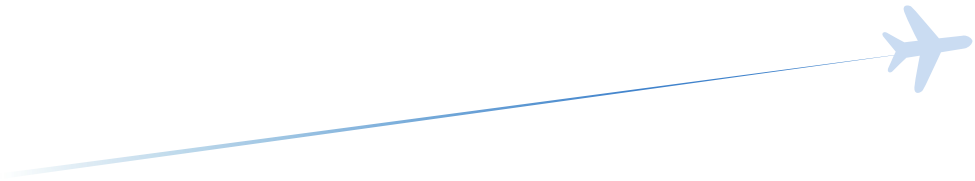

Data Management Plan

D1.1

DART

Grant:	**699299**
Call:	**H2020-SESAR-2015-1**
Topic:	**SESAR-02-2015 Data Science in ATM**
Consortium coordinator:	**University of Piraeus Research Center (UPRC)**
Edition date:	**27 October 2016**
Edition:	**00.01.03**

Founding Members

EUROPEAN UNION EUROCONTROL

SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Jose Manuel Cordero/CRIDA	Project Member	27/10/16
Miguel García Martínez/CRIDA	Project Member	27/10/16

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
George Vouros/UPRC	Project Coordinator	27/10/16
David Scarlatti/BR&T-E	Project Member	27/10/16
George Fuchs/FHFR	Project Member	27/10/16

Approved for submission to the SJU By — Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
George Vouros/UPRC	Project Coordinator	27/10/16
David Scarlatti/BR&T-E	Project Member	27/10/16
Georg Fuchs/FHFR	Project Member	27/10/16
Jose Manuel Cordero/CRIDA	Project Member	27/10/16

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date

Document History

Edition	Date	Status	Author	Justification
0.1	21/07/16	Draft	J.M. Cordero	New Document
0.5	24/07/16	Draft	J.M.Cordero	Data sources description
0.8	26/07/16	Draft	J.M.Cordero/M.Garcia	Sections completion
1.0	27/07/16	Final	J.M.Cordero/M.Garcia	Reviewer comments inclusion. Version for official submission
1.01	25/08/16	Final	J.M. Cordero	Inclusion of changes according to SJU assessment. Version for official submission

1.02	12/09/16	Final	J.M. Cordero	Inclusion of changes according to SJU 2nd round assessment. Version for official submission
1.03	27/10/16	Final	G.Vouros/J.M.Cordero	Inclusion of clarifications according to the meeting with SJU

DART

DATA-DRIVEN AIRCRAFT TRAJECTORY PREDICTION RESEARCH

This document is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 699299 under European Union's Horizon 2020 research and innovation programme.

Abstract

This document describes the Data Management Plan (DMP) of DART project according to the guidelines described in the Guidelines on Data Management in Horizon 2020 document [4]. As such, this DMP describes the data management life cycle for all datasets to be collected, processed or generated by DART project during its research activities. It details all data the project will collect and generate, how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved.[1]

[1] The opinions expressed herein reflect the author's view only. Under no circumstances shall the SESAR Joint Undertaking be responsible for any use that may be made of the information contained herein.

Table of Contents

Executive Summary

The DART Data Management Plan (DMP) details all the data that the project will collect and/or generate, how it will be exploited and made accessible to all project members, how and what data sets will be made available for verification and re-use, and how it will be curated and preserved. In order for the document to be self-contained, and according to the Guidelines on Data Management in Horizon 2020 recommendations, a comprehensive view of the data lifecycle is provided with appropriate definitions of terms being used and assumption under which the plan has been devised, with a succinct description of stakeholders' groups.

This DMP provides information on the data sources to be used and alternative data sources that may be used throughout the project. Although this list may be revised, there is not an update of the DMP foreseen during the project lifecycle. Updates of the data sources (if any) will be reflected in the iterative process for datasets provision, as described in WP1, and clarified in section 2 of this document.

In the scope of the DMP, datasets are considered to be collections of data that are generated, constructed, or provided to DART for achieving its objectives, and may be devised by exploiting data from a single or from multiple data sources. Each dataset has a specific spatial and temporal coverage, and may contain specific features per data source.

DMP describes the data sources to be potentially exploited at high-level, also specifying information on already known metadata per data sources (if it exists), scale of data, modality, and provision methods. Data sources will be described in detail (with the potential inclusion of additional data sources, if necessary) in the document accompanying every dataset to be generated, constructed or provided (this is further detailed in section 2), and are outside the scope this DMP.

This deliverable also specifies a generic policy /methodology for associating data sources being used and datasets devised with metadata, and a specific initial policy for version control. The specific aspects related to data integration for every dataset iteration, addressing the temporal and geographical coherence of each dataset will be described in the document linked to every dataset iteration. These documents are seen as early iterations and parts of the final D1.3 DART Data Pool deliverable, that will comprise of all the datasets iterations during the project lifecycle. These documents will be made available for SJU every time a new dataset is released, and finally in the form of D1.3 deliverable.

Furthermore, this DMP provides information on data sharing, also taking into account limitations of partners, IPR and legal issues, and licensing. No ethical issues have been identified for DART.

Finally, this deliverable concludes with issues concerning archiving and preservation.

1 Introduction

1.1 Purpose of the document

This document describes the data management processes that DART will implement in order to achieve its research objectives, ensuring both the achievement of the project goals in terms of data usability and availability, and in terms of research data quality, sharing and security.

The Data Management Plan (DMP) covers how data will be handled within the project frame, during the research and development phase, but also details the intentions for the archiving and availability of the data once the project has been completed.

DMP describes the overall context of DART research which concerns the pre-tactical decision stage, and in particular DCB processes at local/subregional level. Then, DMP describes data sources to be potentially exploited, well aligned in time and space, specifying information on already known metadata per data sources, scale of data, modality, provision methods, geographical and temporal coverage. Data sources will be described in detail, with the potential inclusion of additional data sources (e.g. EAD has been identified as a potential source and efforts will be devoted to gain access to that data), if necessary.

This deliverable also specifies a generic policy /methodology for associating data sources being used and datasets devised with metadata, and a specific initial policy for version control.

Furthermore, this DMP provides information on data sharing, also taking into account limitations of partners, IPR and legal issues, and licensing. No ethical issues have been identified for DART.

Finally, this deliverable concludes with issues concerning archiving and preservation.

As the project evolves, and as the understanding of needs from different data sources become more concrete, changes in the data situations will be reflected in the iterative process for devising datasets, also described in section 2 of this deliverable.

1.2 Intended readership

This document is intended to be used by DART members.

1.3 Acronyms and Terminology

Term	Definition
ANS	Air Navigation Service
ANSP	Air Navigation Service Provider
ATM	Air Traffic Management
ATC	Air Traffic Control

AU	Airspace User
CFS	Certificate on the Financial Statements
DCB	Demand and Capacity Balancing
DMP	Data Management Plan
GA	General Assembly
H	Humans
Horizon 2020	EU Research and Innovation programme implementing the Innovation Union, a Europe 2020 flagship initiative aimed at securing Europe's global competitiveness.
IPR	Intellectual Property Rights
KPI	Key Performance Indicator
PMP	Project Management Plan
POPD	Protection of Personal Data
PRC	Performance Review Commission
TRL	Technology Readiness Level
SESAR	Single European Sky ATM Research Programme
SJU	SESAR Joint Undertaking (Agency of the European Commission)
SJU Work Programme	The programme which addresses all activities of the SESAR Joint Undertaking Agency.
SESAR Programme	The programme which defines the Research and Development activities and Projects for the SJU.
WBS	Work Breakdown Structure
WP	Work Package

Table 1: Acronyms and Terminology

1.4 Project Introduction

DART responds to the first Call for Proposals of SESAR Exploratory Research projects launched under Part III `Societal Challenges´ of the Horizon 2020 Research Framework Programme (H2020-SESAR-2015-1). DART addresses the Research Topic 02-2015: *Data Science in ATM* and in particular the need to improve understanding on the suitability of applying big data techniques for predicting multiple correlated aircraft trajectories based on data driven models and accounting for ATM network complexity effects.

DART will deliver understanding on the suitability of applying big data techniques for predicting multiple correlated aircraft trajectories based on data-driven models and accounting for ATM network complexity effects.

Data Science is being pervasively applied to many businesses today: it is easy to find many success stories in which the application of modern big-data technologies has made possible approaches out of the mind of anybody a few years ago. It has even make room for provoking proposals like the now

famous form Anderson, Chris: "The end of theory: The Data Deluge Makes the Scientific Method Obsolete." Wired magazine 16.7 (2008): 16-07.

However, it is also well known that many initial promising data-driven solutions have turned out to fail: This, for instance, was the case for the Google Flu project. DART aims to avoid this hype and present to the ATM community an understanding on what can be achieved today in trajectory prediction using data-driven techniques. It is expected that data-driven techniques at least help to improve the performance and accuracy of predictions by complementing classical, model-based prediction approaches. Improved predictions will enable advanced collaborative decision making processes, which finally will lead to more efficient ATM procedures.

DART will blend computer science non-ATM specific expertise from University of Piraeus Research Center (UPRC, Greece) and Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (FRHF, Germany) and ATM state of the art and business needs knowledge from Centro de Referencia de Investigación, Desarrollo e Innovación (CRIDA, Spain) and Boeing Research and Technology Europe (BR&T-E, Spain) in a two years long effort.

Under this prism, the data management plan will identify the data sources to be exploited in order to achieve the project objectives, as well as will identify the DART data life cycle.

The DART Data Management Plan (DMP) details what data the project will collect, generate, how will be exploited or made accessible for verification and re-use, and how it will be curated and preserved.

This deliverable will thus refer to these issues in detail and will seek ways to make a large portion of the data available with respect to partners' agreements, with respect to IPR or privacy laws. No ethical issues have been identified for DART.

1.5 Relation to other Work Packages and Deliverables

This deliverable is related to WP1, WP2, and WP3.

WP1 (Data Management) will create useful, reusable and high quality datasets of recorded aircraft trajectory information, in an iterative process during the project lifecycle; these datasets are themselves a valuable outcome of the project, since the difficulties existing in any data-driven efforts to start with a high quality and well documented dataset from operational source are well known. These datasets will be used both by WP2 and WP3. The datasets will comprise raw data obtained from the ATM surveillance infrastructure and synthetic data generated offline thanks to the exploitation of the initial set of raw data. An important aspect is that the datasets will be accompanied by detailed data and metadata descriptions, when available, and a visual tool that will ease inspection, evaluation and exploitation of each dataset by data scientists. WP1 is responsible for DMP.

WP2 (Single Trajectory Prediction) is devoted to the analysis and evaluation of a wide range of data-driven techniques to the aircraft trajectory prediction problem. The most promising prediction algorithms obtained in WP2 will be evaluated in order to get knowledge about the pros- and cons- of their extensive usage.

WP3 (Collaborative Trajectory Prediction) is devoted to unveil the complexity to be considered in a trajectory prediction due to the influence of the surrounding traffic. Relying on the individual predictions provided by WP2, the application of reinforcement learning algorithms in an agent-based trajectory prediction framework will be studied in order to obtain improvement predictions thanks to the consideration of ATM network effects.

1.6 Approach Taken and Structure of the Deliverable

This DMP has been devised having the following issues in mind:

DMP describes the data sources to be exploited in DART (those initially intended, that might be completed or removed in different dataset iterations), the management of datasets to be constructed, and the lifecycle of data in DART.

As already pointed out, although a dataset may be any coherent set of data, we use this term to denote any set of data from (individual or multiple) data sources, or data created from DART components. These datasets can be used for validation, reusability, dissemination or demonstration purposes. Data sources provide input data to the DART components for realizing their functionality and computing results.

As DART partners address their research objectives in an iterative way, partially driven by the results observed and the quality and adequateness of the datasets provided by WP1, they may decide incorporating data sources not already known or taken onto account, enhance some of the existing ones, or remove them from the datasets. The provided datasets will be continuously refined according to results and needs expressed in terms of requirements.

This iterative process is not intended as an update of this DMP, which has a different scope. However, if subsequent versions of this DMP are considered necessary within DART (according to substantial changes that are considered convenient to be reflected), this deliverable will be updated.

The current version of the plan concerns the main DART data sources according to the ATM expert stakeholders in the project (BR&T-E and CRIDA) criteria, as a result of their experience with ATM data usage and exploitation, thus putting in place hypotheses on the convenience of the data to be provided.

Data sources provided from DART partners or from third parties satisfy specific properties and are subject to constraints and limitations for access and exploitation and should be used with respect to specific IPR and legal restrictions. These, have been specified on the DART Consortium Agreement Signed by partners and may be refined throughout the project.

Specific preparation and curation techniques per data source are also considered to be incorporated into a specific document that references each dataset provision iteration, according to the data lifecycle described in section 2, ensuring coherency.

DART aims for the development of research components for advanced analytics and prediction / forecasting of trajectories. These, in conjunction to the data management components to be developed in WP1 will be rigorously tested and results will be disseminated using specific datasets, which will be agreed among DART partners. These datasets together with research results produced by components in WP2 and WP3 may be made available to the research community via widely-used data repositories, a dedicated Transaction Pipeline to be designed by WP1, or eventually via the DART store.

All data sources being used, datasets, or research results being produced and archived will be described using a specific meta-data schema. The schema/vocabulary to be used will be chosen among well-known alternatives, also detailed below.

Therefore DART DMP must take into account three important issues, distinguishing different needs of data management: (a) Preparation and curation of data sources, (b) Data acquisition, integration and provision via the Transaction Pipeline to be designed and developed, and (c) orthogonally to (a) and (b), preparation of datasets, as well as archiving created results, to be used for validation, reuse and disseminating DART research results.

As the project develops the DMP will be refined by taking into account the data management techniques and the DART data management infrastructure developed, and the datasets to be used and provided, with the consent of all partners.

DMP scope is not intended to include detailed data sources specification, although an overview of the data sources considered for the initial dataset are included for reference. In the same way, DMP scope does not include the scenario definition, addressed by further deliverables, although for clarity and better understanding a research context description has been included.

Finally the construction of distinct datasets and their iterative refinements are not within the scope of DMP, but rather within the scope of D1.3 DART Data Pool. Every dataset, according to the DART data cycle described in section 2, is accompanied with a document detailing the included data, the data sources exploited, as well as the integration criteria to ensure alignment of data sources both geographically and temporarily. This is a key aspect for adequate data-driven learning. This DMP contains some notes and generalities on this topic, although specific aspects will be described in every particular dataset from the iterative data provision process.

The rest of this deliverable is structured as follows:

Data life cycle - section 2 presents the DART data life cycle, and relevant stakeholders.

Basic data information - section 3 provides description of the basic information about the data sources whose use in the DART project has been decided in the first place. Datasets will be described in detail every time that a high-quality dataset is released. These descriptions will be part of the deliverable D1.3, DART Data Pool ,that is intended in an iterative way, from T0+3 to T0+16, according to the workplan.

Metadata - Each data source and dataset will be described with meta-data. These meta-data can be used on the one hand for automating the DART data ingestion, data reusability, but on the other hand for all stakeholders to have a concrete view of the published data. This is further described in Section 4.

Access, sharing and re-use policies – Data sources are associated with limitations for access, legal and IPR constraints. This presents challenges if datasets have to be shared with stakeholders that are not beneficiaries. An important challenge is the integration/interlinking of data from sources having different usage and access policies. Interlinking data with certain constraints and requirements with

data that are publicly and freely available impacts the technological and methodological approaches in order to implement the desired access policy. Section 5 outlines this further.

2 DART DATA LIFECYCLE

2.1 The DART data value chain

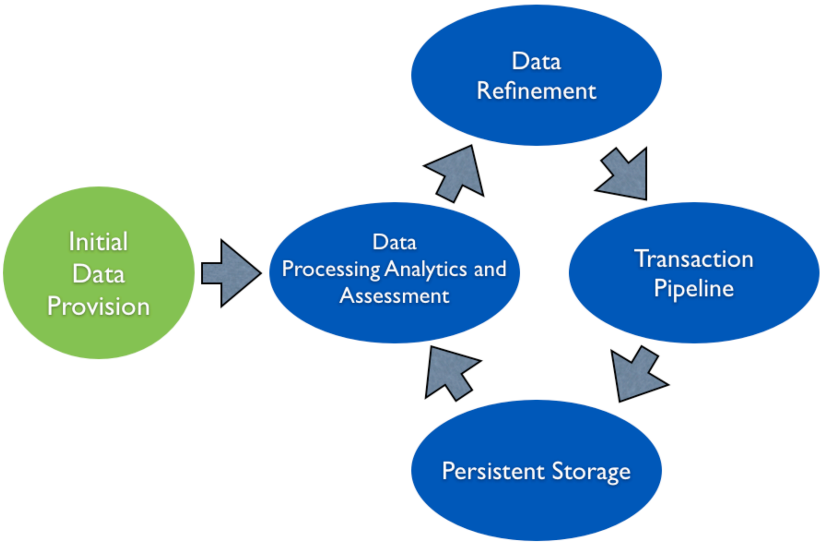

Figure 1. DART data value chain

The DART data value chain (Figure 1) comprises the following stages:

Initial data preparation. WP1 will provide an initial, early dataset according to the DART consortium expertise in ATM data (CRIDA, used to deal and manage ATM operation data, is leading this WP and is responsible for dataset provision), including data from all the necessary data sources, coherently combined in time and space. If early requirements coming from WP2 and WP3 are available on time, they will be considered in this initial dataset generation. This initial dataset will include the precise definition of the data sources to be included. Specifically, multiple, heterogeneous and disperse data sources are expected to be used. Curation and integration guidelines will be included.

This dataset is generated locally, by in-situ processing within WP1 leaders premises and systems. No streaming data provision is foreseen in this stage.

Due to the early stage at which it needs to be delivered, this initial dataset will be provided by means that may differ from the final Transaction Pipeline (that has not yet be designed at this stage).

Data processing, analytics and assessing. The data analytics components include trajectory prediction for single and multiple trajectories. These consume the data provided by the data management component.

WP2 and WP3, as well as the Visual Analytics task within WP1, will process the dataset(s) provided by WP1 with their own developed components and modules, including analytics components for trajectory prediction and visual analytics. Further data curation techniques and integration of data sources – so as data to be aligned in time and space- may be performed at this stage towards the provision of high-quality datasets, while data sources may be filtered (disregarding existing data) to facilitate data exploitation. Thereby, by means of an experimental use of provided datasets, requirements for dataset improvement will be returned to WP1 from WP2 and WP3, in terms of data quality assessment, inclusion of different information (if available), or suggesting different or none curation or integration. The feedback to WP1 dataset provision will be formulated in terms of requirements, recommendations, and suggestions.

WP1, WP2, and WP3 might consider a diversity of techniques in this stage. The quantitative and/or qualitative results of the experiments will be provided and contribute to data quality assessment.

Data refinement. With the outcome of the previous stage, WP1 will try to enhance, refine and improve, both in terms of data provision and data curation/integration, the provided dataset(s). This will be an iterative process involving the previous stage as well as this one, to be repeated periodically (ideally every 3-4 months, what will allow a minimum of three iterations).

Every one of these datasets provided in an iterative way from T0+3 to T0+16 (including initial data) will be accompanied by a reference document that specifically defines the dataset scope, the specific data sources exploited, and the temporal and geographical linkage features to ensure coherence. These documents are not formal deliverables, but they can be seen as early iterations of the final D1.3 DART Data Pool deliverable, that will comprise of all the datasets iterations during the project lifecycle. All of these reference documents will be made available for SJU together with availability notes every time a new dataset is provided. This will allow visibility of the iterative process and feedback in every iteration.

DART Transaction Pipeline. WP1 will design and implement a Transaction Pipeline that will allow secure, efficient and reliable access to these refined datasets. While it will not be available in the early dataset provisions, it will be deployed in a for next iterations, as soon as possible, in order to maximize data flow throughput and facilitating experimentation with data and further data refinement iterations.

DART persistence storage. DART will develop a scalable data store to manage data from disparate sources, providing coherent view on integrated data, ready for efficient utilization by potential higher-level offline and real-time capabilities, to be developed in the project.

The relation of the DART work packages to the stages of data value chain is described on Figure 2, starting from the left of the figure. WP1 will provide detailed information about available data sources as well as information about preparation, curation and accessibility of these data sources. WP1 will also describe the data that will be stored in the DART integrated system store while it will present ways to access integrated data through the Transaction Pipeline. WP2 and WP3 will provide description of analytics results generated and datasets used for validating their computations, and will also provide feedback to WP1 for dataset refinement. In this role, WP1 comprises the Visual Analytics task within WP1. Results generated may also need to be integrated with data in the store and be stored in the DART store.

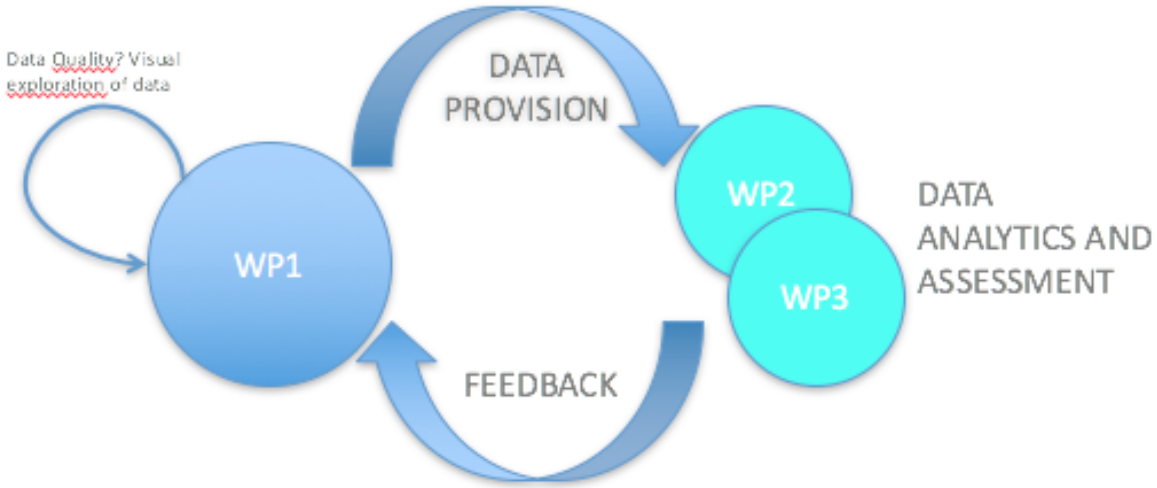

Figure 2. Relation of data value chain to DART workpackages

2.2 Stakeholders and Constraints

Key stakeholders that have been identified and that influence the data management plan decisions as they have specific interest in data in sources, research results, domain specific results (i.e. results connected to aviation operational concerns) are as follows:

Data providers. These represent the organizations which provide the data sources to be integrated into the DART store. These may be DART beneficiaries, or third parties connected to beneficiaries, or organizations providing open data. These parties are represented in the DART project by CRIDA and BR&T-E mainly, as actors within Aviation domain. These DART partners play significant roles, specifying the data sources to be used and assist with the definition of requirements from the data management components, while gathering requirements on data from WP2 and WP3 and providing coherent, high-quality datasets according to agreed requirements.

DART partners. The DART partners collaborate to specifying the requirements for data management and data analytics according to the domain-specific use cases defined and the project research objectives. The quality and the big-data characteristics of the data provided from disparate sources are important features for them, given also that their developments must be validated and evaluated in different scenarios of data growth and quality.

End users. End-users represent user-interest groups that focus on real-life, industrial and user-defined challenges concerning operations regarding moving entities in air. These are represented in the DART project by BR&T-E and CRIDA, including also organizations and domain experts.

Big data analytics researchers and stakeholders of the big data value chain. These are interested in DART developments and research results: Algorithms and novel methods for data management and big data analytics. These are mainly interested to publications produced and seek datasets for validating, evaluating, comparing, and testing analytics methods. Devising such datasets and making them available to the research community is expected to increase the visibility of DART results.

3 THE DART DATA SOURCES

This section presents the initial selection of different ATM data sources to be exploited. According to the requirements raised by WP2 and WP3, new data sources can be included if available (every dataset delivery includes a specific dataset description).

A fine, detailed description of the data in these sources will be provided in the document accompanying each devised dataset jointly to every iteration and provision of the high quality datasets, but here, for the purposes of the DMP, we enlist data sources with a succinct description of their contents, emphasizing mostly to the data providers (i.e the origin of data), how data are acquired, when and where are these acquired, how often, also specifying the contact points for accessing data sources.

All data sources and datasets are to be available during the lifetime of the project (of course open data sources are expected to be available for a longer period) together with the accompanying documentation, as described in section 2.1. These descriptions will be also included in deliverable D1.3, DART Data Pool due on M20.

3.1 TARGET RESEARCH CONTEXT

While scenario definition is not within the scope of this deliverable, is relevant to understand the target operational concept where this research fits for the choice of data sources, even in the case of initial datasets (as stated above, this section covers an initial selection of data sources identified by ATM expertise within WP1). This section contains a high-level description of this research context that will be later refined in a detailed scenario by WP3 scenario definition deliverable scenario in M06.

DART results are considered of application within the European Network. However, within a Continuous Layered Planning concept, they are expected to be potentially applied at different levels. It is a target of DART to ensure that the methods and results obtained are of application for stakeholders, not limited to a particular region.

In particular, due to the expertise available in the project, the ATM operational scenarios targeted are Local/Subregional DCB processes during their planning phase. From an operational level the role of the ANSP at Local/FAB level will be represented by CRIDA (supported by its relationship with EnAire, the Spanish ANSP represented by CRIDA), providing the necessary expertise both in the needs to be addressed and in the data management. The role of the Airspace users (AUs) will be represented by BR&T-E, with a global view of the AU interests (independent of their particular business model).

Thus, from a wide perspective, the target research context regarding the deployment of DART results is a local/subregional DCB process performed at strategic/pre-tactical planning phases, where not

only flow management measures (being Network Manager responsible for them) but also Capacity management and trajectory negotiation processes (in a TBO environment) are possible.

In this context, predictability KPA is expected to be significantly improved by the results provided by DART, affecting other operational KPAs such as Capacity and Efficiency.
This scenario justification is reinforced by the inclusion of Airports, due to the fact that Enaire also manages ATC processes in Airports.

For this reason, operational data from Spanish airspace will be considered (being of the highest possible quality, as a result of operational quality assurance). The project will ensure that no bias appears due to the fact that Spanish airspace data is used, allowing that the conclusions and methods obtained are of application in any other ANSP (even at Network level) within European area, as well as any other target stakeholder such as airlines.

3.2 ATM DATA SOURCES

Aeronautical data is heavily regulated, especially in Europe according to Eurocontrol Standards. For example, Flight Plan filling information follows ICAO FPL2012 format, Radar information is provided following ASTERIX standard (Asterix Cat62 for fused data), datalink between airlines dispatcher and aircraft follows A702-A format, Airspace information is mostly provided in AIXM format. That means that the research results can be applied nationwide in Europe.

Reference datasets candidates to be used (initially considered) as input on investigation will be linked amongst them to ensure coherent geographical and temporal alignment.

The initial dataset will comprise data from the following data sources:.

3.2.1 Weather Data:

Multiple sources provide weather data to Air traffic systems like satellite, met radar and the aircraft itself.

METAR: METAR is a format for reporting weather information.

METARs typically come from airports or permanent weather observation stations. Reports are generated once an hour or half-hour, but if conditions change significantly, a report known as a special (SPECI) may be issued. Some METARs are encoded by automated airport weather stations located at airports, military bases, and other sites. Some locations still use augmented observations, which are recorded by digital sensors, encoded via software, and then reviewed by certified weather observers or forecasters prior to being transmitted. Observations may also be acquired and reported by trained observers or forecasters who manually observe and encode their observations prior to transmission.

Raw METAR is the most common format in the world for the transmission of observational weather data. It is highly standardized through the International Civil Aviation Organization (ICAO), which allows it to be understood throughout most of the world.

METAR information includes RVR, dew point, visibility and surface winds.

Metadata:

N/A

Scale:

Structured data (6 csvs): 3 MB per year and per aerodrome)

Raw data (1 pdf): 20 KB per day and per aerodrome)

Provision method(s):

Structured data: Files for specific time periods and aerodromes can be assembled and delivered (*.csv) .

Raw data: Files (*.pdf) by day and aerodrome can be delivered

Available temporal and geographical coverage:

METAR is available for Spanish airspace, from 2012 to 2016.

Partner(s) responsible: CRIDA & BR&T-E

NOAA: This data source is used mainly to obtain the weather conditions at the position an aircraft is at any given time of the flight.

Weather models use a Grid with a specific resolution. In DART project we'll work with NCEP Grid 4 which has a resolution of 0.5º. (see http://www.nco.ncep.noaa.gov/pmb/docs/on388/tableb.html)

Forecast models can be run several times a day, in DART we'll typically use the latest forecast available previous to the time we are interested in. Forecast models has too a time resolution, or "forecast step", which we expect to be 1 hour.

Metadata :

Data for weather models is typically distributed in "GRIB" format files. GRIB (**GRIdded Binary** or **General Regularly-distributed Information in Binary form**) format allows to compress a lot the weather data and includes metadata about the content of the file, so it is very convenient for transferring the data. The data can be extracted with many available tools (I.e. GRIB API from ECMW available at https://software.ecmwf.int/wiki/display/GRIB/Home).

Scale:

As a reference a global forecast, 6 hours step, for 24h, for 14 isobaric levels at .5º resolution expanded from .grib to .csv can amount about 3.2 Gb.

Provision method(s):

The forecast is available in native/raw format in .grb files. These files may be converted to .csv files using tools like "wgrib2".

Available temporal and geographical coverage:

NOAA is available for the whole European airspace, from 2014 to 2016.

Partner(s) responsible: CRIDA & BR&T-E

SIGMET: is a weather advisory that contains meteorological information concerning the safety of all aircraft. This information is usually broadcast on the ATIS at ATC facilities, as well as over VOLMET stations. They are assigned analphabetic designator from N through Y (excluding S and T). SIGMETs are issued as needed, and are valid up to four hours. SIGMETS for hurricanes and volcanic ash outside the CONUS are valid up to six hours.

Metadata :

N/A

Scale:

Structured data (csvs): 10 MB per year and per aerodrome

Raw data (pdf): 100 KB per day and per aerodrome

Provision method(s):

The files is available in native/raw format in .pdf files. These files may be converted to .csv files.

Available temporal and geographical coverage:

SIGMET is available for the Spanish airspace, from 2012 to 2016.

Partner(s) responsible: CRIDA

TAF: terminal aerodrome forecast (TAF) is a format for reporting weather forecast information. TAFs are issued every six hours for major civil airfields: 0000,0600,1200,1800 UTC, and generally apply to a 24- or 30-hour period, and an area within approximately five statute miles (or 5NM in Canada) from the center of an airport runway complex. TAFS are issued every 3 hours for military airfields and some civil airfields and cover a period ranging from 3 hours to 24 hours.

TAFs complement and use similar encoding to METAR reports. They are produced by a human forecaster based on the ground. For this reason there are considerably fewer TAF locations than there are airports for which METARs are available. TAFs can be more accurate than Numerical Weather Forecasts, since they take into account local, small-scale, geographic effects.

Metadata :

N/A

Scale:

Structured data (csvs): 20 MB per year and per aerodrome

Raw data (pdf): 200 KB per day and per aerodrome

Provision method(s):

The files is available in native/raw format in .pdf files. These files may be converted to .csv files.

Available temporal and geographical coverage:

TAF is available for the Spanish airspace, from 2012 to 2016.

Partner(s) responsible: CRIDA

3.2.2 Radar (Surveillance)

IFS: This data source provides radar tracks of the Spanish airspace controlled by the Spanish ATC provider EnAire. A radar track file consists on tabular data rows with a timestamp key and several rows of geospatial information for each one of these timestamps. The update interval is 5 seconds. The area provided is separated into 5 different regions delivered each one on a different plain text file (ifs files).

Metadata:

On each IFS file some metadata is listed at the beginning and the end of the files. First lines describe some system information without relevance for the rows, indicating the period contained on the file. In fact, this information may be useless too because each day is processed on different files so all files contains the info of a single day that can be discovered by the name of the file.

The raw data files are named with the first letter according to the region and the date on format yyMMdd. On the structured version of the data, each row contains a column with the source file providing traceability to the files.

Scale:

IFS data is available for from 2013 till 2016 (complete years). The covered area is the Spanish airspace. On raw version, one day of tracks is around the 0.5 GB on raw compressed files.

On structured version, one day is around a hundreds millions of records.

Provision method:

Raw version for limited range can be provided through Internet (FTP) or physical media (I.e. DVD).

Structured version could be provided by web services or database direct connection.

Available temporal and geographical coverage:

Available for the Spanish airspace, from 2012 to 2016.

Partner(s) responsible: CRIDA

ADSB: This data source refers to the ADS-B messages broadcasted by many airplanes (practically all airliners) using their transponders. These messages are received by ground based receivers and can

be used to reconstruct the trajectory of the flight. There are several types of messages that can be found but the relevant ones are these about aircraft identification and position.

DART source of ADS-B messages is the ADSBHub network. This network is formed by 81 stations across the globe, 61 of them in continental Europe. The messages received by this network are stored in a human readable format know as "SBS-1 BaseStation port 30003".

Metadata:

N/A

Scale:

ADSB data is available since late 2015 and is continuously recorded, however, not all the sensors are 100% of the time up and the recording system is not 100% of the time up. Power supply and/or network outages can create dates with less or even without data.

One day of messages tops about 2.3 Gb (not compressed).

Provision method(s):

Historical data: Files for specific time periods can be assembled and delivered through Internet or physical media (I.e. DVD).

Real time data: Once a VPN connection is established to BR&T-E Laboratory Network the client can issue a netcat command to receive the real time feed.

Available temporal and geographical coverage:

Available for parts of the European airspace, from 2015 to 2016.

Partner(s) responsible: CRIDA & BR&T-E

3.2.3 Airspace

Sector Configuration: Airspace can be divided in a set of ways, with a different number of pieces (sectors). A sector configuration 9A means that a particular airspace (a region in Spain) is divided in 9 sectors, in a particular way. 9B also mean 9 sectors, but divided in a different way. Typically, due to low traffic at nights, the configuration set at those times is a 1A, meaning that a single sector (thus, a single controller) is in place.

This leads to the fact that configurations available are fixed, but configuration “in place” varies during day, adapting capacity resources (Air Traffic Controllers, mainly, as more sectors open mean more capacity, but also more controllers) to the expected demand.

Metadata:

N/A

Scale:

Configuration: Structured data (csv): 4 KB (1 day)

Sectorisation: Structured data (csv): 37MB (1 period of validity)

Provision method(s):

Structured data: Files (*.csv) for specific time periods and aerodromes can be assembled and delivered.

Available temporal and geographical coverage:

Available for Spanish airspace, from 2012 to 2016.

Partner(s) responsible: CRIDA & BR&T-E

3.2.4 Synthetic trajectories

This data source represents trajectories generated by a Trajectory Predictor (TP). A TP is a software/routine that is included in any software or tool that needs to forecast the future state of the aircraft to perform its tasks. Depending on the particular application that the TP serves, the level of detail (i.e., number of variables and number of aircraft states) that needs to be included in the aircraft state may vary. The trajectory generated by flight management system (FMS) contains multiple aircraft states (e.g., an aircraft state at least each 30 seconds) and each state multiple variables, such as latitude, longitude, altitude, time, calibrated airspeed or mass. These variables are used by other FMS subsystems to generate guidance modes, or monitor the aircraft performances.

The format of a synthetic trajectory depends on the particular TP model and software design and implementation of that model.

DART source of synthetic trajectories messages is a stand-alone model-based TP engine developed by BR&T-E (BR&T-E TPE) to generate trajectories for a set of given input information (flight plan, weather, aircraft model, operational context). For a given particular flight, BR&T-E TPE can be used to generate different alternative synthetic trajectories representing all the possible conditions that the flight may encounter of for what if analysis. Under the aviation use case, BR&T-E TPE will be used to generate reference synthetic trajectories that will serve as benchmark.

In DART we provide the format generated by a stand-alone TP engine developed by BR&T-E (BR&T-E TPE) to generate trajectories.

Metadata :

N/A

Scale:

Synthetic trajectory data will be available under demand for the particular scenario that is going to be studied. In principle, the range of available dates should coincide with the range of available flight plans and surveillance. Ideally, for a particular data set of flight plans and/or corresponding surveillance data, should be a set containing n synthetic trajectories per flight, where n is driven by the particular use case scenario that want to be studied.

One synthetic trajectory for one aircraft containing 300 aircraft states, each of them with 56 different variables would be around 1 MB (XML format; text file of near 20000 lines). This size can be easily reduced either by decreasing the number of variables and/or the sample rate (number of aircraft states)

Provision method(s) :

Historical data: Files processed for specific time periods can be assembled and delivered through Internet or physical media (I.e. DVD).

Available temporal and geographical coverage:

N/A, as is a synthetic data source

Partner(s) responsible: BR&T-E

3.2.5 Flight Plans

Flight Plan is a relevant category as contains the information that triggers a lot of operational decision, both in planning and execution phase, and both on the Air Navigation Service Provision side, and in the Airline one. However, in this initial estimation of data sources is conceived as an input for the synthetic trajectory generation. It will in any case be available so that any other uses of this information are possible.

The Flight Plan is the specified information provided to air traffic services units, relative to an intended flight or portion of a flight of an aircraft.

This data will be extracted from the operational Spanish system, including flight plan updates (all) if necessary.

Metadata

This data source has no metadata information.

Scale

About 1 GB per day.

Provision method

Daily files for Spanish airspace.

Available temporal and geographical coverage:

This data is available for Spanish airspace affected flights, from 2011 onwards.

Partner(s) responsible: CRIDA.

3.3 ALTERNATIVE ATM DATA SOURCES

As a risk mitigation action, the identification of alternative data sources has been addressed. The following table defines, for every data category considered, a primary data source and a secondary (backup) data source. Additionally, an estimation of the likelihood of having a problem with the data sources is included.

Unless a problem is identified with the primary data sources, the alternative ones will not be used. However, in some data categories (i.e., weather), the most convenient primary data source needs to evaluated and analysed according to the project needs.

Data category	Primary Data source	Alternative Data source	Risk estimation
Weather	*NOAA**	*TAF*	Low
	*METAR**	*NOAA*	Low
	*SIGMET**	*NOAA*	Low
	*TAF**	*NOAA*	
Radar	IFS	ADS-B	Low
Airspace	Spanish Operational data (as defined in 3.1.3)	DDR-2	Low
Synthetic Trajectories	BR&T-E Trajectory Predictor using Spanish operational Flight Plans as input	BR&T-E Trajectory Predictor using DDR Flight Plans as input	Low
Flight Plans	Spanish Operational data	DDR-2	Low

*Data source to be finally used as primary needs to be determined

In particular, the synthetic trajectories are provided by the BR&T-E Trajectory Predictor, which processes operational information (flight plans), thus the data source availability is referred to the input of the trajectory prediction tool.

3.4 COMBINING ATM DATA SOURCES

Alignment of the different data sources ensures common geographical and temporal coverage, which is paramount for datasets usage and effective data-driven learning.

The data sources will be combined usually using an ad-hoc reference to ensure that they will refer to the same time and space, as well as to link them when necessary (for instance, radar tracks with flight plan for a particular flight). The specific linkage criteria will depend on the data sources composing the dataset, as well as their features, so it will be particular, ad-hoc, for each dataset iteration.

In particular, the integration of data sources will be done in WP1, where CRIDA is experienced in combining data from disparate data sources, ensuring a temporal and spatial reference.

Typically UTC time is the main reference for temporal alignment, using or correcting the different data sources to fit it. Regarding spatial alignment, coordinates are usually the best cross index. Combined indexes using flight callsign, date, time and aircraft type are usually used. The particular combination method, however, will depend on the specific dataset (and the different data sources it originates from).

This alignment is not seen as a major challenge for most sources, as in terms of Performance Monitoring and Post-operation analysis is nowadays usual within CRIDA. Main challenge is in terms of aligning subjective phenomena (such as those described in SIGMET, related to sectors), with quantitative measures of NOAA grid.

The specific ad-hoc temporal and geographical coverage alignment criteria will be included in every dataset delivery.

4 STANDARDS AND METADATA

4.1 Metadata (Standards and Methodology for Capturing/Creating Metadata)

Data sources to be exploited, as described in previous sections, may be accompanied with metadata descriptions. However there is a diversity of the metadata being used, either at the syntactic or at the semantic levels.

There are two ways to overcome this issue, imposing a coherent and integrated view on data sources provided to the project and to datasets that may be devised by the project (e.g. for evaluation, validation, demonstration etc purposes), together with datasets of results created:

First, an implicit, but quite natural – based on the DART objectives- way: This is based to the fact that all data, from any data source, together with results created will be interlinked and be stored in the DART store through the data transaction pipeline to be defined. Thus, the DART data management components are integral parts of the DMP, since in principle all stakeholders may access the data store and fetch data that populate a coherent schema according to specific spatiotemporal constraints and constraints on other qualities.

Second, an explicit one, where any coherent subset of data throughout the DART data chain is provided either as a file in a specific form, or via an API. Such a subset may concern integrated views of data stored in the DART store, or subset of data in its raw form (i.e. in the form provided by data source providers) or being processed in a specific way (e.g. for anonimization, or for getting just a subset of the variables provided, or for bringing it in a specific/required form), or even data created by any of the DART research components. In this case, each such dataset will be described by metadata properties provided in a separate file, specifying among others the downloadable file, and/or the API for getting this dataset. For this purpose we will use a widely used schema such as DCAT or r3data.org.

These options are yet to be analysed and are considered as a part of the **Transaction Pipeline definition and description, deliverable D1.2**.

In any case, DART will use any of these options but it will not create any schema specific repository for datasets: This is beyond the purposes of the project. However, metadata of datasets will be in separate files in standard forms, so that they can be used by any such data repository, facilitating the discoverability of datasets by third parties and their management within DART.

DCAT [5] is a widely used standard that is RDF designed to facilitate interoperability between data catalogs published on the Web. An application profile of the W3C standard DCAT called DCAT-AP is used within Europe. With this standard dataset description in Europe can be exchanged in a coherent and harmonized context.

A dataset in DCAT is defined as a "collection of data, published or curated by a single agent, and available for access or download in one or more formats".

A dataset does not have to be available as a downloadable file. For example, a dataset that is available via an API can be defined as an instance of dcat:Dataset and the API can be defined as an instance of dcat:Distribution. DCAT itself does not define properties specific to APIs description.

An example of metadata description based on DCAT is given in Figure 3

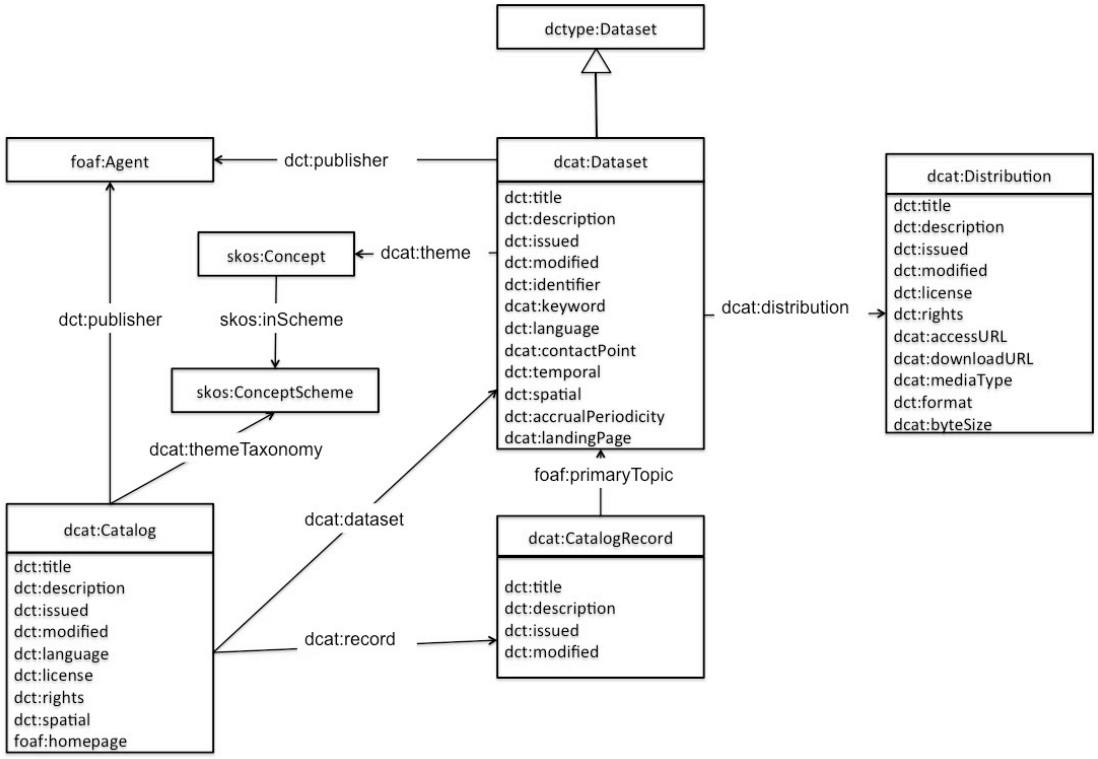

Figure 3. DCAT example

DCAT provides properties for datasets version control (e.g. via dct: issued and dct:modified), spatiotemporal characteristics of data sets (i.e. area dct:spatial and period dct:temporal covered), as well as licensing via the dct:licence property for each dataset distribution.

Although there are several tools for DCAT, these do not seem to be suitable for extracting metadata descriptions from DART datasets. This imposes a workload to creating metadata that we aim to keep minimum, although clearly to the level of quality that allows sharing datasets and increasing data discoverability.

As datasets format and structure is dynamic in this project due to the iterative process, all WPs will need to be aware of changes between consecutive datasets.

4.2 Naming conventions and organization

As mentioned, datasets to be shared, will be stored in a file repository, together with their metadata descriptions which will be provided in a separate file with a link to the dataset file, as specified in the schema to be used.

Metadata files will be organized according to (a) the datasets provided or combinations of these, (b) the geographical coverage, (c) the temporal coverage, (d) their origin and (e) date issued, in corresponding folder structures according these characteristics in the order specified above.

The name of each dataset and metadata file will be a concatenation of these characteristics.

5 DATA SHARING

A number of datasets will be used during DART. Also the research components in WP2 and WP3 may produce an amount of research results.

DART has to consider the possibility to share datasets with respect to the legal and IPR constraints. Specifically, we have to consider where, how, and to whom the data could be made available.

Datasets will be shared either by opening a specific API for third parties to fetch datasets from the DART store, or via making such datasets downloadable as single files. This issue will be covered in D1.2, Transaction Pipeline description.

In any case, the methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data.

Regarding IPR and legal restrictions, these have already specified in the background section of the DART consortium agreement:

Describe Background	Specific limitations and/or conditions for implementation (Article 25.2 Grant Agreement)	Specific limitations and/or conditions for exploitation (Article 25.3 Grant Agreement)
European Air Traffic Data (Flight Plans, Airspace information, Airspace use plan, AIXM + ADR, operational airspace data, AIP's, routes, notams, restrictions, regulations)	BR&T-E can use European Air Traffic Data in DART activities. The transfer of data to other consortium members would require prior approval from the third party data owner.	BR&T-E has to explicitly mention the third party data owner copyright of the data in any dissemination involving data publishing.
Weather data	BR&T-E can use public domain and third party weather data in DART activities. The transfer of the third party data to other consortium members would requires prior approval from the third party data owner.	Public data can be disseminated stating that such material is not subject to copyright protection Third party data can be disseminated following license rules (i.e. not modifying it and with the right attribution)
ADSB messages collected by receivers in Europe	There are no commercial restrictions on how to use this data. Everybody can publish the data for free or to use it for commercial purposes.	Receiver network may require credits on the published dissemination.

CRIDA is a research centre, not a data-generator, which entrusted to use, exploit and maintain the datasets generated by ENAIRE, owner of that data. CRIDA is entitled to use the data for research purposes such as the DART project (where it acts as data provider), and can share them with the consortium only for research purposes.

The following general conditions for accessing and using the data provided by CRIDA apply:

- The Dataset provided will never be the raw data obtained from operational sources, but the result of the fusion and processing performed by CRIDA.
- The Dataset can only be used to achieve the research purposes stated in the DART Technical Annex. No other use is allowed.
- The user is not allowed to create or derive new datasets from the original one.
- The Dataset will be anonymized through its "delocalization" (either in place or time) except in specific cases where the specific confidentiality agreement allows data concerning places and time to appear. Delocalization will be done to disable its traceability to operational events and situations. The user might not change or process the Dataset in any way to remove the anonymity of the data.
- The Dataset will be stored physically in the premises of CRIDA. No physical copy of the Dataset will be provided. The users will not store, copy or otherwise move the Dataset (physically or logically) to databases or systems outside the premises of CRIDA.
- Access to the Dataset will be granted as Needed by the research members of the consortium either for implementing their own tasks under the DART action or for exploiting their own results.
- Access to the Dataset will be granted under a specific confidentiality agreement that will clearly identify the Dataset users, purpose, usage timeframe and any specific clauses that might be needed.
- The Dataset access confidentiality agreement will be signed before the start of the work to set the conditions of this access. CRIDA is committed to provide access in a secure way, royalty-free, in order to achieve the research goals of the project in an efficient way.
- Access to the Dataset will be provided only for the specific purposes and scheduled time windows described in each one of the user confidentiality agreements.
- For security and confidentiality reasons no permanent access to the Dataset will be granted.
- The Dataset will be available through a secure channel as specified by CRIDA.
- Subsets of the Dataset may be stored, copied or otherwise be moved (physically or logically) to databases or systems outside the premises of CRIDA, as Needed by the research members of the consortium either for implementing their own tasks under the DART action or for exploiting their own results. These subsets will be treated as "confidential" and will be provided only under a specific confidentiality agreement that will clearly identify the characteristics of the subset, users, purpose, usage timeframe and any specific clauses that might be needed. This subset will be labelled specifically for the purpose Needed.
-

The above mentioned restrictions apply to the following data sources:

Spanish ATC Platform On-line Flight Plan Data

Including Creation, update and deletion messages for the flight plans in Spanish airspace

Dataset containing information from 2013-2015 (3 years)

Spanish ATC Platform Off-line Flight Plan Data

Including relevant flight messages for all the flights in Spanish airspace (Flight plan creation, deletion and major updates, sector entry, sector leave,…)

Dataset containing information from 2009-2015 (7 years), with focus on 2013-2015

Spanish ATC Radar Data

Including actual radar tracks for all the flights in Spanish airspace

Dataset containing information from 2009-2015 (7 years), with focus on 2013-2015

Spanish Sector configurations

Including actual sector configuration put in place for all the Spanish airspace

Dataset containing information from 2009-2015 (7 years), with focus on 2013-2015

To limit restrictions, DART partners will aim to gain the consent of data providers in limited datasets to be shared, maybe for specific periods and under specific licenses.

6 ARCHIVING AND PRESERVATION

Data sources concerning moving entities' behaviour in the air, are growing in size due to market and service demands. DART will exploit these data sources, together with other geospatial, environmental and weather data sources, and develop scalable methods for processing these data. There is a growing interest in archiving, sensing and performing analytics over mobility and behavioural data. In DART we aim to address issues concerning long term storage of these data as well as research generated data.

These issues related to safety and long-term data storage are complicated when the data evolves or/and they could be merged with other data coming from other sources. So, in this project we need to consider:

- What is the volume of the data to be maintained?
- What is considered long-term (2-3 years, 10 years, etc.)?
- Identification of archive for long-term preservation of data.
- Which datasets will need to be preserved in the archive?
- What about relevant dependent datasets?

Preserved datasets will need to be updated and this means a data preservation policy and process will need to be defined.

A central consideration for any long-term DMP is the cost of preserving that data and what will happen after the completion of the project. Preservation costs may be considerable depending on the exploitation of the project after its finalization. Examples include:

- Personnel time for data preparation, management, documentation, and preservation,
- Hardware and/or software needed for data management, backing up, security, documentation, and preservation,
- Costs associated with submitting the data to an archive,
- Costs of maintaining the physical backup copies (disks age and need to be replaced).

These topics will be discussed during the project. If any relevant conclusions or recommendations are reached, they will be included in potential updates of this DMP. Initially no DMP updates are foreseen to cover these topics, exclusively.

In particular, the Transaction Pipeline deliverable will address these specific topics. However, as a reference this deliverable reflects how data is archived and preserved in the premises of the two data providers of the project, BR&T-E and CRIDA:

Data stored in BR&T-E premises is protected using a VPN access with two-factor authentication. The data is stored with 3 factor replication. A secondary backup is in place. All access to servers require LDAP login. NDAs are signed by all users accessing to the Lab.

Data stored in CRIDA premises is currently accessed from the outside by using a VPN access with authentication. Retrieval of data is done via DB which requires additional LDAP authentication, and permissions are granted according to each profile allowing several levels of access to data. Data is also replicated by a factor of 2, ensuring no data loss happens. Additionally NDAs are signed by all the users with granted access.

References

[1] SJU, Project Execution Guidelines for SESAR 2020 Exploratory Research, Edition 01.00.00, 08/02/2016

[2] DART, DART 699299 Consortium Agreement, v1.0

[3] Participant Portal H2020, Online Manual http://ec.europa.eu/research/participants/docs/h2020-funding-guide/grants/grant-management/communication_en.htm

[4] European Comission, Guidelines on Data Management in Horizon 2020, Version 2.1, 15/02/2016

[5] Data Catalog Vocabulary, W3C Recommendation 16/01/2014 https://www.w3.org/TR/vocab-dcat/

[6] DART, DART 699299 Grant Agreement, v1.0