Grant Agreement No: 687591

Big Data Analytics for Time Critical Mobility Forecasting

datAcron

D8.3 datAcron Data Management Plan (2nd Version)

Deliverable Form	
Project Reference No.	H2020-ICT-2015 687591
Deliverable No.	8.3
Relevant Work Package:	WP 8
Nature:	R
Dissemination Level:	PU
Document version:	
Due Date:	15/07/2018
Date of latest revision:	19/07/2018
Completion Date:	19/07/2016
Lead partner:	UPRC
Authors:	UPRC, IMISG, BRTE, CRIDA, NARI, CMRE
Reviewers:	All partners
Document description:	The datAcron Data Management Plan (datAcron DMP) details what data the project collects, generates, how these can be exploited or are made accessible for verification and re-use, and how these data are curated and preserved. This is the 2nd version of this deliverable.
Document location:	Filestore : /datAcron/WP8/Deliverables/Final

History of changes

Version	Date	Changes	Author	Remarks
0.5	05.06	ToC	George Vouros	
0.8	29.06	Addition of details concerning data sources and datasets	George Vouros, Petros Petrou	
1.0	10.07	Incorporating comments and feedback from partners	George Vouros, Petros Petrou	
1.2	19.07	Incorporating comments and feedback from partners	George Vouros	Adding details about maritime and aviation dataset originators, preparation and curation after comments received from NARI and CMRE. Added references to documents and papers

EXECUTIVE SUMMARY

The datAcron Data Management Plan (datAcron DMP) details what data the project collects, generates, how these are exploited or made accessible to all stakeholders, how and what data sets are made available for verification and re-use, and how they are curated and preserved. In order for the document to be self-contained, a comprehensive view of the datAcron data lifecycle is provided with appropriate definitions of terms being used, and stated assumptions under which the plan has been devised, with a succinct description of stakeholders' groups: These were also part of the 1st version of the datAcron DMP.

The document describes data sources exploited in each of the datAcron domains, i.e. Aviation and Maritime, also specifying information on existing metadata per data sources, size of data sets, modality, and provision methods.

The descriptions of all the datasets in this document are excerpted from WP5 and WP6 Data Preparation deliverables (D5.2, D6.2 and D5.4, D6.4) and from documentation provided by datAcron partners (IMISG, NARI, CMRE, BRTE and CRIDA) proving data and guiding the data exploitation and use during the execution of the project.

Furthermore, information and a generic policy on data sharing is provided, also taking into

account, limitations of partners, IPR and legal issues, licensing and ethical issues.

The deliverable concludes with issues concerning archiving and preservation that should be addressed once the project has been completed.

This is the 2nd version of this deliverable and reflects changes in the data sources and datasets exploited in datAcron, as these have been described in the 1st version of the Data Management Plan. Indeed, as the project evolves, the DMP has been updated to reflect the changes in the data sources exploited in datAcron, while details on datasets are provided. The document identifies which data sets are made available to stakeholders, datasets provision methods, associated metadata and licensing schemes (from providers to datAcron and from datAcron to stakeholders). In addition, policies and methodologies to be used for making datasets available are updated.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS & TERMS

Abbreviation	Term
ATM	Air Traffic Management
BRTE TPE	BRTE Trajectory Predictor Engine
CTOT	Calculated Time of Take-off
DCAT	Data catalogue vocabulary
DMP	Data Management Plan
FMS	Flight Management
NOAA	National Oceanic and Atmospheric Administration
TP	Trajectory Predictor
GA	Grant Agreement 687591
DDR	Demand Data Repository
AIXM	Aeronautical Information Exchange Model
AIS	Automatic Identification System
AIP	Aeronautical Information Publication
RAD	Route Availability Document

1. INTRODUCTION

1.1 Purpose and Scope

A Data Management Plan (DMP) is a formal project document which outlines the handling of the data sources at the different project stages. The H2020 guidelines [3] provide an outline that has to be addressed. The DMP covers how data will be handled within a 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. As the project evolves, the DMP is updated to reflect the changes in the data sources and datasets exploited, as the understanding of data sources and data created becomes more concrete.

datAcron project aims to develop novel methods for (a) **real-time detection and prediction of trajectories** and (b) **detection and prediction of important events** related to moving entities, together with (c) **advanced visual analytics methods**, over multiple heterogeneous, voluminous, fluctuating, and noisy data streams from moving entities, via the (d) **real-time in-situ processing of multiple data streams**, (e) the provision of **integrated views of streaming data with archival data** expressing entities' characteristics, geographical information, patterns of mobility in specific areas, regulations, intentional data (e.g. planned routes) etc., and (f) the provision of advanced solutions for **managing spatio-temporal data**.

Thus, data sources related to datAcron purposes include streaming data sources (mostly concerning surveillance data for moving entities), and archival data sources with historical, contextual and other data concerning the entities themselves, their positioning and movement. These have been identified in deliverables D5.2 "Maritime data preparation and curation (interim)" and D6.2 "Aviation data preparation and curation (interim)" and updated in deliverables D5.4 "Maritime data preparation and curation (final)" and D6.4 "Aviation data preparation and curation (final)"

Technological developments in datAcron are validated and evaluated in user-defined challenges that aim at increasing the safety, efficiency and economy of operations concerning moving entities in the air-traffic management (ATM) and maritime domains: These use case, scenarios and challenges have been specified in deliverables D5.1 "Maritime use case detailed definition" and D6.1 "Aviation use case detailed definition", while evaluation and validation methodologies and criteria have been specified in D5.3 "Maritime experiments specification" and D6.3 "Aviation experiments specification".

Under this prism, the data management plan (DMP) identifies the data sources to be exploited for the purposes of the use cases detailed in D5.1 and D6.1, towards validating and evaluating datAcron methods and tools and the datAcron integrated system prototype, in the Maritime and Aviation domains.

A dataset is any coherent subset of data, but in this document we use this term to denote any set of data from (individual or multiple) data sources, and/or data created from datAcron components.

These datasets can be used for validation, reusability, dissemination or demonstration purposes, also according to GA articles 29.2 "Open access to scientific publications" and 29.3 "Open access to research data". *Data sources* provide input data to the datAcron components for realizing their functionality and computing their results, while datasets are subsets of data that are necessary and sufficient enough for validation, evaluation activities, and for re-use and demonstration of computations.

The datAcron Data Management Plan (DMP) details what data the project collects, generates, how these are exploited or made accessible for verification and re-use, and how are curated and preserved in the lifetime of the project. The datAcron Data Management Plan refers to these issues in detail and specifies ways to make a portion of the data available with respect to partners' agreements, with respect to IPR or privacy laws, and ethical issues specified in D8.5 "Ethics Management Plan".

The current version of the plan concerns the datAcron data sources that fit (are necessary) to the purposes of the use cases and scenarios detailed in D5.1 and D6.1.

1.2 Relation to other Work Packages and Deliverables

This deliverable is related to WP1, WP5 and WP6. WP1 provides the detailed description of how data are to be integrated and stored in the datAcron store, providing coherent, integrated views on data; as well as ways to process, integrate data, and access integrated data. WP5 and WP6 aim to identify the data sources that should be used within the project and prepare datasets so that they are readily available for exploitation according to the datAcron value chain detailed in the next chapter of this document. More specifically DMP leverages D5.2 and D6.2, as well as to their corresponding updates D5.4 and D6.4, since many of the questions identified in DMP need to be answered as part of the data sources' preparation and curation from data providers. Furthermore, this deliverable is related to D8.5 "Ethics Management Plan", since it should consider licensing, IPR and ethical issues.

1.3 Approach Taken and Structure of the Deliverable

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

- As datAcron partners address their research objectives, driven by the use cases and scenarios already detailed in D5.1 and D6.1, in the Maritime and Aviation domains respectively, they have already concluded on data sources that complement the set of sources identified in the first version of the DMP. These sources are mentioned in this version of the DMP. This deliverable is written concurrently with deliverables D5.4 and D6.4, which, with their interim versions D5.2 and D6.2, provide the precise definition of the datasets used, including results of datasets quality assessments, as well as data curation techniques to guarantee continuous data consistency and availability. As such, these four deliverables are connected to this deliverable and descriptions of datasets have been excerpted from them in providing succinct information on datasets in this 2nd version of the DMP.

- As WPs 1,2,3 aim to produce advanced data management components for integrating and managing data from heterogeneous and disparate data sources, as well as providing trajectories and events detected/predicted/forecasted. These components are considered as constituents of the overall data management plan. Thus, they are incorporated into the data management plan, positioning them appropriately into the datAcron data value chain.

- Data sources provided from datAcron 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 ethical, IPR and legal restrictions. Such restrictions have been specified in the datAcron Consortium Agreement signed by partners and may be refined/revisited for particular subsets of data throughout the project.

- datAcron aims to the development of research components for advanced analytics and prediction / forecasting of trajectories and events. These, in conjunction to the data management components developed are rigorously tested, and results can be disseminated using specific datasets, which are agreed among datAcron partners. datAcron has a specific policy for making any of these datasets and computed results available to the research community via a specific widely-used repository.

- All data sources being used, datasets, and research results being produced and archived, are described in a specific way according to a metadata schema, specified in this document.

Therefore datAcron DMP takes into account three important issues, distinguishing different needs of data management: (a) Preparation and curation of data sources by maritime and aviation data providers, (b) Data acquisition, integration and provision of data via components integrated in the datAcron integrated system prototype, and (c) orthogonally to (a) and (b), preparation of specific datasets, as well as archiving created results, to be used for validation, reuse and dissemination of datAcron research results, also according to GA articles 29.2 "Open access to scientific publications" and 29.3 "Open access to research data".

Therefore, DMP has been refined by taking into account the data management techniques and the datAcron infrastructure developed, the data sources and datasets to be used and provided, with the consent of all partners.

The remainder of this deliverable is structured as follows:

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

- **Basic data information** - section 3 provides description of the data sources and datasets used within the datAcron project.

- **Metadata -** Metadata used for data ingestion, data reusability, but on the other hand for all stakeholders to have a concrete view of the data being used, independently on restrictions to access the 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 datAcron beneficiaries. An important challenge is the integration/interlinking of data from datasets having different usage and access policies. Interlinking data with certain constraints and requirements with data that are publicly and freely available, impacts the desired access policy. Section 5 addresses these issues.

- **Archiving and preservation** – section 6 describes the challenging issues that arise regarding the long-term storage of data after project completion.

2. THE datAcron DATA LIFECYCLE.

2.1 The datAcron data value chain

Figure 1. datAcron **data value chain**

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

Data preparation. Use case (WPs 5 & 6) leaders provide the precise definition of the data sources to be used for research, evaluation and validation of research components, together with associated metadata, where they do exist, and/or detailed data descriptions, and detailed information about the origin, scale and provision method(s) per data source. Specifically, multiple, heterogeneous and disparate data sources are expected to be used in maritime and aviation use cases.

In situ-processing and data transformation / integration. The data sources are multiple streaming, as well as archival data sources. The datAcron integrated system incorporates in-situ processing methods that aim to transform data to RDF according to the datAcron ontology, provide data synopses and detect low level events, also integrating where appropriate data-in-motion from streaming sources and archived data, and integrating multiple streams. Requirements for these processes include low-latency and high rates of data compression, without affecting the quality of analytics results.

The data transformation components aim to covert data from (a) low level events' detector, (b) single and multiple streaming data synopses, (c) archival data sources, and (d) results computed by the datAcron analytics components, to RDF triples according to the datAcron ontology. Bringing all data to such a common form aims to facilitate their integration/interlinking, also incorporating semantic information into the process.

The data integration component interlinks data from disparate sources using spatio-temporal link discovery techniques. This produces integrated data views on spatio-temporal data. Integrated data are provided to the analytics components, while they are also stored in the datAcron store.

Persistence storage. datAcron develops a scalable data store to manage data from disparate sources, providing a coherent view on integrated data, ready for efficient utilization by the analytics components developed in the project.

Specifically, datAcron develops a novel data processing framework for supporting the efficient distributed management and querying of RDF spatio-temporal data from disparate data sources (also concerning trajectories and events computed by datAcron analytics components), emphasizing on eager filtering of data using spatial and temporal predicates.

Data processing and analytics. The data analytics components include trajectory and complex event recognition and forecasting, as well as visual analytics. These consume the data provided by the data management component and compute results, as described in D1.7 and D1.12 "Integrated Prototype".

The relation of the datAcron work packages to the stages of data value chain is described in Figure 2, starting the chain from the bottom of the figure. WP5 and WP6 provide detailed information about available data sources (aviation and maritime data), as well as information about preparation, curation and accessibility of data sources. WP1 provides an enriched stream of data-in-motion to analytics components, as well as integrated data stored in the datAcron store, while it presents ways to access that data. WP2, WP3 and WP4 provide description of analytics results generated and relevant datasets for validating, re-producing and disseminating their results, also according to GA articles 29.2 and 29.3 and limitations for data access, legal and IPR constraints. Computed results (e.g. trajectories and events detected/predicted) can also be integrated with data and stored in the datAcron store.
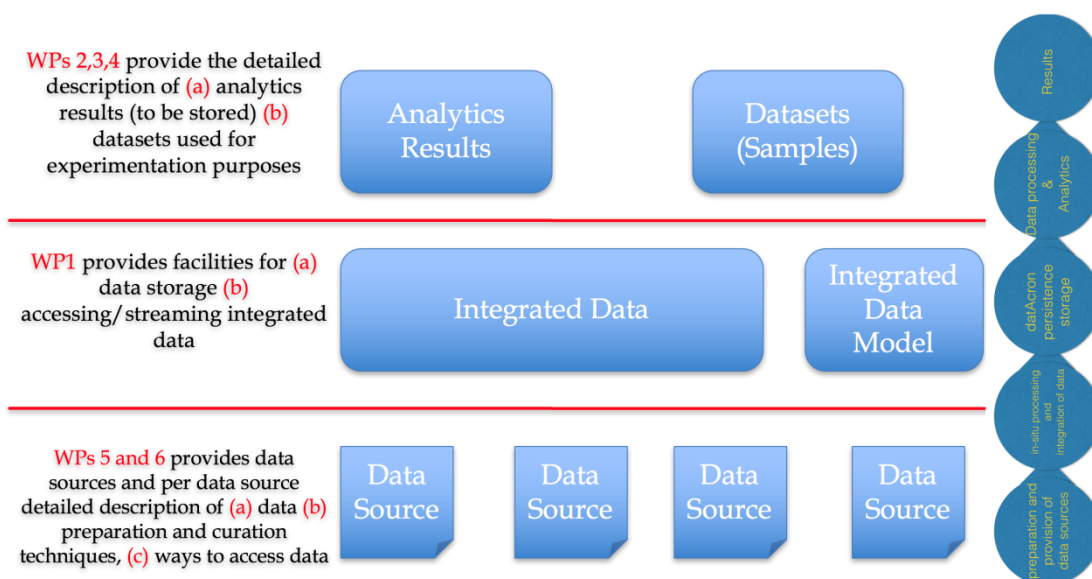

Figure 2. Relation of data value chain to datAcron work packages

It must be pointed out that WP1,2,3 and WP4 provide feedback to WP5 and WP6 concerning the data provided and their suitability for achieving their purposes. This have resulted to updating the data sources, reconsidering the quality of data gathered and/or provided to the project, and/or adding data sources.

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, research results, domain specific results (i.e. results connected to aviation/maritime operational concerns) are as follows:

Data providers. These represent the organizations that provide the data sources to be integrated into the datAcron store. These are datAcron beneficiaries, third parties connected to beneficiaries or to the project via the domain use case interest groups, or organizations providing (open) data related to the purposes of the maritime and aviation use cases. These parties are represented in the datAcron project by WP5 and WP6 leaders and participating partners. datAcron partners playing this role also play significant roles in defining the use cases per domain, in specifying the data sources to be used, in guiding other datAcron partners in how the provided data may be exploited to address the domain use cases, and in assisting with the definition of requirements from the data management components.

datAcron **partners.** The datAcron partners provide data sources and define the use of data sources and datasets within the project, collaborate to design and develop the architecture of the datAcron integrated system, specify the requirements for data management and data analytics according to the domain-specific use cases defined and according to the project research objectives. The quality and the big-data characteristics of the data provided from disparate sources are important features, given also that the methods and algorithms developed in datAcron must be validated and evaluated in different scenarios of data growth and quality.

End-users. End-users represent the use cases interest groups that focus on real-life, industrial and user-defined challenges concerning operations regarding moving entities in sea and air. These are represented in the datAcron project by WP5 and WP6 leaders, including also organizations and domain experts being members of the datAcron domain-specific use-case interest groups.

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

3. THE datAcron DATA SOURCES.

This section presents data sources to be exploited per domain of interest: aviation and maritime. It further provides details on existing metadata per data source (i.e. metadata concerning the content of data sources, mostly according to domain – e.g. whether-standards), scale of data sources and data provision methods.

A fine description of maritime and aviation data sources is provided in D5.4 [15] and D6.4 [16] (and their interim versions D5.2 [14] and D6.2) data preparation deliverables, but here, for the purposes of the DMP, we enlist data sources per domain with a succinct description of their contents, excerpted from these WP5 and WP6 data preparation deliverables, emphasizing mostly to the data providers (i.e. the origin of data), how data are acquired, when and where are these acquired and how often, also specifying the contact points for the data sources.

All data sources are to be available during the lifetime of the project. Open data sources, such as the heterogeneous maritime dataset published by WP5 in Zenodo [6,13] for reuse for training and research also outside the scope of the project, are expected to be available for a longer period. The data documentation available to understand the available data has been provided by datAcron deliverables D5.1, D5.2, D5.4, D6.1 and D6.2, D6.4, and in additional documents prepared by WPs 5 & 6 to facilitate the use of the data by other WPs.

3.1 Aviation Data Sources

Data sources are structured in eight different categories: Weather, Radar, Airspace, Network Manager, Synthetic Trajectories, Aircraft Identification, Flight Plan and Context Information [16].

A scheme of data sources organization in this document is as follows:

Weather	NOAA
Radar	IFS
	ADSB
	DDR
	FlightAware
Airspace	DDR
Network Management	CFMU
Synthetic Trajectories	Synthetic Trajectories
Aircraft Identification	Aircraft Identification
Flight Plans	Network Manager
	DDR
Contextual Information	Network Manager

Weather Data

A single data source (NOAA), is finally considered for this category, as is the only European-wide weather data available.

NOAA (National Oceanic and Atmospheric Administration): This data source is used mainly to obtain the weather conditions at the position an aircraft is at any given time of the flight.

The most relevant variables are the Temperature, the Pressure, and the two horizontal components of the Wind Speed, u and v (a positive u component represents wind blowing **to** the East and a positive v component is wind to the North), since they affect the performance of the aircraft.

Weather models use a Grid with a specific resolution. For aviation in datAcron we 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, for aviation in datAcron we 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 [16].

Metadata:
Data for weather models is typically distributed in "GRIB" format files. GRIB (**GRIdded Binary** or **General Regularly-distributed Information in Binary form**) format enables the compression of the weather data and includes metadata about the content of the file. Thus 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 methods:
The forecasts have been delivered in .grb files to datAcron. These files can be converted to .csv files using tools like "wgrib2".

Provider: NOAA

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: CRIDA

Radar

This data category comprises of three sources of surveillance information: IFS, ADS-B and DDR [16].

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). The files are named with the first letter according to the region and the date on format yyMMdd.

Metadata: Not available.

Scale:

IFS data is available for from 2013 till 2016 (complete years). The covered area is the Spanish airspace, as it also shown in D6.4.
On structured version, one day is around a hundreds millions of records.

Provision methods:
Structured version could be provided by web services, database direct connection or CSV files.

Provider: CRIDA

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: CRIDA

ADSB Messages: 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 for datAcron the relevant ones are these about aircraft identification and position.
datAcron 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:
Messages are contained in a single line of CSV files. ADS-B messages are captured in CSV format. An excerpt of such messages, describing a particular flight, look as follows:

```
MSG,3,,,34324E,,2015/08/03,01:05:03.844,2015/08/03,01:05:07.058,30050,,,45.69032,5.54741,,,0,0,0,0
MSG,6,,,34324E,,2015/08/03,01:05:03.844,2015/08/03,01:05:07.058,30050,,,,,1021,0,0,0,0
MSG,4,,,34324E,,2015/08/03,01:05:03.744,2015/08/03,01:05:07.058,,428.0,51.0,,,-384,,,,
MSG,1,,,34324E,,2015/08/03,01:05:04.344,2015/08/03,01:05:07.058,IBE34CP,,,,,,,,,,
MSG,3,,,34324E,,2015/08/03,01:05:05.943,2015/08/03,01:05:15.337,30025,,,45.69296,5.55218,,,0,0,0,0
MSG,6,,,34324E,,2015/08/03,01:05:05.943,2015/08/03,01:05:15.337,30025,,,,,1021,0,0,0,0
MSG,4,,,34324E,,2015/08/03,01:05:10.643,2015/08/03,01:05:15.337,,426.0,51.0,,,-128,,,,
```

ADS-B messages are formatted according to the SBS Station schema original from Kinetic's SBS-1 & SBS-3 Mode-S. Further details are provided in D6.4 [16].

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 methods:
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&TE Laboratory Network the client can issue a netcat command to receive the real-time feed.

Provider: BRTE

License to datAcron: Public

dataAcron distribution licence: Public

Partner(s) responsible: BRTE

Flightaware Firehorse: This data source was added to the project to solve the coverage issues in some European areas by ADSBHub.
Firehose is a FlightAware API for receiving streaming flight positions (e.g., RADAR, ADS-B, Mode S multilateration (MLAT), datalink, etc) as well as flight status data (e.g., flight plans, departure/arrival messages, flight updates) and surface movement positions. Only access leves 1 are 2 are available for dataAcron project:
1. ADS-B: Worldwide
2. MLAT: Worldwide

Surface data is access level 7 and is not available for the project. This limits coverage around airports

Metadata:
Flightaware firehose data source comes in json files according to schema specified in detail in D6.4.

Scale:
The covered area is the whole world, and thus Europe is covered, as it also shown in D6.4.
On structured version, one day is about 17Gb (not compressed).

Provision methods:
Historical data: Files for specific time periods are assembled and delivered in json format.
Real time data: There is one connection available for the dataAcron project.

Provider: BRTE

License to dataAcron: Internal to dataAcron

dataAcron distribution licence: N/A

Partner(s) responsible: BRTE

DDR: In addition to the above mentioned sources, DDR, or Demand Data Repository (a European wide valuable data source) contains surveillance/radar data, which is embedded in the flight plan data. This data source, including the radar part, is described in the Flight Plan category subsequently, and is only mentioned here for reference as a key radar data source.

Airspace

This category comprises of one single data source, DDR (in particular, the sector configuration part of this repository that also appears in Radar and in Flight plan categories) [16].
Airspace data source can be seen from a dual perspective. On the one hand, describes the existing airspace organization, with no gaps or overlaps, and all the possible ways of combining volumes to generate different operational sector configurations, also with the associated sector capacities, or flights that a sector can manage in a period of time (this static data generically called "adaptation data"), and on the other hand, describes the schedule of effective sector configurations that have been effectively put in place in European airspace.

This Airspace data source category covers only the second of the two. The essential, first one, is covered in the Context Information Category (static data).

DDR Sector Configuration: Air traffic control (ATC) is a service provided by ground-based controllers who direct aircraft on the ground through controlled airspace. The primary purpose of ATC worldwide is to prevent collisions, organize and expedite the flow of air traffic, and provide information and other support for pilots.
Airspace can be divided in a set of ways, with a different number of pieces (sectors).
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: Not available.

Scale:
The range of spatial coverage is the European airspace, from 01/06/2011 to present. As a reference each AIRAC cycle is around 12MB.

Provision methods:
Raw data: Text plain files. ('.cfg', '.cos', '.ncap', '.spc', '.gsl/sls', '.gar/are').

Provider: EUROCONTROL

License to dotAcron: Internal to dotAcron

dotAcron distribution licence: N/A

Partner(s) responsible: CRIDA

Network Management

This category covers the data sources that contain Network Management (also known as Flow Management) information, thus the regulations put in place to ensure a proper Demand Capacity balance in a tactical way. There is a single source considered: CFMU, coming from the Network Management organization (Eurocontrol), thus covering European airspace [16].

CFMU: This data source provides list of flights and regulations that these flights may have. The source is separated in two files, one for flights and other for regulations. When a flight has a regulation, the code of the regulation applied is provided on the row. When a regulation is applied to a flight, a CTOT is set for the flight and a delay over this time without a window [-15;+5] minutes is recorded.

Metadata: Not available

Scale:
CFMU data is available for from 2013 till 2016 (complete years).
dotAcron exploit data concerning the European airspace regulations.
On structured version, one day of information is around 70 MBs on compressed files.

Provision methods:
Structured version could be provided by web services, database direct connection or CVS files.

Provider: CRIDA

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: CRIDA

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.

The datAcron source of synthetic trajectories messages is a stand-alone model-based TP engine developed by BRTE (BRTE TPE) to generate trajectories for a set of given input information (flight plan, weather, aircraft model, operational context). For a given particular flight, BRTE 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, BRTE TPE is used to generate reference synthetic trajectories that serve as benchmark [16].

In datAcron we provide in D6.4 [16] the format generated by a stand-alone TP engine developed by BRTE to generate trajectories.

Metadata: Not available.

Scale:

Synthetic trajectory data are 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 data. Ideally, for a particular data set of flight plans and/or corresponding surveillance data, there should be a set containing *n* synthetic trajectories per flight, where *n* is driven by the particular use case scenario that is 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 methods:

Historical data: Files processed for specific time periods can be assembled and delivered through Internet or physical media (i.e. DVD). The trajectory output format is an XML file containing several fields associated to the aircraft which is represented in the trajectory (**trio:aircraft_identifier_information**) and the trajectory itself (**op_solver:trajectory**). Within the trajectory field, there are several sub-elements which link the input information (intent information) with trajectory states, a reference time and a sequence of aircraft states

containing a certain set of variables. The following table describes each of these fields and subfields. Under the column XO, it is indicated if this element would be always included (X) or is optional (O) depending on the requisites over the synthetic trajectories (certain fields or subfields can be removed from the output trajectory to reduce size)

Provider: BRTE

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: BRTE

Aircraft Identification

This data source provides details concerning aircrafts in a trajectory. In ADSB sources the aircraft is identified by ICAO 24-bit address or (informally) Mode-S "hex code". The ICAO 24-bit address can be represented in three digital formats: hexadecimal, octal, and binary, and typically in ADSB sources is represented in hexadecimal. One of the most important data to obtain given the ICAO address is the model of the aircraft, or more specifically, the ICAO Type Designator, according to DOC 8643 [16].

Metadata:
A csv file is be used to distribute the list of known aircrafts for datAcron.
The file contains the following fields:
- icao character varying (6) - ICAO address in hexadecimal.
- regid character varying - Unique alphanumeric string issued by a National Aviation Authority to identify an aircraft.
- mdl character varying – Aircraft Type according to ICAO DOC 8643
- type character varying- Aircraft model
- operator character varying – last known operator

Scale:
The identification is available for most aircrafts in the trajectories. About 25Mb.

Provision methods:
A csv file is used to distribute the list of known aircrafts for datAcron.

Provider: BRTE

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: BRTE

Flight Plans

Flight Plan is an essential category of data sources as it 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.

Two sources of information are considered: Network manager and (again) DDR [16]. They have similar information but each one of them may be preferable for different Aviation scenarios.

Network Manager Flight Plans: 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.

Metadata:
Flights Plan data is compliant with ICAO 4444 Flight Plan 2012 and is a direct translation to XML format.
Detailed field explanation available at:
http://www.eurocontrol.int/sites/default/files/content/documents/nm/network-operations/HANDBOOK/ifps-users-manual-current.pdf

Scale:
The size of the XML file for one flight with all updates can be around 400Kb. A file that refers to a flight with a subset of changes can be about 150Kb.

Provision methods:
A xml file for every Flight Plan update and per flight is used to distribute Flight related information.

Provider: CRIDA

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: CRIDA

DDR Flight Plans: This dataset is focused on the historical traffic data stored in ALLFT+, which contains the flight plans information.
The format of the files is plain text, where each line contains all information of a single flight. Details on this source are provided in D6.4 [16].

Metadata: Not available

Scale: From 01/06/2011 to present. About 130 MB per day.

Provision methods: Daily CSV files for European airspace.

Provider: EUROCONTROL

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: CRIDA

Contextual information data sources

This contextual information category is the complementary one to the airspace data. This category contains a single data source and it includes purely static data, describing the operation environment: It describes the existing airspace organization, with no gaps or overlaps, and all the possible ways of combining volumes to generate different operational sector configurations, also with the associated sector capacities, or flights that a sector can manage in a period of time [16].

Network Manager Contextual Information: The Contextual Information provided at European level by Eurocontrol is intended to provide services related to the management and sharing of airspace data (e.g. airspaces, routes, aerodromes, etc.)
The airspace data consists of two types of information:

- Airspace Structure Information for retrieving up-to-date airspace data from the CACD database. The CACD database is the repository for the environment data (a.k.a. airspace data) used in the network management systems to perform Flight Planning and Flow Management. This data includes AIP (Aeronautical Information Publication) concepts (such as Routes, Points and Aerodromes), and non-AIP concepts (such as Flows, RAD (Route Availability Document) Restrictions and Traffic Volumes).
- Airspace Availability Information for querying and modifying the airspace availability information; this includes the Flexible Use of Airspace.

The Airspace services make use of AIXM 5.1/ADR-E (http://www.aixm.aero) types when possible (ADR-E stands for ADR Extension). The following main information areas are in the scope of AIXM:

- Aerodrome/Heliport including movement areas, services, facilities, etc.
- Airspace structures
- Organizations and units, including services
- Points and NavAids
- Procedures
- Routes
- Flying restrictions

Metadata:
A xml file for every AIXM feature or set of AIXM features update for is used to distribute context information.
The structured Airspace Information uses a subset of AIXM 5.1, and includes three tables about Airports, as also detailed in D6.4.

Provision methods:
XML files are distributed with current airports data.

Provider: EUROCONTROL, CRIDA & BRTE

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner(s) responsible: CRIDA & BRTE

3.2 Maritime data sources

The maritime use case should be supported by data sources including [14]:

- Automatic Identification System (AIS, www.navcen.uscg.gov/?pageName=AISmain) messages broadcasted by ships;
- Maritime regulations, specifying the legislation and the rules for navigation and fishing;
- Marine protected/closed areas, where fishing and sea traffic may be (temporarily) forbidden;
- Traffic separation schemes and Nautical charts, useful to define vessel routes;
- Vessel routes and Fishing areas estimated from historical traffic data;
- Registry data on vessels and ports;
- Records of past events, such as incidents and illegal activities reports;
- Meteorological and oceanographic data (METOC) on atmospheric and sea state conditions and currents;

An overview of maritime data sources is provided in the following table (extracted from D5.4 "Maritime Data Preparation and Curation (final)" [15]).

Type	Source	Provenance	Format	Spatial extent	Temporal extent	Volume	Velocity
Surveillance	Automatic Identification System	Naval Academy	Flat files (comma-separated values)	Western part of the Channel and France	6 months	19.680.743 messages (1.05 GB)	~76 messages per min (in average)
	Automatic Identification System	IMISG	Flat files (text)	Europe	1 month	81.722.110 messages (8.11 GB)	~1.830 messages per min (in average)
	Automatic Identification System	IMISG	Stream of decoded messages in JSON	Europe	-	~400 KB / min (in average)	~3.700 messages per min (in average)
Weather	Sea state	SHOM, IFREMER	Flat files (comma-separated values)	Western part of the channel and France	6 months	79.652.684 forecasts (3.02 GB)	1463 forecast files, 1 file / 3 hours
	Weather forecast	Met Office (United Kingdom) based on data provided by NOAA (United States)	Flat files (comma-separated values)	Western part of the channel and France	6 months	71.516 observations (5 MB)	1 observation / hour, from 16 weather stations
Contextual	Geographical	Various, including S57 nautical charts	ESRI shapefiles	Europe	-	22 different features (1.4 GB)	-
	Port Registers	World Port Index, SeaDataNet	ESRI shapefiles	World	-	5754 different ports (70 MB)	-
	Vessel Registers	Europe, Agence National des Fréquences	Flat files (comma-separated values)	-	-	166.683 distinct ships	-

Part of the data listed in this table and the related documentation is available for sharing in Zenodo (see [6] and details in [13]).

Surveillance information

Vessel position reports are coming from multiple sensors and sensor networks, but and in the execution of context of datAcron we focus on AIS (Automatic Identification System) data:

AIS (Automatic Identification System) data.

"According to the European Commission regulations several types of ships are obliged to broadcast AIS messages, including: ships of 300 gross tonnage and upwards in international voyages; 500 and upwards for cargoes not in international waters and passenger vessels; and, more recently, smaller fishing vessels. [...]
Two main classes of messages are distinguished:

- Kinematic messages from which 2D vessel routes can be derived.
- Static messages providing ship meta-information such as ship identifiers (MMSI and IMO number), name, type, and dimension of vessel, and route-based information,

such as destination (Port of Call), danger, Estimated Time of Arrival (ETA), draught." [14]

"The AIS data for the datAcron project has been sourced from a range of terrestrial AIS (T-AIS) and satellite AIS (S-AIS) sources. The terrestrial sources are collated from various sources in Europe and decimated to limit the amount of data. The satellite data is obtained from the ORBCOMM constellation of satellites of various generations and include an AIS receiver on the International Space Station (ISS), a range of older generation satellites and 11 new generation satellites that go to make up 19 sources of satellite data." [15]

Each of the satellites can only download AIS data when there is a ground station within their coverage footprint. There are three satellite ground stations in Morocco, Italy and Norway, servicing the area of interest for which the datAcron project is consuming AIS data. The overlap of the satellite footprint, the orbit of the satellite and the coverage of a ground station primarily affects the delay between when an AIS signal is received on the satellite to when it is available to the datAcron partners via the satellite data collection and processing network. datAcron has defined the following area of interest for AIS data to be obtained and provided:

North West Coordinates: 52 Degrees North, 12 Degrees West
South East Coordinates: 30 Degrees North, 30 Degrees East (which is expanded further east, including all the Mediterranean Sea).

"The AIS data is provided in its raw and unparsed format as received from the AIS data sources, and can contain any of the 27 different message types as described in the ITU-R.M 1371-4 or NMEA 4.0 specification." [15]

Metadata:
AIS messages provided by IMISG include a comment or TAG block which provides metadata and additional information for the IEC 61162-1 message, with the following format:

Identifier	Description
s:	The source of the message.
c:	The unix timestamp of the message when received (seconds since midnight, January 1st, 1970)
T:	The human readable timestamp of the message when received in yyyy-mm-dd hh.nn.ss
e:	The message error flag. Bits in this identifier are set to '1' if any of the 15 errors described in Annex A are true. If no error is present in the data, is this field excluded from the message output
i:	Proprietary data and contains the following fields, separated by a '|' character: <table><thead><tr><th>Identifier</th><th>Description</th></tr></thead><tbody><tr><td>X=</td><td>Data source RX / TX capability = always set to '0'</td></tr><tr><td>D=</td><td>Data source 'delayed data flag' = always set to '1'</td></tr><tr><td>T=</td><td>Proprietary timestamp of the message</td></tr><tr><td>P=</td><td>The IP address and port where the message was received by the MSA</td></tr><tr><td>R=</td><td>The direction of the message</td></tr></tbody></table>

Scale:
The data growth on the IMISG existing satellite AIS systems is about 7GB per day and this is expected to increase to about 12GB per day based on the increased number of satellites collecting AIS data and additional data available within the message structure.

Limiting the data to the Mediterranean area only the AIS data per day is expected to be about 15% of the above for the initial phase (1.8GB per day). As and when the metadata is included, the data served is expected to increase by a further 30% on top of what is available in the initial phase (2.3GB per day).

datAcron also plans to include some terrestrial AIS data that could add a further 3GB per day only as archival data.

Provision methods:
As a file:
The dataset where provided as a flat file as a series of AIS messages with the following specifications:

- Area of interest: North West Coordinates are (52 Degrees North, 12 Degrees West) and
  South East Coordinates are (30 Degrees North, 37 Degrees East).
- Start time: 2016-01-01 00:00:00 UTC
- End time: 2016-01-31 23:59:59 UTC
- ITU Message types selected: ITU123, ITU5, ITU9, ITU18, ITU19, ITU21, ITU24
- Total number of AIS messages: 80169806
- Total File size: 8.5G

As a stream:
IMIS is delivering an AIS data steam and delivering the data via a TCP/IP data stream. The characteristics of the data stream is given as:

- Area of interest: North West Coordinates are (52 Degrees North, 12 Degrees West) and
  South East Coordinates are (30 Degrees North, 37 Degrees East).
- ITU Message types provided: ITU123, ITU5, ITU9, ITU18, ITU19, ITU21, ITU24
- Average AIS message velocity: A total of 25K-30K over 15 minutes.
- Average number of vessels transmitting AIS: 35 000 transmitting AIS messages over 15min intervals for one day.

Provider: IMISG

License to datAcron: Internal to datAcron

datAcron distribution licence: N/A

Partner responsible: IMISG

Brest Area Flat File:
"This dataset covers a time span of six months, from October 1st, 2015 to March 31st, 2016 and provides ships positions within Celtic sea, the Channel and Bay of Biscay (France). The bounding box has the following coordinates: Longitude between -10.00 and 0.00 and Latitude between 45.00 and 51.00. [...] The position of the terrestrial receiver used for NARI AIS data recording, in shapefile format." [15]

"Raw AIS messages are differentiated in 27 types that have been subdivided into two main classes:
dynamic and static messages.

- AIS dynamic messages: Dynamic messages provide information on position, speed, heading, course over ground, rate of turn. Here, the following ITU Message types were

selected: ITU 1, ITU 2, ITU 3, ITU 18, ITU 19. The information extracted from these messages contains: the ship identifier (MMSI), the coordinates of the vessel and the associated speed, its heading, the course over ground. Dynamic data also include locations of aids to navigation (ATON) provided by message ITU 21 and search and rescue (SAR) provided by message ITU 9.

• AIS static messages: These messages provide ship meta-information such as ship identifiers (MMSI and IMO number), name, type, and dimension of the vessel, its destination, its estimated time of arrival (ETA), its draught. The following
ITU message types were selected: ITU 5, ITU 19, ITU 24.

• AIS status, codes and types: some data fields of AIS messages are encoded, usually with integer codes. For a clear understanding and analysis of AIS data, the encoding enumerations with associated information are required. Several files explaining this equivalences have been integrated in the dataset:

   − Status: The navigational status (e.g. moored, under way) provided in the dynamic messages is coded by an integer. The corresponding types have been detailed in a status file (in CSV format).

   − Country Codes: The list of the country codes corresponding to the first three digits of each MMSI number (e.g. 227 is France) is detailed in a CSV file.” [15]

Provider: NARI

License to datAcron: CC-BY-NC-SA-4.0

datAcron *distribution licence:* CC-BY-NC-SA-4.0

Partner(s) responsible: NARI

METOC Data

“The reference source of harmonized oceanographic data in Europe is by far the Copernicus Marine Environment and Monitoring Service (CMEMS)[1], developed by the EU as part of the European Programme for the establishment of a European capacity for Earth Observation and Monitoring. This operative service provides an interactive catalogue of updated oceanographic products produced by the network of oceanographic centres in Europe.” [14]

In particular, weather and ocean datasets used in the maritime use case include:

National Oceanic and Atmospheric Administration (NOAA) datasets[2], including “global meteorological and oceanographic datasets from cooperating networks of ships and buoys” [14], sourced by Med Office (UK).

Scale:
It concerns the period 2015-10-01 00:00:00 UTC to 2016-03-31 23:59:59 UTC and covers Longitude between -10.00 and 0.00 and Latitude between 45.00 and 51.00

Provision method:
As a csv file (http://rp5.ua/archive.php) - sourced by Met Office (United Kingdom) based on data provided by NOAA (United States)

---

[1] Copernicus Marine Environment and Monitoring Service: marine.copernicus.eu

[2] www.ndbc.noaa.gov/data

Provider: NOAA (rp5.ua)

License to datAcron: Public

datAcron distribution licence: Public

Partner(s) responsible: NARI

Sea sate based on WAVEWATCH III model: This is a public dataset, details of which are made available at [7].

Scale:
It concerns the period 2015-10-01 00:00:00 UTC to 2016-03-31 23:59:59 UTC
and covers Longitude between -10.00 and 0.00 and Latitude between 45.00 and 51.00

Provision method:
As a csv file.

Provider: SHOM - IFREMER

License to datAcron: Public.

datAcron distribution licence: Public

Partner(s) responsible: NARI

Contextual Information data sources.

Several data sources, mainly European sources or sources provided from third parties, are used to contextualize the Maritime scenarios, as well as to further characterize the other scenarios included in the fishing activities monitoring use case, as it has been detailed in D5.1. These data sources (including Brest port, Aircraft stations, Anchorage areas, Brest bay, Dumping Ground, Fairway, Inshore Traffic Zone, Recommended Track, Surveillance Centers, Ushant TSS, Western Rescue Stations, Local Shellfish Vessels) have been integrated in the reference dataset with utilisation constraints – as detailed in D5.4 [15]. These are archival data provided in files, they can be exploited only with a non-profit usage and are limited to the datAcron project. Their names and format are as follows:

Geographical and Environmental datasets (see [15] for proper reference to original datasets):

"**Brest port**: A polygon representation of the Brest port (shapefile).
Ports of Brittany: Names and coordinates of 222 ports of Brittany (shapefile).
SeaDataNet port index: Names and coordinates of about 5000 halieutic ports throughout the world (shapefile).
World port index: Names and coordinates of about 3700 ports throughout the world (shapefile).
Aircraft stations: 14 aircraft stations extracted from S57 AIRARE (airport/airfield) objects (as shapefile).
Anchorage areas in Brest bay as shapefile (extracted from S57 ACHARE objects).
Brest bay: Two polygons defining the Brest Bay as shapefile (the extended bay polygon also include entrance area).
Dumping Ground: The shapefile contains S57 DMPGRD objects (dumping ground).

Europe coastline: High resolution European coastline (polylines and polygons, as shapefile), created by the European Environmental Agency (EEA) for highly detailed analysis.

European Maritime boundaries: The shapefile contains maritime boundaries that include territorial waters, bi- or multi-lateral boundaries as well as contiguous and exclusive economic zones.

Fairway: The shapefile contains S57 FAIRWY objects (Fairway).

World Seas: The shapefile contains names and polygones of world seas.

Inshore Traffic Zone: The shapefile contains S57 ISTZNE objects (Inshore Traffic Zone).

Recommended Track: The shapefile contains S57 RECTRC objects (Recommended Track).

Surveillance Centers: The shapefile contains 59 coastguard stations (extracted from S57 CGUSTA objects).

Ushant TSS: The shapefile contains the Ushant traffic separation zone (obtained from S57 TSEZNE object).

Western Rescue Stations: The shapefile contains S57 RSCSTA objects (Rescue station).

Exclusive Economic Zones: Two shapefiles contain Exclusive Economic Zones Boundaries (polygones and polylines). Areas beyond this boundary can be classified as high seas.

 FAO fishing areas: The shapefile contains fishing areas estimation provided by the Food and Agriculture Organization of the United Nation.

 Fishing constraints: The shapefile contains two geographic areas where shellfish fishing activity is forbidden in a the time window of the dataset.

 Natura 2000 areas: Contains a collection of maritime Natura 2000 areas is shown for all the countries of the European Union.

 Fishing locations: The shapefile contains usual fishing grounds at European scale between September 2014 and September 2015." [15]

Vessel oriented datasets:

Fishing Vessels Fleet Register: European Union fishing vessels with Member State flags.

ANFR Fleet Register: Fleet register of the French Frequencies Agency

Local Shellfish Vessels: A list of 63 known shellfish vessels with license in Brest area.

ICCAT Blacklist vessels: A list of vessels presumed to have carried out illegal, unreported, and unregulated (IUU) fishing activities in the ICCAT convention area and other areas.

Partner(s) responsible: NARI.

3.3 Data created

Data created during the project by means of the project-specific computing components are as follows:

Analytics components results

These concern trajectories detected and predicted, computed by the components developed in WP1 and WP2, low-level events computed by the in-situ components (WP1) or high-level events computing by the event recognition and forecasting components (WP3). The methodology for creating these datasets are within the core of datAcron research objectives and are described in the description of action for the project.

Metadata:
There are not metadata for these data sources / datasets. However, datAcron has developed an ontology for describing the data for trajectories and events, their interlinking, as well as their interlinking with other data sources (contextual and weather). This ontology has been specified in deliverables D1.4 and D1.9 on "Data integration, management". The description of such datasets, when necessary, is made according to the schema for describing metadata for any data source and data set (as specified in Section 4 of this deliverable).

Scale:
Trajectory and events data are available under demand for any particular scenario that is studied. In principle, the range of available dates should coincide with the range of available surveillance data. Ideally, for a particular data set of surveillance data should be a set containing n compressed and enriched trajectories per moving entity, and m specific events detected or forecasted, where n and m are driven by the particular use case scenario that is studied.
The typical size of each trajectory and event is difficult to be estimated, but their size depends on the compression achieved (e.g. number of moving entity states), as well as by the number of variables in each state (e.g. contextual and weather variables associated).

Provision methods:
Trajectories and/or events can be fetched from Kafka streams and/or retrieved by specific queries addressed to the datAcron store, with respect to spatiotemporal constraints and specific properties of moving entities, contextual and weather data.
However, specific datasets – to be decided among partners responsible for data sources and for the computation of analytics results – may be devised and stored in a specific form (e.g. as RDF or csv files) accompanied with specific metadata.

***Partner(s) responsible*: UPRC (WP1 & WP2), NCSD'D' (WP3), FRHF (WP4).**

Surveillance data synopses

These concern synopses of moving entities trajectories, aiming to reduce the velocity and volume of surveillance data without compromising the accuracy of analytics components to be developed. One of the key project objectives and key performance indicators is to achieve a rate of compression that is greater than 95%, subject to the accuracy of detection and forecasting components (both for trajectories and events).

Metadata:
There are not metadata for these data sources / datasets created. However, datAcron has developed an ontology for describing the data for synopses of trajectories, and their

interlinking to other data sources (contextual and weather). The metadata used for describing such datasets is according to the schema for describing any data source and data set (Section 4).

Scale:

Trajectory synopses are available as streaming data, and according to the loosely-coupled architecture for datAcron, they are communicated among modules via the Apache Kafka messaging system, while they are also stored in the datAcron store. In addition to that, such synopses are available under demand for any particular scenario that is going to be studied. Ideally, for a particular data set of surveillance data should be a set containing a synoptic trajectory data per moving entity and trajectory.

The typical size of each trajectory synopses depends on the compression achieved, and if such a trajectory has been linked to other data, it also depends by the number of variables in each state (e.g. contextual and weather variables associated).

Provision methods:

Synopses of trajectories are made available either as streaming (via Apache Kafka streams) or archival data (as text files) to any other components of the datAcron integrated system, or to third parties, given an agreement for such a provision among partners exists. These can be retrieved by specific queries addressed to the datAcron store with respect to spatiotemporal constraints and specific properties of moving entities, contextual and weather data, as well as via access to the appropriate Apache Kafka streams, in the formats specified in D1.6 and D1.11 on integrated prototype software design. Specific datasets – to be decided among partners responsible for data sources and for the computation of synopses – may be devised and stored in a specific form (e.g. as RDF or csv files) accompanied with specific metadata.

***Partner(s) responsible:* UPRC, FRHF (WP1 & WP2).**

4. STANDARDS AND METADATA

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

Data sources to be exploited, as described in the previous section and detailed in interim deliverables D5.2, D6.2, and the final ones D5.4 and D6.4 on data preparation and curation for the maritime and aviation domains, may be accompanied with domain-specific metadata descriptions. However, there is a diversity of the metadata being used, either at the syntactic or at the semantic levels, while we need metadata for registering sources and datasets in a common repository for locating, accessing, and sharing data.

In datAcron there are two complimentary ways to locate, access and share data:

First, a quite natural and direct way – based on the datAcron data value chain: This is based to the fact that all data, from any data source, together with results created (computed), are integrated/interlinked and are made available as streams and/or stored in the datAcron store. Thus, in principle all stakeholders, given the appropriate licence, may access the data store and fetch data under a coherent schema, according to specific spatiotemporal constraints and constraints on other qualities represented by means of RDF property values and semantic types.

Second, any coherent subset of data throughout the datAcron data chain is provided as a downloadable file in a specific form, or via an API. Such a dataset may concern integrated views of data, 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 getting just a subset of the variables provided, in a required form, etc), or data created by any of the datAcron research components. In this case, each such dataset is properly registered and described according to metadata properties mapped to widely used schema such as DCAT (Data catalogue vocabulary).

DCAT [4] is a widely used standard that is RDF designed to facilitate interoperability between data catalogs published on the Web, although initially intended for governmental data only. An application profile of the W3C standard DCAT called DCAT-AP (DCAT Application Profile) [5] is used within Europe.

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 (or a store query endpoint) can be defined as an instance of dcat:Dataset and the API (or endpoint) can be defined as an instance of dcat:Distribution. DCAT itself does not define properties specific to APIs description.

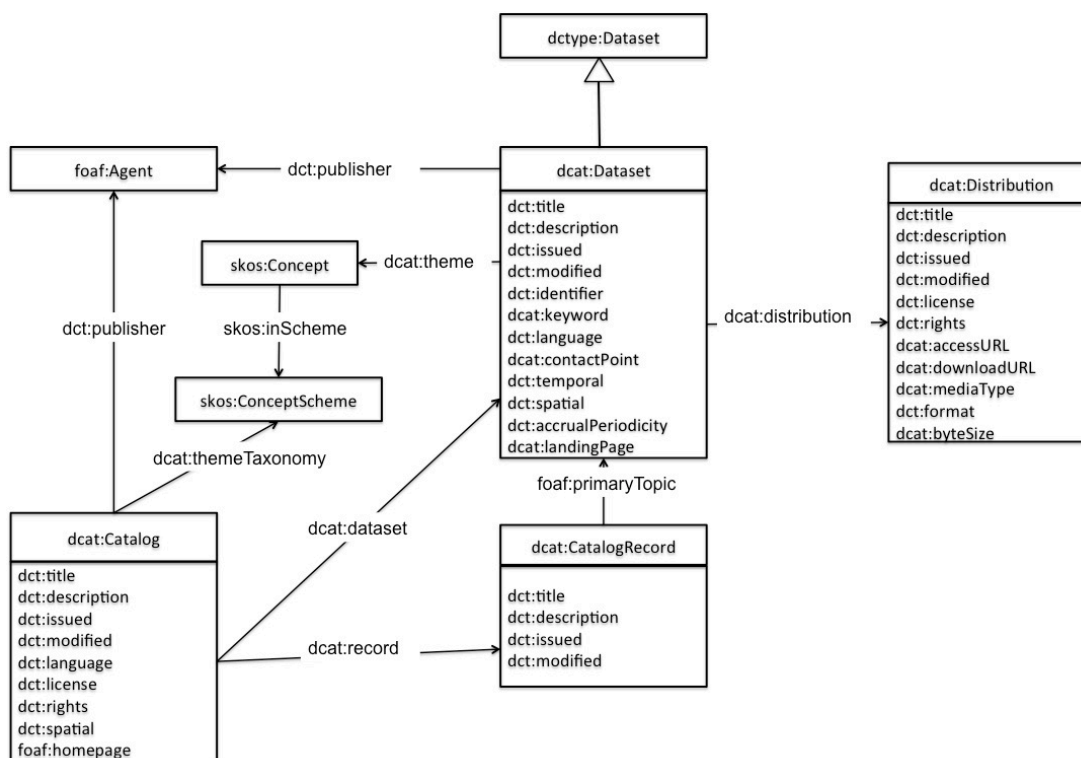An overview of metadata description based on DCAT is given in Figure 3.

Figure 3. DCAT overview

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 datAcron datasets. This imposes a workload to creating metadata that we aim to keep at a minimum level, although clearly to the level of quality that allows managing and sharing datasets.

Among the data management tools available, datAcron has registered all datasets provided by data providers and datAcron partners in CKAN[3], installed in the datAcron cluster hosted in University of Piraeus. Also, all these datasets, with the appropriate naming conventions and descriptions have been uploaded to the datAcron filestore, in WP5 (for the Maritime domain) and WP6 (for the Aviation domain) folders, where all partners have access. In addition to these datasets, WP2 trajectory synopses generated from surveillance data from both domains have been uploaded in the datAcron filestore (in appropriate folders in WP2) and registered in datAcron CKAN. The same holds for maritime complex events from WP3.

Specifically, in datAcron we have described datasets registered in the CKAN by considering the following properties, which have been mapped to CKAN fields as follows:

Dcat Property	CKAN Field	Description
`dcat:Dataset.dct:title`	Title	dataset name
`dcat:Dataset.dct:description`	Description	short description of the dataset

---

[3] http://ckan.org/

`dcat:Dataset.dcat:contactPoint`	Organization + Maintainer	datAcron beneficiary responsible for the dataset
`dcat:Dataset.dct:temporal`	Period Covered	temporal period of the data
`dcat:Dataset.dct:spatial`	Area Covered	spatial coverage of the data
`dcat:Dataset.dct:issued`	Date Issued	date at which the dataset issued within datAcron
`dcat:Dataset.dct:licence`	Licence (for using it)	license scheme to which the dataset is provided to datAcron by the Provider
`dcat:Dataset.dct:identifier`	URL	unique identifier of the dataset, specified by the datAcron filestore location of the dataset
`dcat:Dataset.datAcron:Provider`	Provider	Provider of the dataset
`dcat:Distribution.dct:licence`	Distribution Licence / CKAN Licence	license scheme under which datAcron distributes the dataset
`dcat:Distribution.dct:format`	Format	format in which the dataset is provided
`dcat:Distribution.dcat:bytesize`	Bytesize	size of the dataset in bytes
`dcat:Distribution.dcat:accessURL`	Filestore location	link to a point to access data

4.2 Organization of datasets.

Datasets provided are stored together with their metadata descriptions providing a link to a dataset file or to an API, according to the metadata schema used.

CKAN organizes data as follows: datasets are named as packages that is actually the root folder of a dataset. Packages contain resources which are files (e.g. csv, json, rdf).
There are distinct organizations which are the packages owners, and users can upload or manage datasets. Organizations are the data providers that are datAcron beneficiaries.

The pipeline to upload a dataset is, first has to create a package and provide the fields: title, url, organization, visibility (private or public).

datAcron datasets are described using the properties specified in Section 4.1 and are kept private.

5. DATA SHARING

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

Datasets can be shared either by opening a specific API or a query endpoint for third parties to fetch data from the datAcron store, or providing access rights to the datAcron CKAN so as stakeholders to download datasets as single files.

The methods used to share data provided by beneficiaries depend 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 datAcron consortium agreement.

BRTE specific background information:

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)	BRTE can use European Air Traffic Data in datAcron activities. The transfer of data to other consortium members would require prior approval from the third party data Provider.	BRTE has to explicitly mention the third party data Provider copyright of the data in any dissemination involving data publishing.
Weather data	BRTE can use public domain and third party weather data in datAcron activities. The transfer of the third party data to other consortium members would requires prior approval from the third party data Provider.	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.
FlightAware Data[4]	FlightAware data can be used just for research, publication of results is fine with the acknowledge credit to FlightAware but data can't be redistributed without permission.	Receiver network may require credits on the published dissemination.

---

[4] New data source, not included in the Consortium Agreement: Its specification in this deliverable clarifies the agreed use within the scope of datAcron.

CRIDA specific background information:

CRIDA is a research center, not a data-generator, which entrusted to use, exploit and maintain the datasets generated by ENAIRE, Provider of that data. CRIDA is entitled to use the data for research purposes such as the datAcron 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 datAcron 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 datAcron 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 datAcron 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 enter, 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

IMISG specific background information:
As far as the AIS related data sources from IMISG are concerned, the following background is identified and agreed upon for the Project.

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)
Terrestrial & Satellite AIS data. These will be enriched with metadata, and include archival and real-time data made available via REST query services and CSV data files.	Access is on a royalty-free basis to background, for beneficiaries needed to implement their own tasks under the action.	Access is under fair and reasonable conditions to background needed for beneficiaries to exploit their own results. The raw data cannot be published although samples can be provided.

NARI and CMRE specific background information
(not included in the Consortium Agreement)**:**

As far as the data sources provided by NARI are concerned, the following background is identified and agreed upon for the Project:

- Terrestrial AIS data from NARI datasets, enriched with metadata and included archival data made available via CSV data files, are available as open data according to CC-BY-NC-SA. Specific reference to the original availability of these data in the Zenodo repository [6,13] should be maintained as specified at [6,13]. In particular, the access through the Zenodo repository [6,13] will remain the main access for this dataset. Any distribution not acknowledging the original source at [6,13] will be considered a violation of the terms and condition for data access (i.e. violation of "attribution"). For use in publications, credits should be given according to the conditions specified in [6,13], i.e., cite as:
-
  "RAY, Cyril, DRÉO, Richard, CAMOSSI, Elena, & JOUSSELME, Anne-Laure. (2018). Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance (Version 0.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1167595"

- To the other data available for download in [6,13] provided by NARI, the same conditions as above apply.
- To the other data available for download in [6,13] but originated from third parties, the conditions specified by the data originators and made available together with each data subset in [6,13] apply.
- To the other data prepared by NARI and CMRE and currently available in the datAcron filestore to projects partners only, access restrictions from third parties and limitation of use apply. Those data will not be available for sharing beyond the execution of the project.
- To the data prepared by CMRE (vessel routes, degraded data for evaluation), NATO policies [8-12] and CMRE terms and conditions for data and information reuse apply. Use in scientific publications needs to be pre-approved and acknowledged. Those data might be available after the execution of the project on the basis of existing NATO policies and rules for data release [8-12] and on the basis of CMRE terms and conditions for data reuse.

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

Concerning licensing, the datAcron Ethics Management Plan D.8.5 (section 3.3, Identification and documentation of licensing options) makes a clear identification of all options that the project must consider for making datasets available to 3rd parties.

An important challenge is the integration/interlinking of data from data sources and datasets having different usage and access policies. Interlinking data with certain constraints and requirements with data that are publicly and freely available impacts the access policy applied. This is not problematic when the aggregated data is subject to the same or more restrictive access, usage and dissemination conditions as the source data themselves.

There should be great care to avoid problematic situations where data are being distributed throughout channels to stakeholders that do not satisfy the conditions stipulated by one of the sources. To prevent incorrect usage, managing the access, usage and dissemination conditions of the newly created datasets is important. That information forms the cornerstone of the correct implementation of the required access, usage and dissemination policies.

datAcron applies the following strategy:

- The data providers ensure that for each dataset, the access, sharing and reuse policy is known. In case the dataset integrates data from multiple data sources, the more restrictive access, sharing or reuse conditions apply. This information is reported to the datAcron Executive Board, together with a request to apply a specific access/share/reuse policy to a dataset.
- The datAcron Executive Board, in respect of existing conditions and legislation on information data sharing and reuse, decides with the data providers on the specific policy to be applied for each dataset, and decides on the license to be applied. The decision must be taken in 3 weeks from the date of request.
- The exposure of the dataset is made according to the above-made decision. The licencing scheme agreed in datAcron is CC-BY-NC-SA-4.0. Metadata are updated accordingly.
- The metadata of the created outcome is always public. This ensures transparency of the knowledge that is gathered.

For the data generated during the remaining of the execution of the project, and excluding data generated from the heterogeneous maritime dataset at [6,13], for which the conditions

at [6,13] apply, the workflow to be followed for deciding on granting open access to specific parts of the research data (datasets) and under which restrictions, license etc. is as follows:

Given a specific publication or digital research data generated in the action by a beneficiary, the beneficiary must aim to deposit at the latest on publication time of the manuscript, or at the latest one month after the generated dataset (if this is not associated to a publication), the research data (datasets) in the deposited scientific publication, together with the metadata. To do so

- Metadata descriptions of the datasets are specified to the datAcron repository, together with either a link to a dataset file, or to an API for fetching the associated data from the datAcron store.
- The beneficiary ensures that for each dataset, the access, sharing and reuse policy information is known. In case the dataset integrates data from multiple data sources, the more restrictive access, sharing or reuse conditions apply.
  This information, together with the (link to the) dataset is provided to the Executive Board at least 4 weeks prior to the deadline for registering the dataset.
- The datAcron Executive Board decides on the specific policy to be applied for each dataset. The licencing scheme agreed in datAcron is CC-BY-NC-SA-4.0.
  The decision must be taken in 3 weeks from the date of request.
- The exposure of the dataset is made according to the above-made decision. Metadata are updated accordingly.
- The metadata of the created outcome is always public. This ensures transparency of the knowledge that is gathered.

6. ARCHIVING AND PRESERVATION

Data sources concerning moving entities' behavior in the air and in the sea, are growing in size due to market and service demands. datAcron exploits such data sources, together with other geospatial, environmental and weather data sources, and develop scalable methods for processing, managing and analyzing these data. There is a growing interest in archiving, sensing and performing analytics over mobility and behavioral data. In datAcron, we address issues concerning long term storage of these data, as well as research generated data.

These issues related to 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 do consider the following issues:

- What is the volume of the data to be maintained?

Partners that provide data to the project aim to maintain all data provided to datAcron, either in the form already provided or in forms already stored and curated according to own organization policies.

datAcron partners, except from data Providers, cannot preserve and exploit data provided to the project beyond the duration of the project. This does not hold for shared data (under specific licensing schemes) on in specific cases where special agreement for data preservation/exploitation is in place.

- What is considered long-term (2-3 years, 10 years, etc.)?

datAcron partners that provide data to the consortium aim to maintain data according to own organization policies, but surely within a 5-years horizon.

- Identification of archive for long-term preservation of data.

datAcron partners aim to maintain own data according to own organization policies (i.e. to in-house data stores / archives).

In cases of shared datasets (e.g those already identified in the Maritime domain) archives in repositories (e.g. OpenAIRE) will remain the main point of access for these datasets.

In particular, the heterogeneous maritime dataset available at [6,13] will remain available beyond the 5-year horizon at the following url http://doi.org/10.5281/zenodo.1167595 for research and training purposes beyond the end of datAcron, according to the conditions specified at [7,14]. Updated versions of the dataset, including additional datasets and related documentation, will be linked to the existing version. The Zenodo url http://doi.org/10.5281/zenodo.1167595 and the corresponding DOI will always link to the updated version, while enabling access to previous versions.

datAcron partners that do not own data can store and maintain those made publicly available and those shared from data providers, except in cases where special agreements are made among partners, out of the datAcron scope.

- Which datasets will need to be preserved in the archive?

After closing the project, the datAcron archive may include the datasets that partners have agreed to share (the licencing scheme agreed in datAcron is CC-BY-NC-SA-4.0).

In any case (i.e for any such dataset), the already established archives in widely-used

repositories will remain the main points of access for any such shared dataset.

- What about relevant dependent datasets?

There is no specific policy on this.

Up to now, except from the public datasets made available to ᴅᴀᴛᴀᴄʀᴏɴ, the only datasets shared to stakeholders (thus, those datasets that ᴅᴀᴛᴀᴄʀᴏɴ partners would be able to exploit after the project duration) include the maritime open access datasets. Depending on the final sets of open access datasets, partners will consider costs, including:

- 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).

All the above issues will be further detailed in the last version of the DMP.

7. CONCLUSIONS

The datAcron Data Management Plan (datAcron DMP) details what data the project collects, generates, how these are exploited or made accessible to all stakeholders, how and what data sets are made available for verification and re-use, and how it will be curated and preserved during the lifetime of the project, in accordance to the datAcron data value chain. In order for the document to be self-contained, a comprehensive view of the datAcron data lifecycle is provided with appropriate definitions of terms being used and stated assumptions under which the plan has been devised, with a succinct description of stakeholders' groups.

Data sources exploited per domain are described, also specifying information on existing domain-specific metadata per data source, size of data sets, modality, and provision methods. The deliverable specifies details on data sources and datasets according to a metadata scheme and how these are registered in the datAcron CKAN repository.

Furthermore, information and the datAcron policy on data sharing is provided, also taking into account, limitations, IPR and legal issues applying to the data sources, together with licensing and ethical issues.

The deliverable concludes with issues concerning archiving and preservation.

8. REFERENCES

[1] Resource Description Format, http://www.w3.org/RDF/ & http://www.w3.org/standards/techs/rdf#w3c_all

[2] UK DMP checklist [http://ukdataservice.ac.uk/manage-data/plan/checklist.aspx]

[3] Guidelines on Data Management in H2020, Version 2.1, 15 Feb. 2016 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[4] DCAT Vocabulary [http://www.w3.org/TR/vocab-dcat/]

[5] DCAT-AP [https://joinup.ec.europa.eu/asset/dcat_application_profile/description]

[6] Cyril Ray, Richard Dréo, Elena Camossi, Anne-Laure Jousselme (2018). Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance (Version 0.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1167595

[7] Edwige Boudiere, Christophe Maisondieu, Fabrice Ardhuin, Mickael Accensi, Lucia Pineau-Guillou, Jeremy Lepesqueur (2013). A suitable metocean hindcast database for the design of Marine energy converters . International Journal of Marine Energy , Volume 3-4, December 2013, pages e40-e52. http://doi.org/10.1016/j.ijome.2013.11.010

[8] [NIMP C-M(2007)0018] NATO Information Management Policy (NIMP)

[9] [PDIM C-M(2008)0113] Primary Directive on Information Management (PDIM)

[10] [C-M(2002)60] The management of non-classified NATO Information

[11] [C-M(2002)49] Security within the North Atlantic Treaty Organization

[12] [AC/35-D/2002 Rev 42] Directive on the Security of Information

[13] Cyril Ray, Richard Dréo, Elena Camossi, Anne-Laure Jousselme, Clément Iphar. Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance. Data in Brief. Elsevier (submitted)

[14] Cyril Ray, Elena Camossi, Anne-Laure Jousselme, Melita Hadzagic, Christophe Claramunt, Ernie Batty, *Maritime Data Preparation and Curation (interim)*, datAcron Deliverable 5.2, June 2016.

[15] Cyril Ray, Clément Iphar, Richard Dréo, Waldo Kleynhans, Elena Camossi, Anne-Laure Jousselme, Maximilian Zocholl, Ernie Batty, Quentin Roche, Arnaud Metzger, *Maritime Data Preparation and Curation (final)*, datAcron Deliverable 5.4, July 2018.

[16] Garcia Martinez Miguel, Calvo Fernandez, Esther, Costas Pablo, Fresno Cambre Laura, López Ballestín Elena, Iglesias Martinez Enrique, Lopez Leones Javier, Scarlatti David, Cordero José Manuel, *Aviation Data Preparation and Curation (final)*, datAcron Deliverable 6.4, July 2018.