City/Region Data	○	○	○	○	○	○	○
Environmental Data	○	○	○	○	○	○	○
Health and Epidemics Data	○	○	○	○	○	○	○
Web Data	○	○	○	○	○	○	○



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

11. **What is the volume of data that your organization collects for the following domains (based on the time frame in your previous answer)? (if applicable)**

Mark only one oval per row.

	Megabytes	Gigabytes	Terabytes	Do Not Collect
Aircraft Data	○	○	○	○
Airport Data	○	○	○	○
Aviation Business Data	○	○	○	○
City/Region Data	○	○	○	○
Environmental Data	○	○	○	○
Health and Epidemics Data	○	○	○	○
Web Data	○	○	○	○

12. **What is the source of your collected data for the following domains? (if applicable)**

Mark only one oval per row.

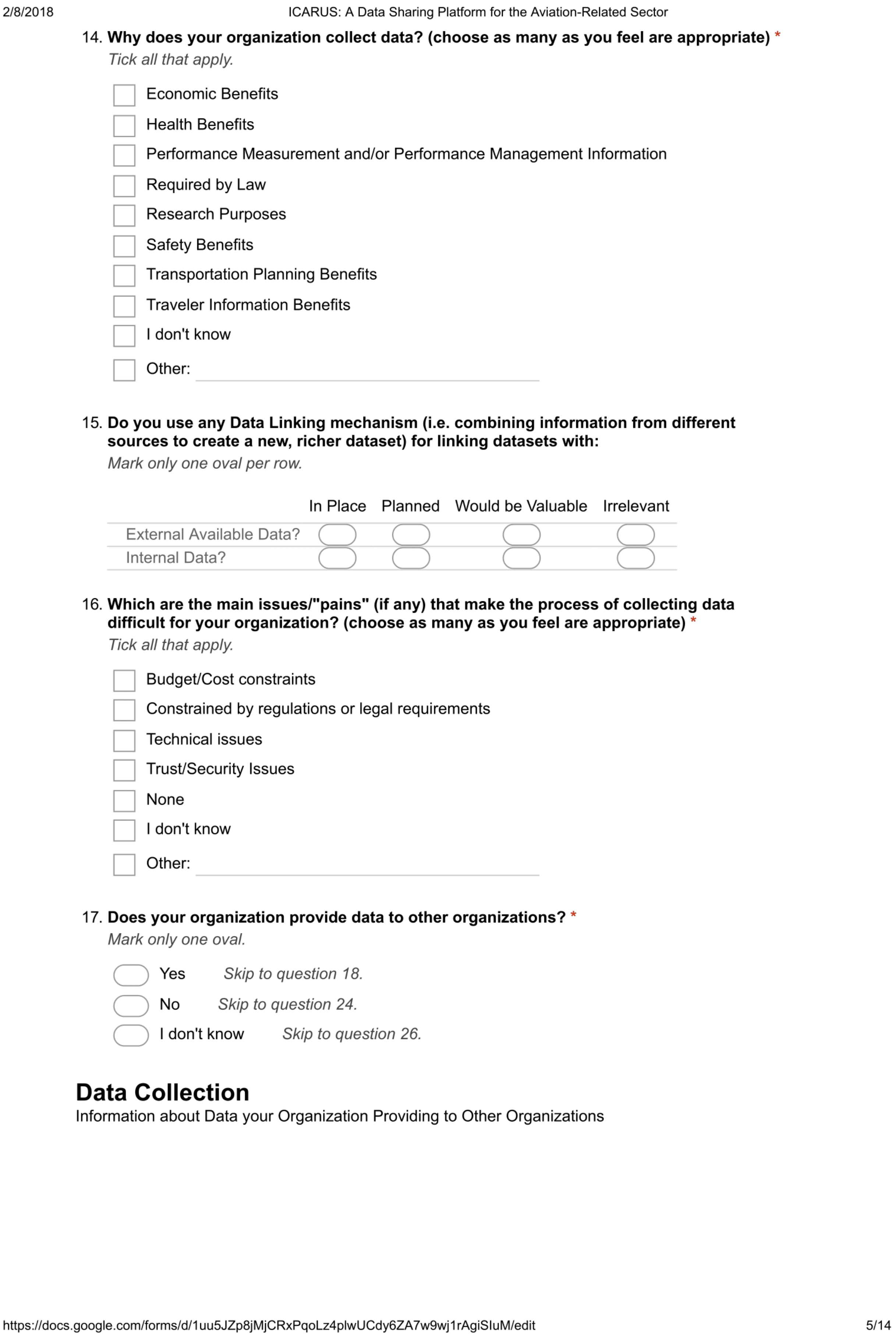
	In-House	Outsourced (i.e. provided by other data providers)	Both	Do Not Collect
Aircraft Data	○	○	○	○
Airport Data	○	○	○	○
Aviation Business Data	○	○	○	○
City/Region Data	○	○	○	○
Environmental Data	○	○	○	○
Health and Epidemics Data	○	○	○	○
Web Data	○	○	○	○

13. **How does your organization collect data (if applicable)? (choose as many as you feel are appropriate)**

Tick all that apply.

	Custom In-House Mechanism	Via APIs	Via Email	Via Intermediate Third Party (e.g. Data Marketplace)	Via Portable Devices (e.g. external disk, memory stick)	Not Applicable
Aircraft Data	□	□	□	□	□	□
Airport Data	□	□	□	□	□	□
Aviation Business Data	□	□	□	□	□	□
City/Region Data	□	□	□	□	□	□
Environmental Data	□	□	□	□	□	□
Health and Epidemics Data	□	□	□	□	□	□
Web Data	□	□	□	□	□	□

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSIuM/edit 4/14



14. **Why does your organization collect data? (choose as many as you feel are appropriate)** *

Tick all that apply.

□ Economic Benefits
□ Health Benefits
□ Performance Measurement and/or Performance Management Information
□ Required by Law
□ Research Purposes
□ Safety Benefits
□ Transportation Planning Benefits
□ Traveler Information Benefits
□ I don't know
□ Other: ______________________________

15. **Do you use any Data Linking mechanism (i.e. combining information from different sources to create a new, richer dataset) for linking datasets with:**

Mark only one oval per row.

	In Place	Planned	Would be Valuable	Irrelevant
External Available Data?	○	○	○	○
Internal Data?	○	○	○	○

16. **Which are the main issues/"pains" (if any) that make the process of collecting data difficult for your organization? (choose as many as you feel are appropriate)** *

Tick all that apply.

□ Budget/Cost constraints
□ Constrained by regulations or legal requirements
□ Technical issues
□ Trust/Security Issues
□ None
□ I don't know
□ Other: ______________________________

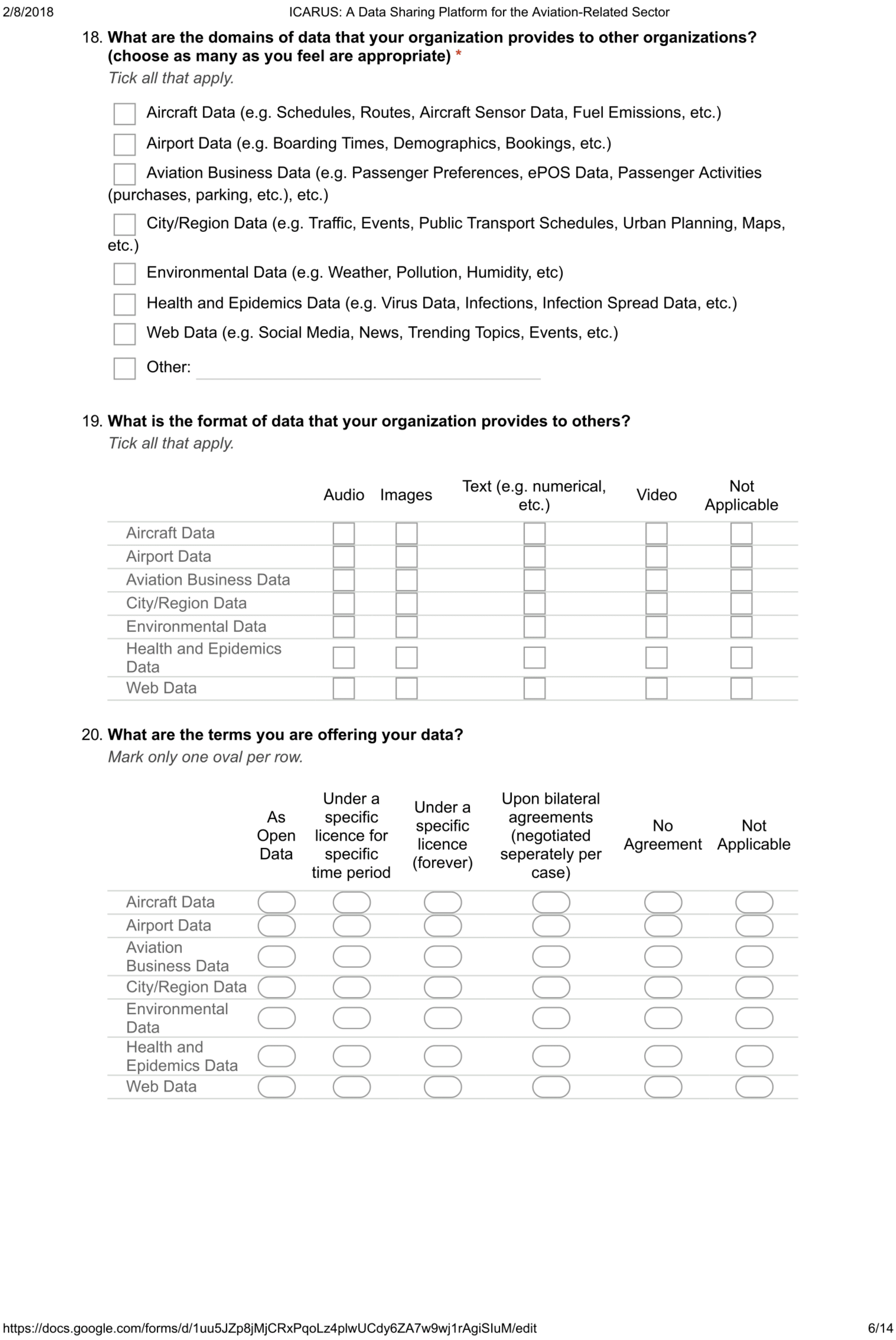
17. **Does your organization provide data to other organizations?** *

Mark only one oval.

○ Yes *Skip to question 18.*
○ No *Skip to question 24.*
○ I don't know *Skip to question 26.*

Data Collection

Information about Data your Organization Providing to Other Organizations



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

18. **What are the domains of data that your organization provides to other organizations? (choose as many as you feel are appropriate)** *

Tick all that apply.

□ Aircraft Data (e.g. Schedules, Routes, Aircraft Sensor Data, Fuel Emissions, etc.)

□ Airport Data (e.g. Boarding Times, Demographics, Bookings, etc.)

□ Aviation Business Data (e.g. Passenger Preferences, ePOS Data, Passenger Activities (purchases, parking, etc.), etc.)

□ City/Region Data (e.g. Traffic, Events, Public Transport Schedules, Urban Planning, Maps, etc.)

□ Environmental Data (e.g. Weather, Pollution, Humidity, etc)

□ Health and Epidemics Data (e.g. Virus Data, Infections, Infection Spread Data, etc.)

□ Web Data (e.g. Social Media, News, Trending Topics, Events, etc.)

□ Other: ______________________

19. **What is the format of data that your organization provides to others?**

Tick all that apply.

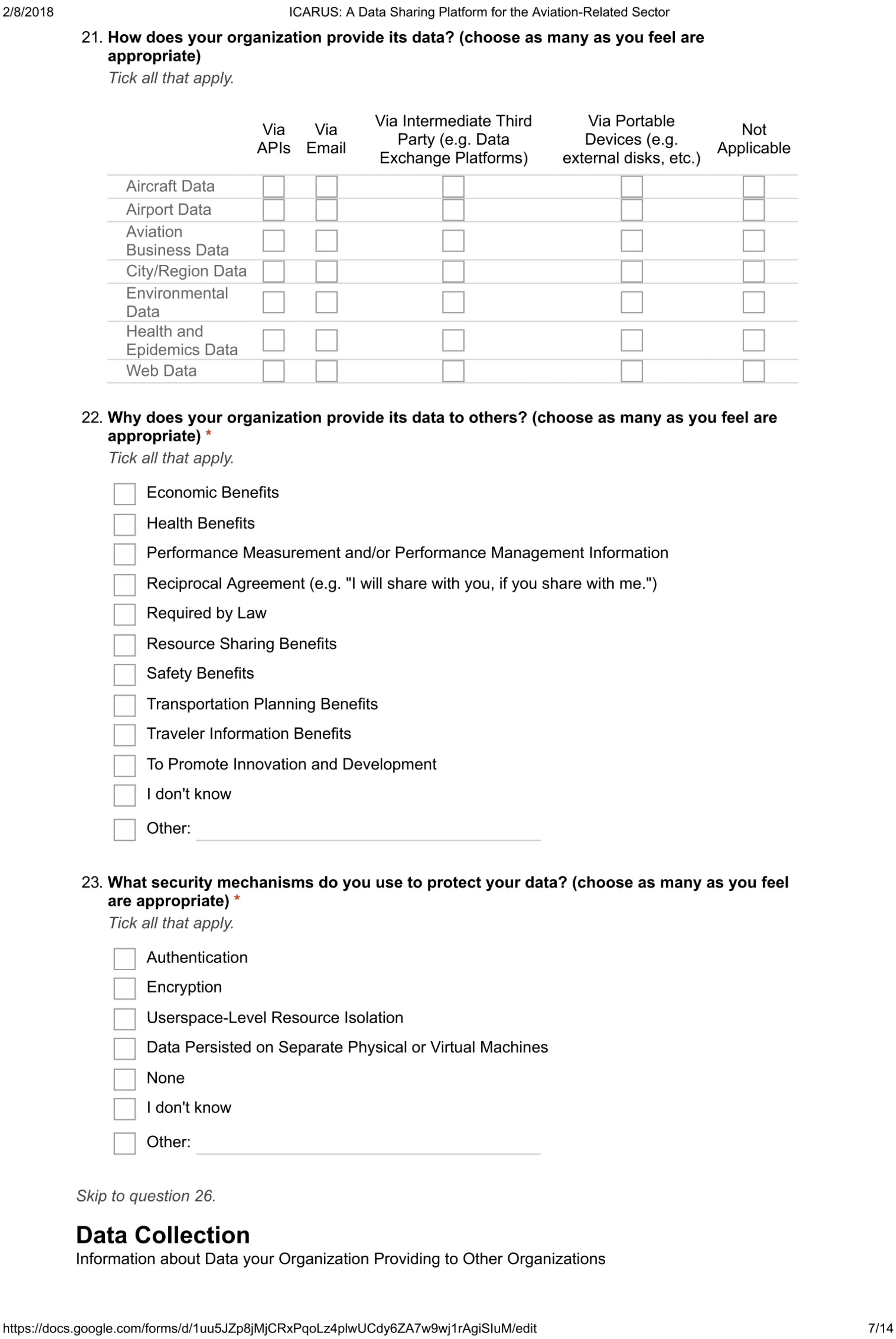
	Audio	Images	Text (e.g. numerical, etc.)	Video	Not Applicable
Aircraft Data	□	□	□	□	□
Airport Data	□	□	□	□	□
Aviation Business Data	□	□	□	□	□
City/Region Data	□	□	□	□	□
Environmental Data	□	□	□	□	□
Health and Epidemics Data	□	□	□	□	□
Web Data	□	□	□	□	□

20. **What are the terms you are offering your data?**

Mark only one oval per row.

	As Open Data	Under a specific licence for specific time period	Under a specific licence (forever)	Upon bilateral agreements (negotiated seperately per case)	No Agreement	Not Applicable
Aircraft Data	○	○	○	○	○	○
Airport Data	○	○	○	○	○	○
Aviation Business Data	○	○	○	○	○	○
City/Region Data	○	○	○	○	○	○
Environmental Data	○	○	○	○	○	○
Health and Epidemics Data	○	○	○	○	○	○
Web Data	○	○	○	○	○	○

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSIuM/edit 6/14



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

21. **How does your organization provide its data? (choose as many as you feel are appropriate)**

Tick all that apply.

	Via APIs	Via Email	Via Intermediate Third Party (e.g. Data Exchange Platforms)	Via Portable Devices (e.g. external disks, etc.)	Not Applicable
Aircraft Data	□	□	□	□	□
Airport Data	□	□	□	□	□
Aviation Business Data	□	□	□	□	□
City/Region Data	□	□	□	□	□
Environmental Data	□	□	□	□	□
Health and Epidemics Data	□	□	□	□	□
Web Data	□	□	□	□	□

22. **Why does your organization provide its data to others? (choose as many as you feel are appropriate)** *

Tick all that apply.

□ Economic Benefits
□ Health Benefits
□ Performance Measurement and/or Performance Management Information
□ Reciprocal Agreement (e.g. "I will share with you, if you share with me.")
□ Required by Law
□ Resource Sharing Benefits
□ Safety Benefits
□ Transportation Planning Benefits
□ Traveler Information Benefits
□ To Promote Innovation and Development
□ I don't know
□ Other: ____________________

23. **What security mechanisms do you use to protect your data? (choose as many as you feel are appropriate)** *

Tick all that apply.

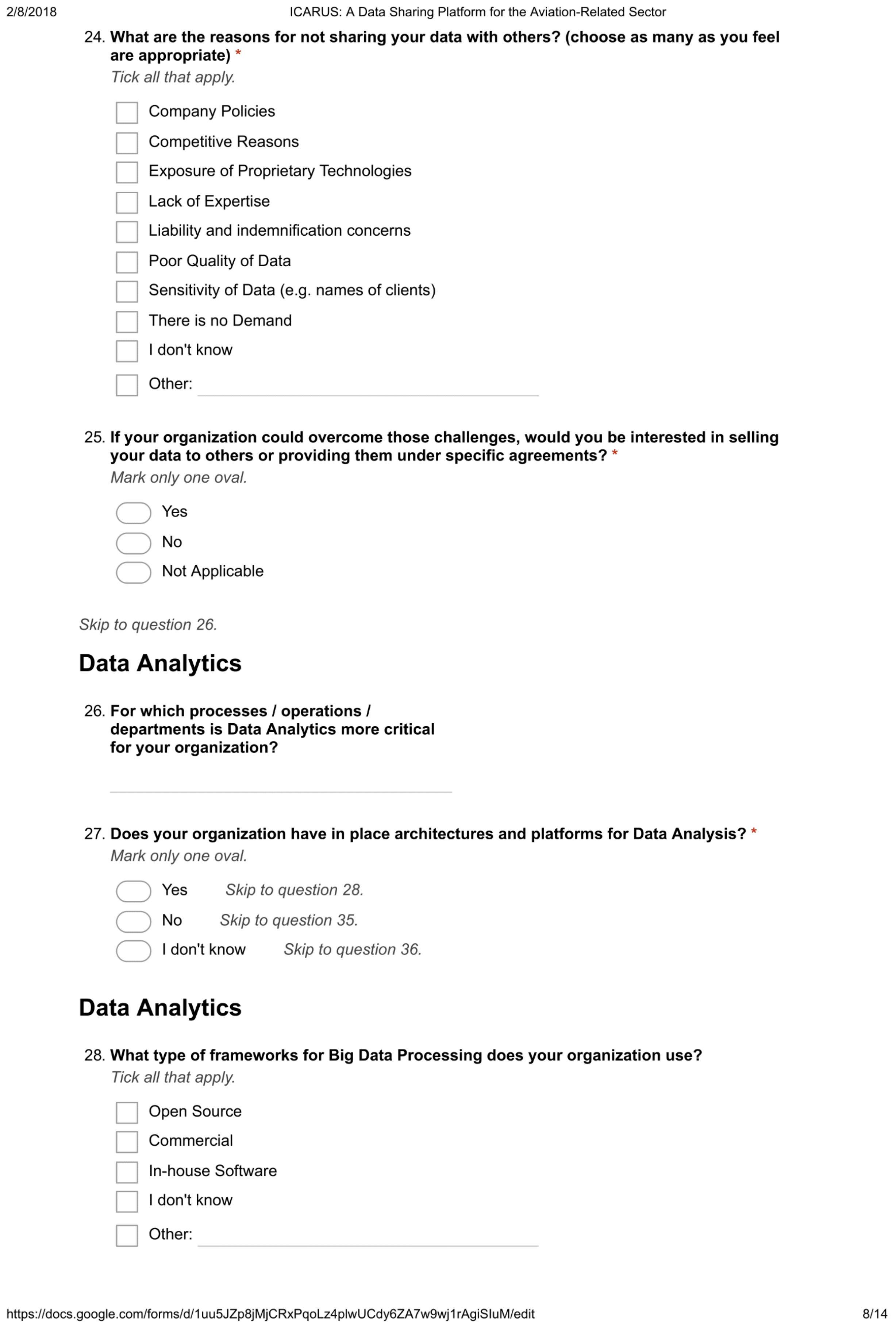
□ Authentication
□ Encryption
□ Userspace-Level Resource Isolation
□ Data Persisted on Separate Physical or Virtual Machines
□ None
□ I don't know
□ Other: ____________________

Skip to question 26.

Data Collection

Information about Data your Organization Providing to Other Organizations

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSIuM/edit 7/14



24. **What are the reasons for not sharing your data with others? (choose as many as you feel are appropriate)** *

Tick all that apply.

□ Company Policies
□ Competitive Reasons
□ Exposure of Proprietary Technologies
□ Lack of Expertise
□ Liability and indemnification concerns
□ Poor Quality of Data
□ Sensitivity of Data (e.g. names of clients)
□ There is no Demand
□ I don't know
□ Other: ____________________

25. **If your organization could overcome those challenges, would you be interested in selling your data to others or providing them under specific agreements?** *

Mark only one oval.

○ Yes
○ No
○ Not Applicable

Skip to question 26.

Data Analytics

26. **For which processes / operations / departments is Data Analytics more critical for your organization?**

____

27. **Does your organization have in place architectures and platforms for Data Analysis?** *

Mark only one oval.

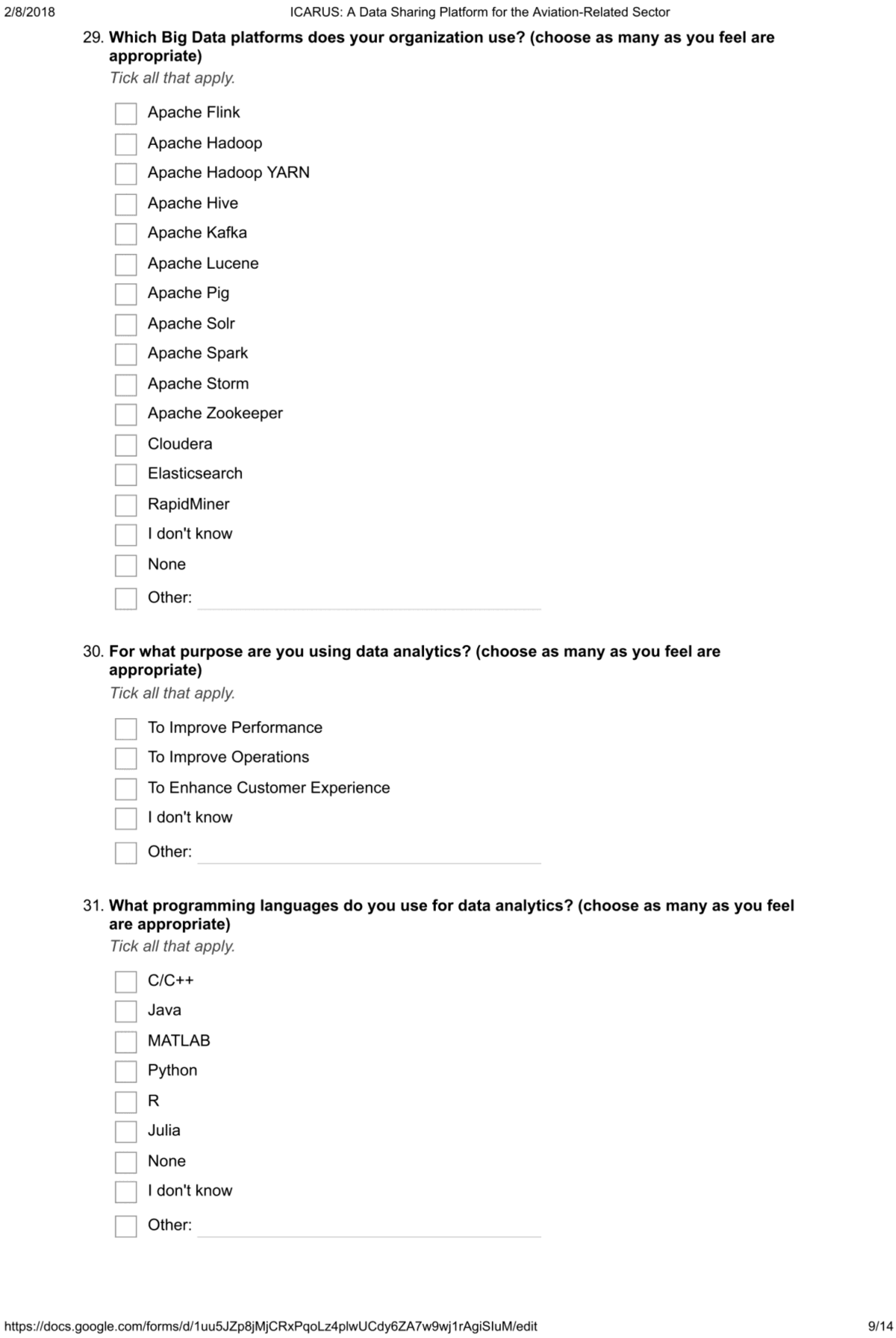
○ Yes *Skip to question 28.*
○ No *Skip to question 35.*
○ I don't know *Skip to question 36.*

Data Analytics

28. **What type of frameworks for Big Data Processing does your organization use?**

Tick all that apply.

□ Open Source
□ Commercial
□ In-house Software
□ I don't know
□ Other: ____________________



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

29. **Which Big Data platforms does your organization use? (choose as many as you feel are appropriate)**

Tick all that apply.

- □ Apache Flink
- □ Apache Hadoop
- □ Apache Hadoop YARN
- □ Apache Hive
- □ Apache Kafka
- □ Apache Lucene
- □ Apache Pig
- □ Apache Solr
- □ Apache Spark
- □ Apache Storm
- □ Apache Zookeeper
- □ Cloudera
- □ Elasticsearch
- □ RapidMiner
- □ I don't know
- □ None
- □ Other: ____________________

30. **For what purpose are you using data analytics? (choose as many as you feel are appropriate)**

Tick all that apply.

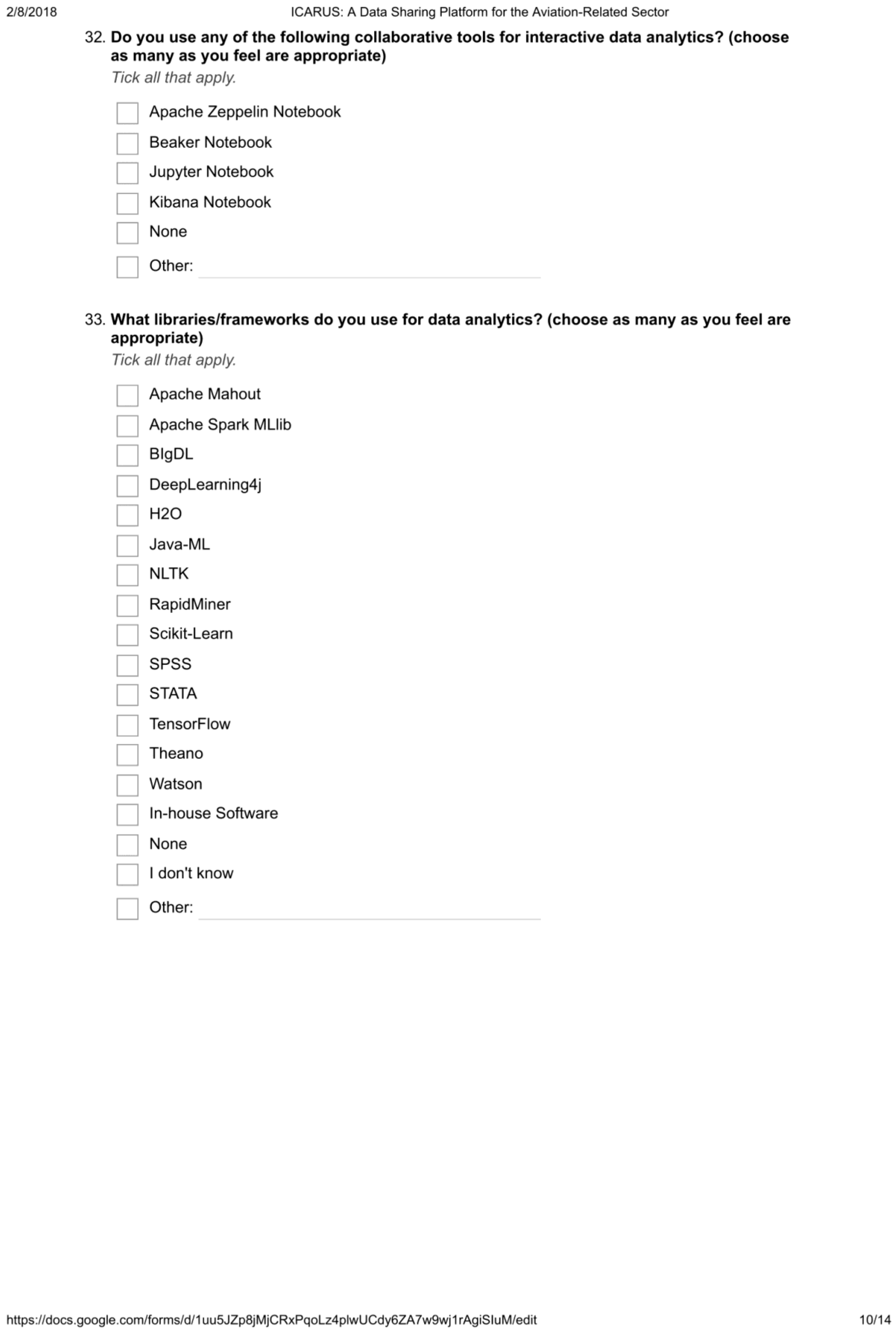
- □ To Improve Performance
- □ To Improve Operations
- □ To Enhance Customer Experience
- □ I don't know
- □ Other: ____________________

31. **What programming languages do you use for data analytics? (choose as many as you feel are appropriate)**

Tick all that apply.

- □ C/C++
- □ Java
- □ MATLAB
- □ Python
- □ R
- □ Julia
- □ None
- □ I don't know
- □ Other: ____________________

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSIuM/edit 9/14



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

32. **Do you use any of the following collaborative tools for interactive data analytics? (choose as many as you feel are appropriate)**

Tick all that apply.

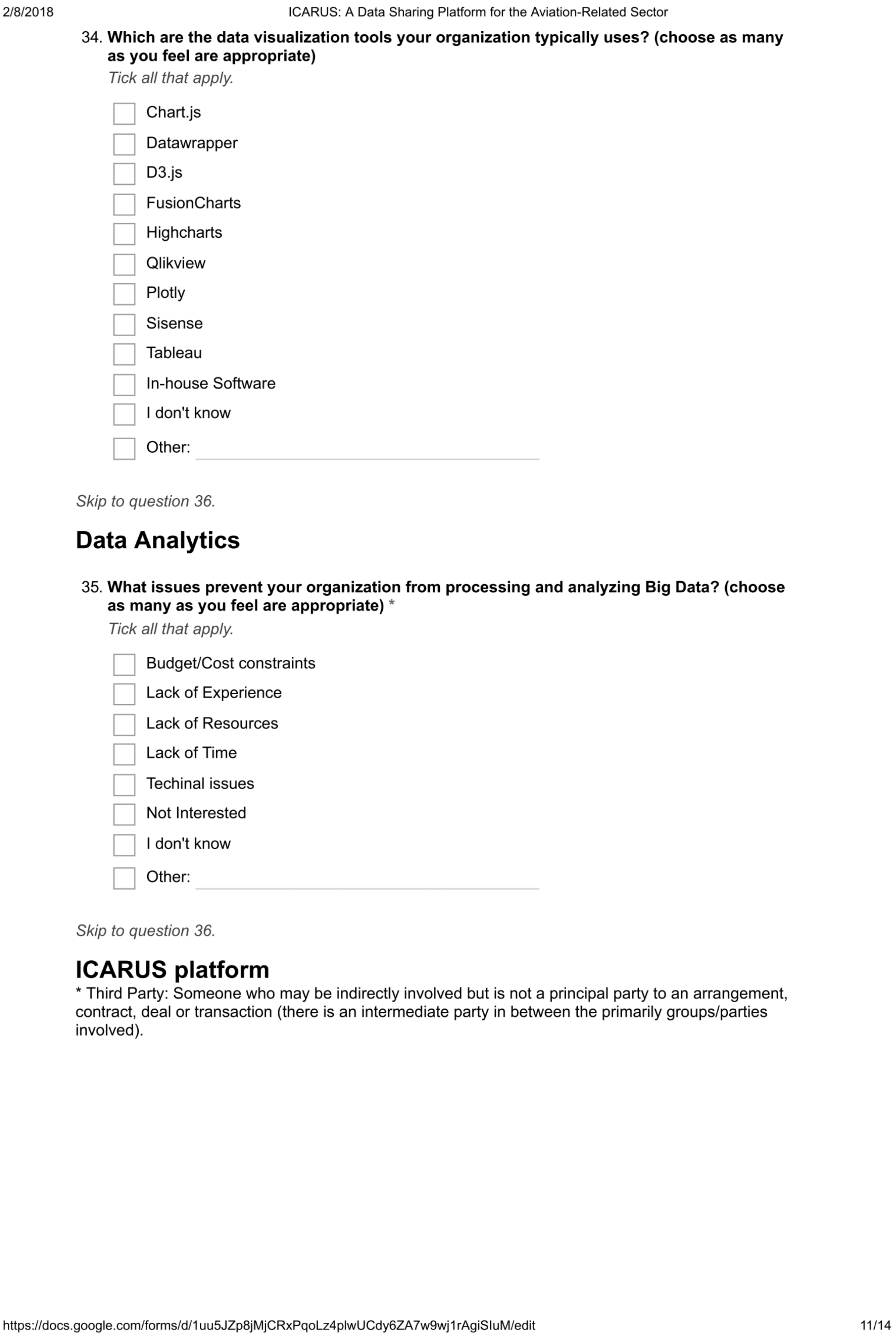
□ Apache Zeppelin Notebook
□ Beaker Notebook
□ Jupyter Notebook
□ Kibana Notebook
□ None
□ Other: ______________________

33. **What libraries/frameworks do you use for data analytics? (choose as many as you feel are appropriate)**

Tick all that apply.

□ Apache Mahout
□ Apache Spark MLlib
□ BIgDL
□ DeepLearning4j
□ H2O
□ Java-ML
□ NLTK
□ RapidMiner
□ Scikit-Learn
□ SPSS
□ STATA
□ TensorFlow
□ Theano
□ Watson
□ In-house Software
□ None
□ I don't know
□ Other: ______________________

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSIuM/edit 10/14



34. **Which are the data visualization tools your organization typically uses? (choose as many as you feel are appropriate)**

Tick all that apply.

□ Chart.js
□ Datawrapper
□ D3.js
□ FusionCharts
□ Highcharts
□ Qlikview
□ Plotly
□ Sisense
□ Tableau
□ In-house Software
□ I don't know
□ Other: ____________________

Skip to question 36.

Data Analytics

35. **What issues prevent your organization from processing and analyzing Big Data? (choose as many as you feel are appropriate)** *

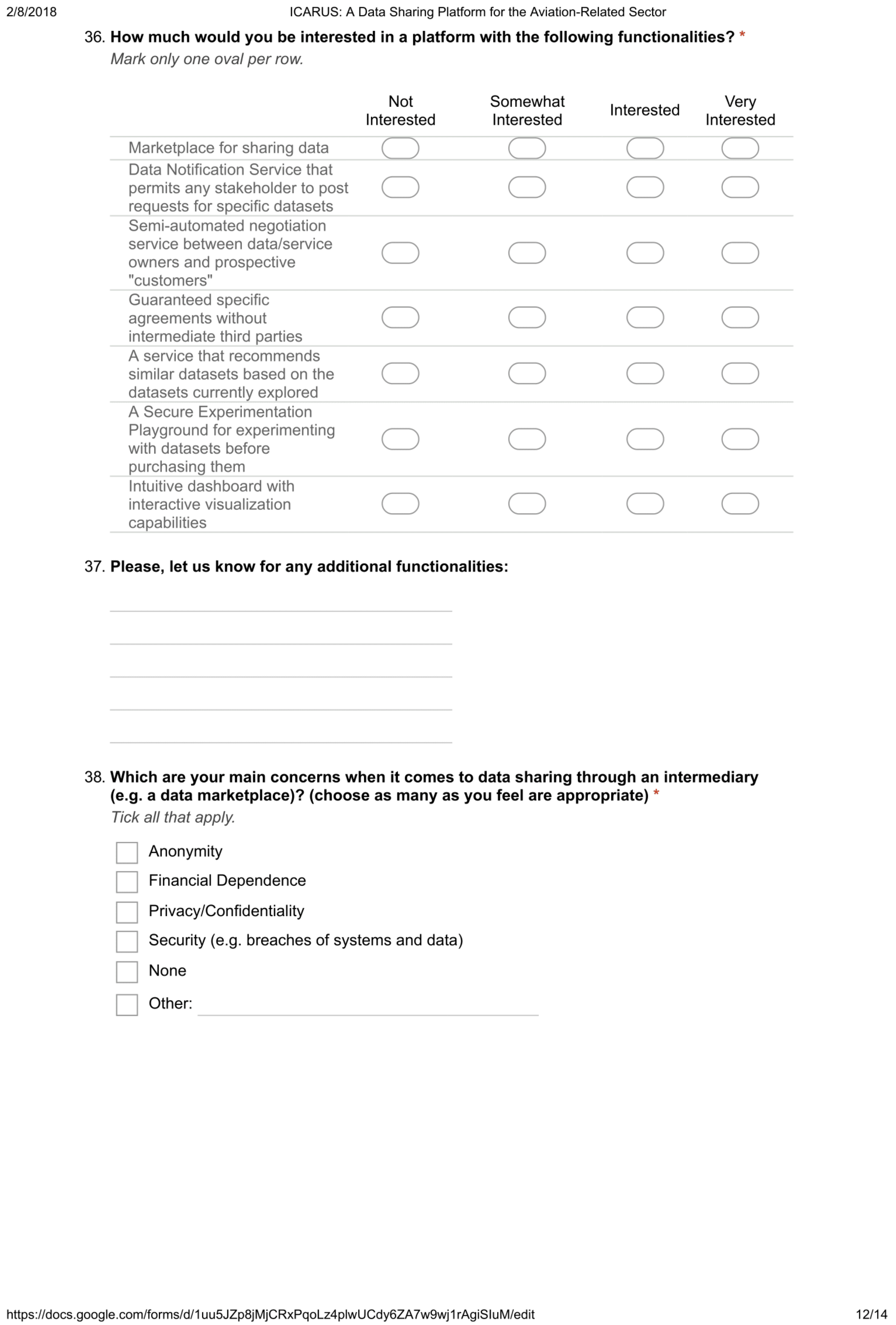
Tick all that apply.

□ Budget/Cost constraints
□ Lack of Experience
□ Lack of Resources
□ Lack of Time
□ Techinal issues
□ Not Interested
□ I don't know
□ Other: ____________________

Skip to question 36.

ICARUS platform

* Third Party: Someone who may be indirectly involved but is not a principal party to an arrangement, contract, deal or transaction (there is an intermediate party in between the primarily groups/parties involved).



36. **How much would you be interested in a platform with the following functionalities?** *

Mark only one oval per row.

	Not Interested	Somewhat Interested	Interested	Very Interested
Marketplace for sharing data	○	○	○	○
Data Notification Service that permits any stakeholder to post requests for specific datasets	○	○	○	○
Semi-automated negotiation service between data/service owners and prospective "customers"	○	○	○	○
Guaranteed specific agreements without intermediate third parties	○	○	○	○
A service that recommends similar datasets based on the datasets currently explored	○	○	○	○
A Secure Experimentation Playground for experimenting with datasets before purchasing them	○	○	○	○
Intuitive dashboard with interactive visualization capabilities	○	○	○	○

37. **Please, let us know for any additional functionalities:**

____

____

____

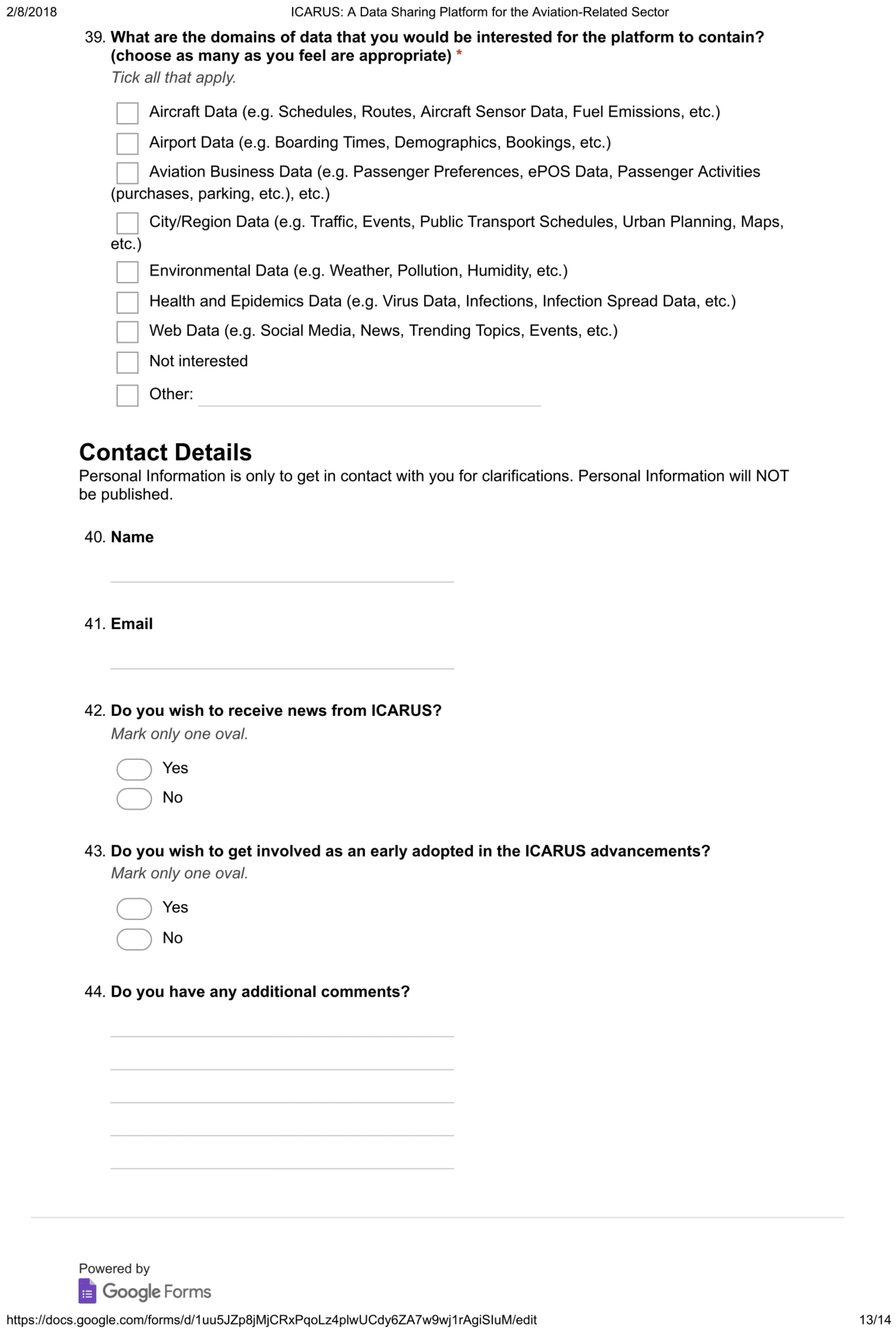
____

____

38. **Which are your main concerns when it comes to data sharing through an intermediary (e.g. a data marketplace)? (choose as many as you feel are appropriate)** *

Tick all that apply.

- □ Anonymity
- □ Financial Dependence
- □ Privacy/Confidentiality
- □ Security (e.g. breaches of systems and data)
- □ None
- □ Other: ____



2/8/2018 ICARUS: A Data Sharing Platform for the Aviation-Related Sector

39. **What are the domains of data that you would be interested for the platform to contain? (choose as many as you feel are appropriate)** *

Tick all that apply.

□ Aircraft Data (e.g. Schedules, Routes, Aircraft Sensor Data, Fuel Emissions, etc.)

□ Airport Data (e.g. Boarding Times, Demographics, Bookings, etc.)

□ Aviation Business Data (e.g. Passenger Preferences, ePOS Data, Passenger Activities (purchases, parking, etc.), etc.)

□ City/Region Data (e.g. Traffic, Events, Public Transport Schedules, Urban Planning, Maps, etc.)

□ Environmental Data (e.g. Weather, Pollution, Humidity, etc.)

□ Health and Epidemics Data (e.g. Virus Data, Infections, Infection Spread Data, etc.)

□ Web Data (e.g. Social Media, News, Trending Topics, Events, etc.)

□ Not interested

□ Other:

Contact Details

Personal Information is only to get in contact with you for clarifications. Personal Information will NOT be published.

40. **Name**

41. **Email**

42. **Do you wish to receive news from ICARUS?**

Mark only one oval.

○ Yes

○ No

43. **Do you wish to get involved as an early adopted in the ICARUS advancements?**

Mark only one oval.

○ Yes

○ No

44. **Do you have any additional comments?**

Powered by

Google Forms

https://docs.google.com/forms/d/1uu5JZp8jMjCRxPqoLz4plwUCdy6ZA7w9wj1rAgiSIuM/edit 13/14