



Data Management Plan V2
Work Package 8 Task 8.3 Deliverable 8.4

Authors


Nikos Giatrakos, Antonios Deligiannakis
Athena Research & Innovation Center (Athena)

Arnau Montagud, Miguel Ponce de León
Barcelona Supercomputing Center (BSC)

Holger Arndt, Stefan Burkard
Spring Techno (Spring)

Konstantina Bereta, Konstantinos Chatzikokolakis,
Marios Vodas, Dimitris Zisis
MarineTraffic (MT)

Elena Camossi, Gabrielle Ferri, Raffaele Grasso
NATO STO CMRE (CMRE)

 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.: WP8 D8.4
		Rev.: 1.0
		Date: 30/06/2020
		Class.: Public



Distribution list:

Groups:	Others:
WP Leader: Athena Task Leader: Athena	Internal Reviewer Partner: BSC, Spring, MT, CMRE INFORE Management Board INFORE Project Officer

Document history:

Revision	Date	Section	Page	Modification
0.1	14/05/2020	1-3	5-18	Creation
0.2	20/05/2020	4	19-29	Creation
0.3	21/05/2020	-	-	Submitted for internal review to Spring, BSC
0.4	22/05/2020	1-4	5-29	Internal review comments by BSC and Spring incorporated
0.5	26/05/2020	5	30-45	Creation
0.6	01/06/2020	6-9, 5	46-49, 30-45	Creation, modification
0.7	02/06/2020	-	-	Submitted for internal review to MT, CMRE
0.8	24/06/2020	1-9,10	5-50	Internal review comments incorporated
0.9	26/06/2020	1, 3-5	5, 12, 26, 43	Updated links to dataset uploads
1.0	29/06/2020	All	All	Self-review and final version

Approvals:

First Author: Nikos Giatrakos (Athena) Date: 29/06/2020

Internal Reviewers: Miguel Ponce de León (BSC), Stefan Burkard (Spring), Konstantina Bereta (MT), Raffaele Grasso (CMRE) Date: 24/06/2020

Coordinator: Antonios Deligiannakis (Athena) Date: 30/04/2020



 Project supported by the European Commission Contract no. 825070	<h3>WP8 T8.3</h3> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public




Table of contents:

1	Executive Summary	5
2	Introduction.....	6
3	Life Science Data and Management Procedures	7
3.1	Life Science Data Summary	7
3.1.1	Origin of data – Purpose of data collection – Relation to the objectives of the project	7
3.1.2	Types and format of data generated – Reused data – Expected data size	8
3.1.3	Data utility – to whom it is will be useful	10
3.2	FAIR Life Science Data.....	10
3.2.1	Making data findable including provisions for metadata	10
3.2.1.1	Standards for metadata creation, discoverability of data (metadata provision), approach towards search keyword	10
3.2.1.2	Identifiability of data, naming conventions, clear versioning	11
3.2.2	Making data openly accessible.....	12
3.2.3	Making data interoperable.....	12
3.2.4	Data re-use (through clarifying licenses)	18
3.2.5	Data security.....	18
3.2.6	Allocation of resources.....	18
4	Financial Data and Management Procedures	19
4.1	Financial Data Summary	19
4.1.1	Origin of data – Purpose of data collection – Relation to the objectives of the project	19
4.1.2	Types and format of collected data – Reused data – Expected data size.....	20
4.1.3	Data utility – to whom it is will be useful	22
4.2	FAIR Financial Data.....	22
4.2.1	Making data findable including provisions for metadata	22
4.2.1.1	Standards for metadata creation, discoverability of data (metadata provision), approach towards search keyword	22
4.2.1.2	Identifiability of data, naming conventions, clear versioning.....	25
4.2.2	Making data openly accessible.....	26
4.2.3	Making data interoperable.....	26
4.2.4	Data re-use (through clarifying licenses)	28
4.2.5	Data security.....	29
4.2.6	Allocation of resources.....	29
5	Maritime Data and Management Procedures	30
5.1	Maritime Datasets – Collection Purpose – Relation to project objectives	30
5.2	Maritime Data utility – to whom it is will be useful	31
5.3	Datasets and Dataset Summary: Origin of Data – Types and format of collected/generated data – Reused data – Expected data size	31
5.3.1	AIS Raw Data	31
5.3.2	Kafka Streams (AIS Derived Data Streams).....	33
5.3.3	Patterns of Life.....	33
5.3.4	Acoustic Data.....	34
5.3.5	Satellite Image Data	35
5.3.6	Vehicle Status Data	36
5.3.7	Thermal Camera Data	36
5.3.8	Composite Maritime Event Data.....	36
5.4	FAIR Maritime Data.....	37
5.4.1	Making data findable including provisions for metadata	37
5.4.1.1	AIS Raw, Kafka Streams and Patterns of Life data.....	37
5.4.1.2	Acoustic Data	38

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3 Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



5.4.1.3	Copernicus Data	39
5.4.1.4	Thermal Camera and Vehicle Status data	42
5.4.1.5	Composite Maritime Event Data	42
5.4.2	Making data openly accessible.....	43
5.4.3	Making data interoperable.....	44
5.4.4	Data re-use (through clarifying licenses)	45
5.4.5	Data security.....	45
5.4.6	Allocation of resources.....	45
6	Scientific Publication Data and Zenodo Repository Status.....	46
7	Expert User Requirements/Feedback Data.....	47
8	Ethical aspects.....	48
9	Conclusions.....	49
10	References.....	50

 Horizon 2020 European Union Funding for Research & Innovation	Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.: WP8 D8.4
			Rev.: 1.0
			Date: 30/06/2020
			Class.: Public



1 Executive Summary

This deliverable provides the second version of the Data Management Plan in INFORE at Month 18 of the project. The deliverable will receive its final form in Month 36. To build the Data Management Plan we utilized the respective template of the European Research Council (ERC): ERC DMP¹ provided by the Digital Curation Center. We also followed the guidelines for FAIR (findable, accessible, interoperable, reusable) data management in H2020². Since we aim to use the current document and its future versions as a single point of reference and documentation for the datasets used or produced in INFORE, we provide an as detailed as possible description of datasets and data management procedures.

Participating in the Open Research Data Pilot of the European Commission, INFORE provides access to portions of data from all three use cases (Life Science, Financial, Maritime) of the project. We have setup an INFORE community at Zenodo¹⁵. Zenodo is implementing the FAIR principles³ and is indexed by OpenAIR. In particular, we currently share:


- Life Science Data: <https://doi.org/10.5281/zenodo.3922263>, <https://doi.org/10.5281/zenodo.3921049>, <https://doi.org/10.5281/zenodo.3923070>
- Financial Data: <https://doi.org/10.5281/zenodo.3886895>
- Maritime – AIS Data: <https://doi.org/10.5281/zenodo.3754481>
- Maritime – Acoustic Data: to be uploaded by Month 19 of the project
- Maritime – Composite Maritime Event Data: <https://doi.org/10.5281/zenodo.2557290>

References to these datasets will be provided at the project web site and our contributions also appear in <https://datasetsearch.research.google.com/>. Moreover, all publications in the scope of the project have been uploaded under the INFORE community at Zenodo. All datasets in INFORE use cases do not involve sensitive information.

¹ https://dmponline.dcc.ac.uk/public_templates

² http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

³ <https://about.zenodo.org/principles/>

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public




2 Introduction

In INFORE, the developed technology will be tested in three different application domains. INFORE’s architectural components are fueled by each application’s input data and are specialized in the execution of complex workflows over multiple Big Data platforms and HPC infrastructures.

The Life Sciences Use Case builds a virtual laboratory for studying cancer evolution under the effect of combinational drug therapies. To do so, it uses data stemming from in-silico models of multi-cellular system evolution under circumstances found in in-vivo tumors. The Financial Use Case provides datasets involving stock market data arriving in high-velocity streams. The aim is to forecast price swings of stocks, predict systemic risk (i.e., great linkage between major market participants) and to aid in distinguishing investment opportunities. Finally, the Maritime Use Case aims at improving Maritime Situational Awareness, i.e. the ability to perceive and forecast activities and threats in maritime environments. To achieve that it fuses a variety of data including positioning – AIS (Automatic Identification System) data, quantities sensed by autonomous unmanned vehicles navigating at sea and satellite image data. The aim of incorporating these data in the analysis is to correlate the heterogeneous data sources towards the identification and forecasting of activities of “dark targets” that (intentionally) hide from AIS monitoring systems. Finally, an additional source of data for each use case involves requirement collection and expert user feedback.

This deliverable is organized so that it abides by the ERC DMP template¹. For each dataset we firsts provide an elaborate data summary and we then proceed with explaining our provisions for making the corresponding dataset FAIR. In that spirit, Section 3 refers to Life Science data, Section 4 elaborates on Financial data, while Section 5 presents maritime data and management planning procedures. Section 6 refers to project publications, Section 7 comments on expert user requirements and feedback data. Finally, Section 8 clarifies ethical aspects.

 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.: WP8 D8.4
		Rev.: 1.0
		Date: 30/06/2020
		Class.: Public

3 Life Science Data and Management Procedures

3.1 Life Science Data Summary

3.1.1 Origin of data – Purpose of data collection – Relation to the objectives of the project

Our goal is to develop a virtual laboratory to conduct in-silico simulations which: i) integrate heterogeneous sources of experimental data as well as biological knowledge; ii) generate hypotheses about underlying mechanisms of biological processes determining tumour growth, drug resistance and drug synergies in cells; iii) in-silico design, test and optimization of treatments based on combination of drugs. The above spans all tasks within the scope of WP1 of the project. For this, we integrate two simulation frameworks, an agent-based modelling one, namely PhysiCell [1] [2], and a Boolean modelling one, namely MaBoSS [3] [4], into a software called PhysiBoSS [5] [6].

PhysiCell is developed in Paul Macklin's lab⁴. MaBoSS and PhysiBoSS are developed in the Computational Systems Biology of Cancer group at Institut Curie⁵ (Paris, France). PhysiBoSS (from the merging of PhysiCell and MaBoSS) is an adapted version of PhysiCell to integrate inside each cell a Boolean model computation. All the aforementioned frameworks constitute open-source software and are freely available [2] [4] [6].

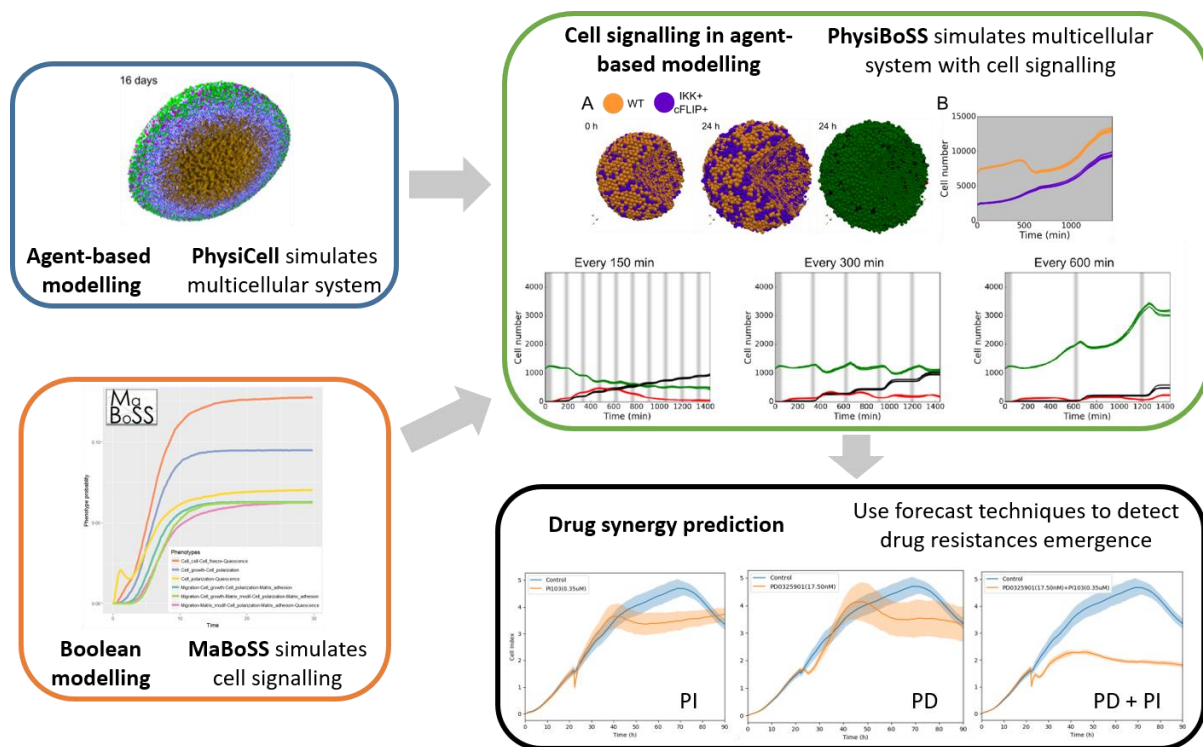


Figure 1: Data generating software and purpose of data generation in the project.

For the purposes of the project, a large number of simulations is being conducted over High Performance Computing (HPC) infrastructure, available to INFOR by the Barcelona Supercomputing Center (BSC) partner. What we seek for is to create a framework so that we can run simulations studying the behaviour of tumours of realistic sizes (thousands or millions of interacting cell - agents⁶), with respect to the effect certain drug combinations have on the evolving tumours. The effect is mainly, but not only, determined by three time series data

⁴ <http://physicell.mathcancer.org/>

⁵ <https://sysbio.curie.fr/>

⁶ PhysiCell has been parallelized with OpenMP, and its performance scales linearly with the number of cells. Simulations up to 105-106 cells are feasible on quad-core desktop workstations; to run large simulations corresponding to tumours of realistic sizes can only be possible with HPC infrastructure.

<p>Project supported by the European Commission Contract no. 825070</p>	<p>WP8 T8.3 Deliverable D8.4</p>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

on the number of non-apoptotic cell death (Non-ACD), apoptotic or proliferating cells. In this model, effective drugs' activity forces tumour cells into Non-ACD and fewer to apoptosis, as detailed in [5] .

The concept is to first selectively tag a sufficient number of performed simulations as “useful” or not, by employing machine learning (active learning, in particular) approaches. For now, the “usefulness” of a simulation is determined based on the evolution of the number of proliferating, apoptotic and non-ACD cells and on the shape of the corresponding time series, in various timesteps of the simulation process. Based on the tagged simulations, INFORE will automatically (at runtime), tag multiple, large-scale simulations run in parallel, forecasting whether they are going to be useful or not. In case a simulation is forecasted to be useful, the applied drug combination is judged as effective and the simulation is further monitored to interactively determine successive drug combinations. If a simulation is forecasted to be “unuseful”, it is ceased. By correctly forecasting and ceasing unpromising simulations, computational resources will be saved and devoted to new simulation instances with different parameters, cutting down the time needed to explore high-dimensional parameter spaces and to find in-silico hypothesis. Furthermore, utilized simulation models and parameters, as is the case with the agent-based model of PhysiBoss, will be calibrated in the scope of the project for a particular cell line/type by using the drug experiments corresponding to that particular cell line. With respect to the above goals, at a later stage, we will attempt to incorporate experimental evidence on drug synergies, collected from public available resources such as the Drug Combination DataBase (DCDB)⁷.

The workflow from training the machine learning model, tagging and forecasting the fate of live simulation instances will be drawn using the Graphical Editor of the INFORE architecture, the process will be optimized by the INFORE optimizer and will be assisted by the Machine Learning and Forecasting Component of the architecture.

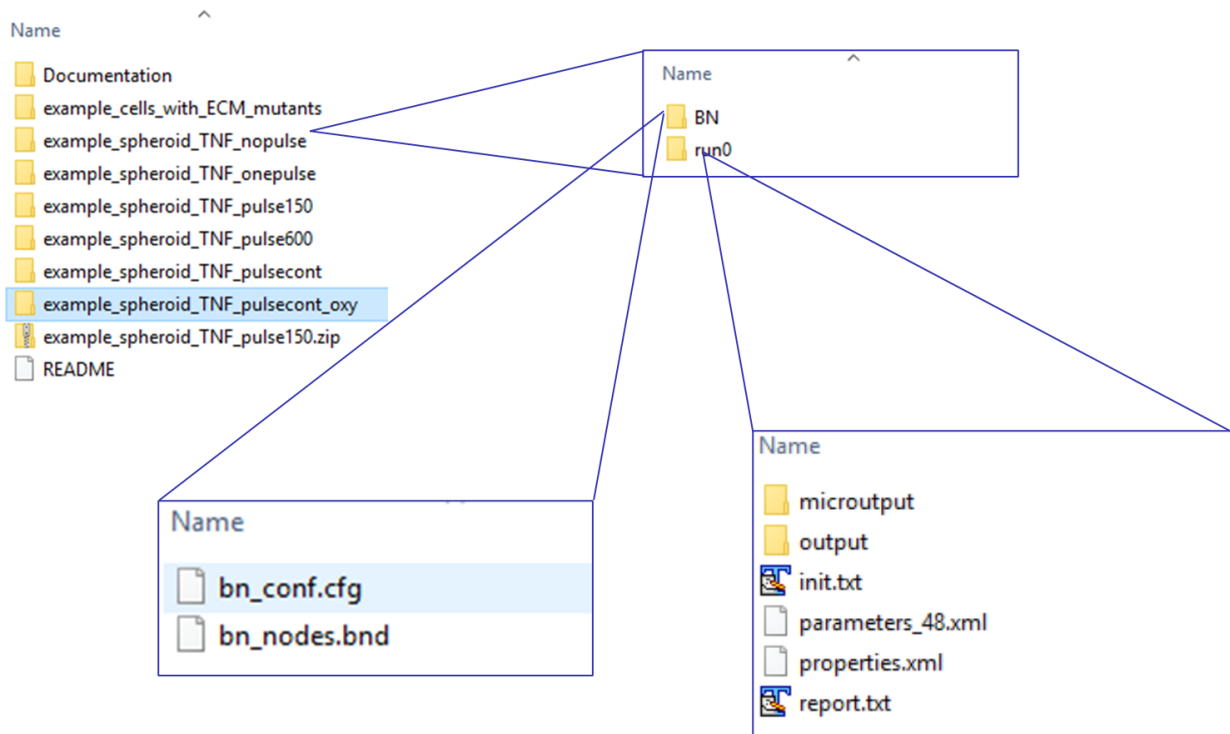


Figure 2: Structure of folders for a single tumor simulation

3.1.2 Types and format of data generated – Reused data – Expected data size

The size of the produced data depends on the size of the simulated tumor and the simulation parameters as outlined in the data dictionaries, later on in this deliverable. Indicatively, we note that simulating tumors of realistic sizes can produce data at a rate of approximately 100 GB/min [7] .

⁷ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4275564/>

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

All the information related to a single run simulation is kept into a separate folder which we term the root folder/directory. Naming conventions for all relevant data are provided in Section 3.2.1.2. The root folder contains at least two folders as illustrated in Figure 2. The first folder named BN includes two ASCII files: `bn_conf.cfg` and `bn_nodes.bnd`. These files describe the configuration of MaBoSS as well the boolean model itself and are documented in [4] under the Cell Fate model repository. The second folder, `run0` in Figure 2, termed `run` for short, contains all input and output files of the simulation. In case the same simulation, with respect to drug applications, is run multiple times using different parameters, respective folders are stored in (regular expression – alike convention is used for generic namings in this document) `/run\d*/` - named folder. The internals of the `run` folder are shown at the bottom of Figure 2 as well. For all input and output files data dictionaries are provided as will be explained in Section 3.2.3.

simulated time	num cells	num division	num death	wall time	oxygen	tnf
0	1137	1137	0	0.0211054	0	0
30	1226	89	0	12.8422	0	0
60	1238	12	0	19.6248	0	0
90	1252	14	0	23.808	0	0
120	1269	17	0	15.299	0	0
150	1286	17	0	15.7466	0	0

Figure 3: Example of content in the report.txt output file

Input files: The `parameters.xml` file describes the input configuration of the simulation. Details on the internals and the schema of this file follow in Section 3.2.3. The `init.txt` file in Figure 2 allows the user to create an initialization file containing information of the initial state of the simulation, cell types, positions, phases, etc. This text file is referenced in a `<initial_configuration>` element of `parameters.xml`. The output of a simulation, in `/cells_%05d.txt/` files, as described below, can also be **reused** as the input `init.txt` file to another simulation. A `/cells_%05d.txt/` file is a plain-text file where fields are separate by ";". Rows correspond to individual cells and fields correspond to attributes depicting the form of the cell at a particular time step. The `properties.xml` file⁸ in Figure 2 is automatically generated, composing the union of `parameters.xml` and MaBoSS configurations.

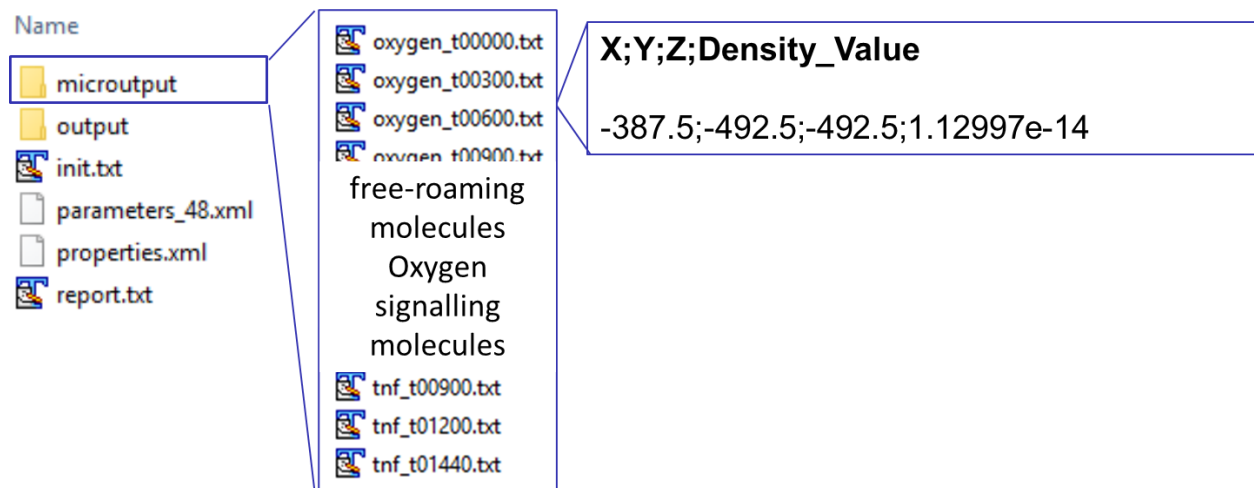


Figure 4: Structure and content of the microuput folder

Output files: The `report.txt` file includes the information about the number of cells that are alive or have died in the various time steps (expressed in simulation time and wall time) of the simulation along with the concentration

⁸ The `/_d*/` in the file naming corresponds to the degree of parallelism under which the simulation is running as will be explained later on in this section.

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

of various densities (free-roaming molecules) such as oxygen and TNF⁹. The `microutput` folder inside the `run` folder includes a series of `<densityName>_t\d{5}.txt/` files describing the position (x,y,z in a 3D simulation space) and the concentration of the corresponding density at various timesteps (included in the `t\d{5}` part of naming convention) of the simulation as illustrated in Figure 4. Finally, the `output` folder, as shown in Figure 5, contains text files named `/cells_\d{5}.txt/` and includes information about the status of the individual cells. Respective data dictionaries follow at Section 3.2.3.

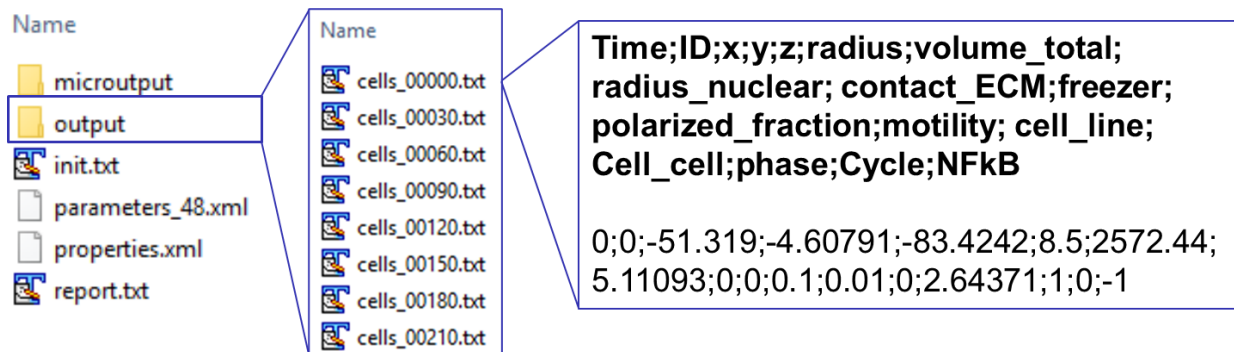


Figure 5: Structure and content of the output folder

3.1.3 Data utility – to whom it is will be useful

The data will be particularly useful for the relevant scientific communities in their efforts to develop personalized cancer therapies that will improve the treatment and the quality of life of cancer patients. Moreover, the target audiences regarding the life science datasets include researchers in Life Sciences and in Computer Science fields related to Big Data management, databases, data mining, machine learning and artificial intelligence fields.

3.2 FAIR Life Science Data

3.2.1 Making data findable including provisions for metadata

3.2.1.1 Standards for metadata creation, discoverability of data (metadata provision), approach towards search keyword

PhysiBoSS uses XML format for metadata related to the input `parameters.xml` file, as proposed in the MultiCellIDS standardization initiative¹⁰. At its current status, this parameter file complies with the earlier MultiCellXML project¹¹. Given this, the simulation data are searchable and discoverable using metadata tags and keywords provided in the simulation `parameters.xml` file. Section 3.2.3 comments on data dictionaries used and relates XML tags and keywords with their meaning and valid value ranges. Moreover, keywords can be extracted based on the naming conventions used, which follow a regular expression-alike format. For instance, if a simulated tumor is of spheroid shape, this keyword is included in the naming of the root folder, as will be described shortly. To parse the XML files, PhysiBoSS integrates in its code the Tinyxml2 files¹², which are open sourced and freely available¹³.

BSC INFOR partner currently works in the development and upgrade of a new version of PhysiBoSS, following the addon architecture proposed by Paul Macklin’s lab⁴. The main idea is to uncouple the development and maintenance of both tools so PhysiBoSS’s capabilities can evolve by adding new features, while being able to easily

⁹ A TNF inhibitor is a pharmaceutical drug that suppresses the physiologic response to tumor necrosis factor (TNF), which is part of the inflammatory response.

¹⁰ <http://multicellids.org/>

¹¹ <http://multicellxml.sourceforge.net/>

¹² <https://github.com/gletort/PhysiBoSS/tree/5fd3fcc834064c4d51fb18fef360b760dee1440f/src/tinyxml>

¹³ <http://www.grinninglizard.com/tinyxml2/index.html>

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3 Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



keep the PhysiCell core functionalities up to date. Second, because PhysiBoSS has diverged from the current PhysiCell release, there are currently two flavors for parameters (inputs) and outputs formats:

- The PhysiCell native formats (MultiCellIDS + matlab files + SVF)
- The current PhysiBoSS formats (MultiCellXML + txt (;separated) as already described)

So, in the new version of PhysiBoSS, the details of which are planned to be included in Deliverable D8.6 (Data Management Plan V3) at Month 36 of the project, the aim is to support (at least for a while) both input/output formats. Nonetheless, in the future BSC plans to proceed with the latest MultiCellIDS standard.

Finally, inspecting the standards included in RDA Metadata Repository¹⁴, the "Minimum Information for Biological and Biomedical Investigations" standard seems suitable, at a later stage of the project, to store the experimental data from the combination drug experiments. However, a final decision on this issue is still to be made.

3.2.1.2 Identifiability of data, naming conventions, clear versioning

Simulation data are shared through the INFORÉ page at Zenodo¹⁵, which is implementing the FAIR principles. Zenodo uses standard dataset identification mechanisms including Digital Object Identifiers (DOIs). In case new versions of shared datasets exist, they receive their own DOI and their Zenodo description will include the link to the older version(s). DOIs, metadata and keywords at Zenodo override naming conventions with respect to making data findable.

We now describe the naming conventions used for the Life Science datasets in regular expression-alike format. Please also recall our discussion in Section 3.1.2 about the contents of simulation files and folders.

Root folder:

```
</Cellpopulationproperties>\*_?<Cellpopulationmajorparameters>\*_?<Densities Applied>\*_?<DensitiesMajorParameters>\*_?/
```

For instance, in a folder named `spheroid_TNF_pulsecont_oxy`, `Cellpopulationproperties=spheroid`, `Cellpopulationmajorparameters=null`, `DensitiesApplied 1=TNF`, `DensitiesApplied 2=oxygen`, `DensitiesMajorParameters 1= pulsecont`, `DensitiesMajorParameters 2=null`. We have simulations of a tumor spheroid, with different oxygen tolerance conditions, with one continuous pulse of TNF cytokine applications. The difference among parameter files is the "oxygen_necrotic" value, which controls the threshold above which cells commit to necrosis due to lack of oxygen.

MaBoSS files: are always named `bn_conf.cfg` and `bn_nodes.bnd` inside the BN folder of the simulation as described in Section 3.1.2.


Simulation files: Each simulation can be run a number of times under different parameters, threshold values etc. PhysiBoSS will then create a separate folder for each run named as `/run\d*/`. The input parameter file of the simulation is named as `/parameters_?\d*?.xml/` where `?\d*?` involves the degree of parallelism used when the simulation was run. `output`, `microutput`, `report.txt`, `properties.xml`, `init.txt` files and folders have fixed names inside a `/run\d*/` folder. The `microutput` folder inside the `/run\d*/` folder includes a series of `<densityName>_t\d{5}.txt/` files where `t\d{5}` denotes the simulation timestep at which the concentration of the corresponding density is reported. Finally, the `output` folder, contains text files named `/cells_\d{5}.txt/`.

Simulation data samples that are being made available within the INFORÉ consortium are uploaded at project code and data platforms as detailed in Deliverable D8.2 Quality Assurance Plan, submitted at Month 6 of the project. For instance, platforms such as Confluence¹⁶ and Bitbucket¹⁷ keep versions of uploaded data.

¹⁴ <http://rd-alliance.github.io/metadata-directory/standards/>

¹⁵ <https://zenodo.org/communities/infore-project/>

¹⁶ <https://www.atlassian.com/software/confluence>

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.: WP8 D8.4
		Rev.: 1.0
		Date: 30/06/2020
		Class.: Public



3.2.2 Making data openly accessible

Life Science data have been made available at Zenodo and under the INFORE community:

<https://doi.org/10.5281/zenodo.3922263>, <https://doi.org/10.5281/zenodo.3921049>,
<https://doi.org/10.5281/zenodo.3923070>

As described in Section 3.1.2 all data are stored in ASCII files and require no special software to get read and processed. All data generation software is provided open source by the contributors. PhysiCell is developed in Paul Macklin's lab⁴. MaBoSS and PhysiBoSS are developed in the Computational Systems Biology of Cancer group at Institut Curie⁵ (Paris, France). PhysiBoSS (from the merging of PhysiCell and MaBoSS) is an adapted version of PhysiCell to integrate inside each cell a Boolean model computation. All the aforementioned frameworks constitute open-source software [2][4][6]. Moreover, for visualizing simulation outcomes the open-source tools of Paraview^{18,19} and POV-Ray²⁰ can be used. To parse the XML files, PhysiBoSS integrates in its code the Tinyxml2 files²¹, which are open sourced and freely available¹² in the code of PhysiBoSS. Finally, for easing the construction of the `init.txt` file a separate executable is provided²². All related documentation is included in the above cited references and footnotes.

To stream simulation data in the INFORE architecture in an online fashion (instead of reading PhysiBoSS output files stored on hard disk), Athena has developed code that directs the output of PhysiBoSS to Kafka [8]. Athena uses the `cppkafka` library, which is freely available on Github²³. `Cppkafka` is a version of Apache Kafka, that allows C++ applications (like PhysiBoSS) to produce and consume messages using Apache Kafka protocol. `Cppkafka`'s design is based on the `rdkafka` library²⁴. According to its creator, `librdkafka` is a C library implementation of the Apache Kafka protocol with high performance. Thus, `cppkafka` library consists a `librdkafka`'s wrapper and provides high level consumer and producer interface, as well as, some other utilities like a buffered producer (which simplifies producer error handling) and a compacted topic consumer. `Cppkafka` lets us create Kafka producers/consumers and respective configuration with very little code. Athena will examine the possibility of making this code freely available after further testing and integration with INFORE architecture.

The simulation data is reproducible using PhysiBoSS, but large-scale simulations require HPC infrastructure to be possible. There is no restriction in the Life Science data that INFORE can share, except from the fact that data used in unpublished scientific papers will be made available after the acceptance for publication of the corresponding papers.

Having mentioned the above, in INFORE we selectively make available simulation data which we judge will be most useful for target audiences described in Section 3.1.3. The maximum allowed size of a dataset at Zenodo is 50 GB, but multiple datasets can be uploaded and there is no specific limit for communities²⁵.

3.2.3 Making data interoperable

In order to make data interoperable PhysiBoSS wiki²⁶ provides metadata dictionaries for the `parameters.xml` file. Here, we provide the metadata vocabularies provided by PhysiBoSS for easy reference and we also include data dictionaries for the output files.

The `parameters.xml` file describes the various simulation parameters, expressed via respective XML tags (elements) related to:

¹⁷ <https://bitbucket.org/product>

¹⁸ <https://github.com/gletort/PhysiBoSS/wiki/Paraviewing>

¹⁹ <https://www.paraview.org/>

²⁰ <http://www.povray.org/>

²¹ <https://github.com/gletort/PhysiBoSS/tree/5fd3fcc834064c4d51fb18fef360b760dee1440f/src/tinyxml>


²² https://github.com/gletort/PhysiBoSS/wiki/PhysiBoSS_CreateInitTxtFile

²³ <https://github.com/mfontanini/cppkafka>

²⁴ <https://github.com/edenhill/librdkafka>

²⁵ <https://help.zenodo.org/>

²⁶ <https://github.com/gletort/PhysiBoSS/wiki/Parameters#-simulation->

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

<simulation> parameters that refer to global properties of the simulation, e.g. the numerical time step, are included in the simulation parameters XML element. Table 1 provides the metadata dictionary related to child elements of the <simulation> XML element.

Table 1: Metadata Dictionary of Simulation Parameters

Parameter name	Default value (range)	Description
time_step	0.01 min (0.001-1)	Diffusion time scale (smallest time scale). If no entities are diffusing, can be higher. From here ²⁷ .
mechanics_time_step	0.1 min (0.01-5)	Time scale of motion, cell volume changes. From here ²⁷ .
cell_cycle_time_step	6 min (0.5-10)	Cell cycle time scale change of cell phase. From here ²⁷ .
maximal time	4320 min	Duration of simulated time
output_intervals	60 min	Frequency at which cells position and states are written to file.
output_densities	600 min	Frequency at which microenvironment densities concentration are written to file.
write_passive_cells	0 (0 or 1)	If write the position of passive cells in output files or not
write_ratio_voxels	0.5 (0-1)	Proportion of microenvironment voxel values to write to output files (writing all of them can be pretty heavy)
number_of_threads	10	Number of threads for parallel computing. Depends on the machine used.
friction_passive_cells	0.0001 (0.00001-1)	How easy will it be to move passive cells (if high, fixed cells).
mode_cell_cycle	0 (0 or 1)	Mode of calculation of the cell cycle: 0, cycling is defined as in ²⁷ . 1: cycling is dependent on the boolean network outputs.
number of densities	>= 1	Number of diffusing entities. By default, there is just one (oxygen).
density_0	oxygen	Name of the diffusing densities. Increment the indices in "density_0" to define more than 1 (e.g. <density_1> tnf </density_1>)...
bounding_box_xmin	-200 μm (10-10000)	Definition of the boundary box surrounding all the simulation space
bounding_box_xmax	+200 μm (10-10000)	Definition of the boundary box surrounding all the simulation space
bounding_box_ymin	-200 μm (10-10000)	Definition of the boundary box surrounding all the simulation space
bounding_box_ymax	+200 μm (10-10000)	Definition of the boundary box surrounding all the simulation space
bounding_box_zmin	-200 μm (10-10000)	Definition of the boundary box surrounding all the simulation space
bounding_box_zmax	+200 μm (10-10000)	Definition of the boundary box surrounding all the simulation space
minimum_voxel_size	30 μm (1-50)	Spatial discretisation of the microenvironment. Length of one side of the cube (voxel).
svg_coloring_mode	0 or 1	How to color the cells in svg output: 0, cytoplasmic radius and nuclear radius are colored differently; 1, cells are colored according to their phase


<cell_properties> common to all cells of one cell line are given. This XML element is repeated for each cell line, i.e., depending on how many cell lines the simulated cell population is composed of, this XML element appears an equal amount of times. Table 2 provides the metadata dictionary related to child elements of the <cell_properties> XML element.

²⁷ <https://github.com/sysbio-curie/PhysiBoSS/wiki/Parameters#-references->

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

Table 2: Metadata Dictionary of Cell Properties Parameters

Parameter name	Default value (range)	Description
motility_amplitude_min	5 (0.0001-20)	Mobility of the cell, varying from min value to max according to cell's internal state. Empirical unit: 0.0001 very slow (fixed), 20 fast motion
motility_amplitude_max	5 (0.0001-20)	Mobility of the cell, varying from min value to max according to cell's internal state. Empirical unit: 0.0001 very slow (fixed), 20 fast motion
polarity_coefficient	0 (0-1)	Amount of polarization of the cell: if 0, movement is totally random. If 1, movement is totally defined by the cell's current polarity axis
persistence	0.5 (0.0001-1000)	Persistence of the polarity axis: when value is high, cell will keep its polarity axis constant for a long time. For low value, it will aligned to its previous velocity very fast
mode_motility	0 (0 or 1)	How motility is calculated. If 0, motion is always random. If 1, motion is biased towards its polarity axis according to its polarization coefficient
homotypic_adhesion_min	0.17 (0-10)	Cell-cell adhesion coefficient, for cells of the same cell line (0 means cells ignore each other). Discussed in ²⁷ . The value of the coefficient varies from min to max value according to its internal state (recruitment of cadherins)
homotypic_adhesion_max	0.17 (0-10)	Cell-cell adhesion coefficient, for cells of the same cell line (0 means cells ignore each other). Discussed in ²⁷ . The value of the coefficient varies from min to max value according to its internal state (recruitment of cadherins)
heterotypic_adhesion_min	0.17 (0-10)	Cell-cell adhesion coefficient, for cells of different cell lines (0 means cells ignore each other). Discussed in ²⁷ . The value of the coefficient varies from min to max value according to its internal state (recruitment of cadherins)
heterotypic_adhesion_max	0.17 (0-10)	Cell-cell adhesion coefficient, for cells of different cell lines (0 means cells do not attract each other). Discussed in ²⁷ . The value of the coefficient varies from min to max value according to its internal state (recruitment of cadherins)
cell_cell_repulsion	10 (0-50)	Hard-core repulsion coefficient of 2 cells that overlap. Discussed in ²⁷ .
ecm_adhesion_min	0 (0-50)	Strength of cell adhesion to the ECM (depends also on the local ECM density). 0 means it does not adhere to the matrix. The value of this coefficient varies from min to max value according to its internal state (recruitment of integrins)
ecm_adhesion_max	0 (0-50)	Strength of cell adhesion to the ECM (depends also on the local ECM density). 0 means it does not adhere to the matrix. The value of this coefficient varies from min to max value according to its internal state (recruitment of integrins)
ecm_degradation	0 (0-50)	Coefficient of cell degradation of the ECM (activity of the MMPs). Empirical unit
cell_basement_membrane_repulsion	10 (0-100)	Hard-core repulsion between membrane (if defined) and a cell. If 0, the cell ignores the outer membrane (can cross it). Discussed in ²⁷ .
cell_ecm_repulsion	10 (0-100)	Hard-core repulsion between ECM and a cell. If 0, the cell ignores the ECM (can cross it). Similar to cell-basement membrane repulsion coefficient
max_interaction_factor	1.4 (1-2)	Factor of distance (compared to cell radius) until which a cell can reach (with filopodia for example). Discussed in ²⁷ .
contact_cell_cell_thresh	1 (0-5)	Threshold of contact from which cell will be considered as

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public




Parameter name	Default value (range)	Description
old		having neighbor, for input of the boolean network. If 0, then the input "Neighbor" of the network will always be 1
contact_cell_ECM_threshold	1 (0-5)	Threshold of contact between the cell and the ECM density from which cell will be considered as adhering to the matrix, for input of the boolean network. If 0, then the input "ECM" of the network will always be 1
protein_threshold	1 (0-5)	Threshold of amount of <i>protein</i> bound to the cell (internal accumulator) above which the cell is considered to internalize this protein (input of the corresponding entity will be 1 in the boolean network). The name <i>protein</i> can be used to set the default for all densities. The name of a particular density can also be precised (e.g. <i>tnf_threshold</i>)
cell_radius	8.5 μm (1-30)	Radius of a cell, just after division (minimal radius). Default value corresponds to a 3T3 fibroblast cell
nucleus_radius	5 μm (0.5-10)	Radius of the nucleus of the cell, just after division
fluid_fraction	0.75 (0-1)	Fraction of the cell that is liquid vs solid. See in ²⁷ .
phenotype_number	0 (0;1;2;3;4)	Choice of cell phenotype (definition and properties of the cell phases). Available choices are: 0, MCF7 like cell as in ²⁷ ; 1, MCF7 with fast deaths; 2, MCF7 with a cycle duration of 24h; 3, 3T3 like cell; 4, G0 cells (cells cannot change phase, stay quiescent)
secretion_rate	0.1 fg/cell/min (0-5)	Speed of a density production by a cell when the cell secretion is active. This value was chosen from a range of plausible value for TNF in ²⁷ .
uptake_rate	0.0025 /voxel/min (0-0.1)	Quantity of density (e.g. TNF) consumed by a cell when it is available. Empirical value

<network> properties related to the Boolean network computation describing the cell's signaling states. The majority of the parameters concerning the Boolean network structure are defined separately in MaBoSS network files, with the MaBoSS conventions in .cfg and .bnd ASCII files stored in the BN sub-folder inside the root folder. In the <network> XML element, however, the parameters that can be additionally defined are those specific to PhysiBoSS simulations: update time step of the network and the definition of mutation which is specific to each cell line. The same MaBoSS configuration (bn_conf.cfg and bn_nodes.bnd) files are used for all cell lines, but the parameters as the transition rates can be varied across cell lines to simulate mutations in specific genes up (or down)-regulations.

Table 3: Metadata Dictionary of Network Parameters

Parameter name	Default value (range)	Description
network_update_step	10 min (0.1-100)	Frequency of updates of the network state. Compute a given number of MaBoSS iterations (defined in MaBoSS files) every network_update_step times.
mutation_0		Definition of a mutation: change the transition rate for one gene up(down)-regulation for one cell line
mutation_0->symbol_name	e.g. \$High_IKK [4]	Defines the transition rate value to change. Must be defined in the MaBoSS network file with the exact same name
mutation_0->cell_line	0 (0,1,2,...)	Defines for which cell line this value of the transition rate is defined
mutation_0->rate	0.0 (0-1 in general)	The value of the transition rate (0 will never happen, 1 is default speed)
network_update_step	10 min (0.1-100)	Frequency of updates of the network state. Compute a given number of MaBoSS iterations (defined in MaBoSS files)

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

Parameter name	Default value (range)	Description
		every <code>network_update_step</code> times.
<code>mutation_0</code>		Definition of a mutation: change the transition rate for one gene up(down)-regulation for one cell line
<code>mutation_0->symbol_name</code>	e.g. <code>\$High_IKK</code> [4]	Defines the transition rate value to change. Must be defined in the MaBoSS network file with the exact same name
<code>mutation_0->cell_line</code>	0 (0,1,2...)	Defines for which cell line this value of the transition rate is defined
<code>mutation_0->rate</code>	0.0 (0-1 in general)	The value of the transition rate (0 will never happen, 1 is default speed)
<code>network_update_step</code>	10 min (0.1-100)	Frequency of updates of the network state. Compute a given number of MaBoSS iterations (defined in MaBoSS files) every <code>network_update_step</code> times.

<initial_configuration> of the population can be given separately describing the state of the cells (the position of all initial cells, their state, size etc). As already discussed, with PhysiBoSS code, an executable `PhysiBoSS_CreateInitTxtFile` can be given, which allows the user to create the `init.txt` file containing this information for given parameters and chosen modes. This file is referenced in the <initial_configuration> element. The output of a simulation can also be used as the input initial file to another simulation.

Table 4: Metadata Dictionary of Initial Configuration Parameters

Parameter name	Default value (range)	Description
<code>sphere_radius</code>	100 μm (10-10000)	If no initial file is provided, length of the initial sphere of cells to create. To choose from experimental conditions to mimic.
<code>load_cells_from_file</code>	<code>init.txt</code>	Name of the file containing the initial position of all the cells
<code>create_ecm_from_file</code>	<code>ecm.txt</code>	Name of the file containing the initial value of ECM concentrations (must contains a list of position (x,y,z) and density value).
<code>number_of_passive_cells</code>	0 (0-100000)	If no initial file is provided, number of passive cells to add (around the active cells sphere).
<code>time_passive_cells</code>	1000000 min (0-maxtime)	Change the repulsive capacity of the passive cells during the simulation, after the time given (before not repulsing=phantom cells, after repulsing cells=obstacle). By default, value above max time of simulation: no change
<code>mode_injection</code>	-1 (-1 or 0 or 1)	How injection of a density is simulated. Options are: -1, no injection; 0, injection in all the voxels of the space; 1; injection only in the voxels at the outer boundary of the space
<code>time_add_densityName</code>	1000000 min (0-maxtime)	Time interval between density injection (e.g. TNF injection) (default value is higher than simulation maximal time, so no injection)
<code>duration_add_densityName</code>	10 min (0-maxtime)	Duration of the pulse injection when there is injection to do
<code>time_remove_densityName</code>	1000000 min (0-maxtime)	Time at which to clear the microenvironment of all the density (e.g. TNF) (corresponding by example to a washout/change of solution)
<code>densityName_concentration</code>	7.2 $\text{fg}/\mu\text{m}^3$ (0-100)	Concentration of injected density (name of the density followed by <code>_concentration</code> : e.g. <code>tnf_concentration</code>)
<code>membrane_shape</code>	none	Addition of an outer "membrane" that constrains all the cells inside this geometry. Can be: none, duct or sphere
<code>membrane_length</code>	0	If a shape is defined, length of the geometry (e.g. radius for the sphere)

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



Regarding the output data and respective dictionaries, as already discussed, an output and microuput sub-directories are created in the root directory of the simulation.

A `report.txt` file is created inside the root folder giving a quick summary of the simulation, with the number of cells that divided or died in between (defined in the input XML file) output times. This file is simple to interpret as shown in the following exemplary content (first row includes headers):

```

simulated time num cells numdivision numdeath walltime oxygen tnf
0 1137 1137 0 0.0251834 18999 274.461
30 1228 91 0 22.5977 0.000994473 28.8518
60 1240 12 0 14.1297 5.12626e-11 0.484111
90 1246 6 0 13.4879 2.64152e-18 0.0439171

```

In the folder `output`, text files named `/cells_{d}.txt/` are generated during the simulation, containing the current cells states (position/size/cycle state). One such file is created per simulation time point with the results of each of the studied variables. This semicolon-separated file has a header row with the names of the variables and one row for each agent (cells, in this case).

Table 5: Data Dictionary for cell state in `/cells_{d}.txt/` files


Output variable name	Description
Time	Simulation time
ID	Cell-agent unique identifier
x	Domain coordinate (X)
y	Domain coordinate (Y)
z	Domain coordinate (Z)
radius	Float value indicating the radius of the corresponding cell-agent (μm^3) at a given simulation time
volume_total	Float value indicating the volume of the corresponding cell-agent (μm^3)
radius_nuclear	Float value indicating the nuclear radius of the corresponding cell-agent (μm^3)
contact_ECM	A boolean value indicating if the cell-agent is attached to the extracellular cell matrix (ECM)
freezer	Indicates the level of cell freezing: 0 not, 1 cannot change volume, 3 (for bitwise operations) cannot change volume and move
polarized_fraction	A number (p) between 1 and 0, showing how polarized the cell is
motility	Percentage of current motility amplitude (evolve between 0 and 1)
cell_line	Integer identifier pointing to the cell line type to which the cell-agent belongs (e.g. 0 correspond to the default MCF7 cell line)
Cell_cell	Integer indicating the total number of contacts with other cells
phase	The current phase of the cell cycle phase in which the cell-agent is at current simulation time
Cycle	Integer identifier pointing to the cell cycle model attached to the agent.
NFkB	NFkB (nuclear factor kappa-light-chain-enhancer of activated B cells) is a protein complex that controls transcription of DNA, cytokine production and cell survival.

Example of output file for cells at time 0 (first row includes headers):

```

Time;ID;x;y;z;radius;volume_total;radius_nuclear;contact_ECM;freezer;polarize
d_fraction;motility;cell_line;Cell_cell;phase;Cycle;NFkB
0;0;-46.758;-10.7294;-
85.806;10.0174;4210.69;6.02332;0;0;0.1;0.01;0;2.65594;0;0;-1

```

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



```
0;1;-46.5751;7.86895;-  
82.8054;9.81311;3958.3;5.90048;0;0;0.1;0.01;0;3.66865;0;0;-1  
0;2;-31.2033;-37.4872;-  
84.2829;9.5;3591.36;5.71221;0;0;0.1;0.01;0;2.99975;1;0;-1
```

Similarly, time-specific, semicolon-separated text files are built for the different densities (free-roaming molecules on the extracellular space, such as oxygen, signalling molecules, microenvironment density, etc) in the microuput folder. An exemplary output file for micro-environment density at time 0 is given below. The first three columns correspond to spatial coordinates and the fourth to the value of the density (no header line):

```
-417.5;-492.5;-492.5;0.0630239  
-357.5;-492.5;-492.5;0.0630185
```

The above metadata and data dictionaries facilitate interoperability, but to facilitate inter-disciplinary interoperability we also cite references to scientific papers, links and other sources discussing basic concepts that are relevant to the data.

3.2.4 Data re-use (through clarifying licenses)

The Life Science data space is made publicly available under FOSS licensing as described in Deliverable D7.3 Initial Business and Exploitation Plan, Dissemination Report, submitted on Month 18 of the project.

The data that have already been uploaded at Zenodo will be made incrementally updated with new contributions as soon as the consortium partners assess that a new data portion relevant to the target audiences (see Section 3.1.3) can be made available.

Since the Life Science data are generated in a controlled simulation environment the produced data are of high quality with no particular lack of veracity. There may be incomplete data in cases when simulations were abruptly stopped either as non-useful or due to software (e.g. lack of memory) or hardware (e.g. connectivity to workers) issues.

There is no restriction to the use of data by third parties even after the end of the INFORE project. The data are made available to allow others join forces and move forward the research in the area. As already noted, data out of which novel scientific results are reached, will be made available when the corresponding papers get accepted for publication or a technical report is made publicly available to repositories such as Zenodo and ArXiv.

3.2.5 Data security

Data recovery as well as secure storage and transfer of data procedures follow the procedures described in the technical guidelines of the PRACE project²⁸ for the INFORE consortium and MareNostrum 4 user's guide²⁹ for BSC partner.

There is no sensitive data used in the project (see also Section 6).


3.2.6 Allocation of resources

For Life Science data, costs for servers used throughout the duration of the project for scientific research are covered as "other costs" in the overall project budget. These include the costs of data curation and preservation prior or after data sharing. Since this data constitutes informational wealth of BSC and its scientific mission, useful simulation data will be preserved and curated by BSC at own costs after the end of the project.

The Project Coordinator is responsible for data management issues throughout the project's lifespan.

²⁸ https://prace-ri.eu/wp-content/uploads/Technical_Guidelines_Call_21.pdf

²⁹ <https://www.bsc.es/support/MareNostrum4-ug.pdf>

 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



4 Financial Data and Management Procedures

4.1 Financial Data Summary

4.1.1 Origin of data – Purpose of data collection – Relation to the objectives of the project

The goal of the financial data analysis in INFORE is to analyse high speed stock market data streams, train machine learning models that extract valid patterns used to perform real-time suggestion and forecast of investment opportunities, systemic risk (i.e., great linkage between major market participants) prediction and price swings. Indicatively, to detect systemic risks, i.e., stock level events that could trigger instability or collapse of an entire industry or economy, requires discovering and interactively digging into correlations among thousands of stock streams. The problem involves identifying the highly correlated pairs of stock data streams under various statistical measures, such as Pearson’s correlation over N distinct, high speed data streams, where N is a very large number. To track the full correlation matrix results in a quadratic explosion in space and computational complexity which renders interactivity, for very large N, infeasible. The problem is further exacerbated when considering higher-order statistics (e.g., conditional dependencies/correlations). Similarly, stock correlations play an important role in order to timely identify and suggest investment opportunities. For instance, having up-to-date information on clusters of correlated stocks, one may use this information to choose to invest on other stocks in the same cluster when one or more of the cluster members exhibit an increasing trend.

The idea in the scope of INFORE is to use the Graphical Editor Component of the INFORE architecture and draw workflows specifying the steps of the real-time stock market streams’ analysis. These workflows will be automatically optimized by the INFORE optimizer and get deployed by the Manager Component of the architecture. The workflows make use of the online machine learning and forecasting facilities and heavily use the Synopsis Data Engine Component of the architecture (see Deliverable D6.1 submitted on Month 12 of the project) in order to ensure interactivity and real-time monitoring of risks and investment opportunities.

Real-time and near real-time data from the financial area include:

- Foreign exchange rates
- Futures on indices and commodities
- Bond markets
- Stocks from worldwide exchanges and market indices.

Additional sources for financial data may also involve international interest rates and cryptocurrencies.

The Financial data streams in the scope of INFORE are collected via the Data API³⁰ (see Figure 6) that has been developed and owned by the Spring Techno partner, before and independently of the current project. As shown in Figure 1 (upper right part of the API in the figure), the user can start/stop monitoring any listed or the entire list of supported stocks using the Start Quote and Stop Quote buttons. Moreover, market depth data can also stream in by using the Start Depth button. Each quote arrives as a separate stream and, as will be explained later on in this section, the collected streams involve Level 1³¹ (termed quote) and Level 2³², Level 3³³ (termed depth) data.


Besides collecting real time data streams, Spring Techno has also made available to the project a repository of quote and historical stock data. The description of these data follows in the upcoming section. At its current state the repository includes quotes monitored from the start of the project and historical data from the past 17 years. Although historical data are not part of the streaming analytics INFORE wishes to provide, they can be useful in incorporating knowledge, such as seasonalities, in the real-time analysis process as well as act as an additional

³⁰ http://www.springtechno.com/J-Data_API_Description.pdf

³¹ <https://www.investopedia.com/terms/l/level1.asp>

³² <https://www.investopedia.com/terms/l/level2.asp>

³³ <https://www.investopedia.com/terms/l/level3.asp>

 <p>Project supported by the European Commission Contract no. 825070</p>	<p>WP8 T8.3 Deliverable D8.4</p>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

dataset for serving experimentation purposes for research and papers developing novel ideas in the scope of the project.

Upon being analysed, the raw data described above will yield derived datasets such as time evolving clusters or correlations of stocks. Such data are already being generated as results of research papers in the scope of the project [9] and will be considered in Deliverable D8.6 (Data Management Plan V3) at Month 36 of the project.

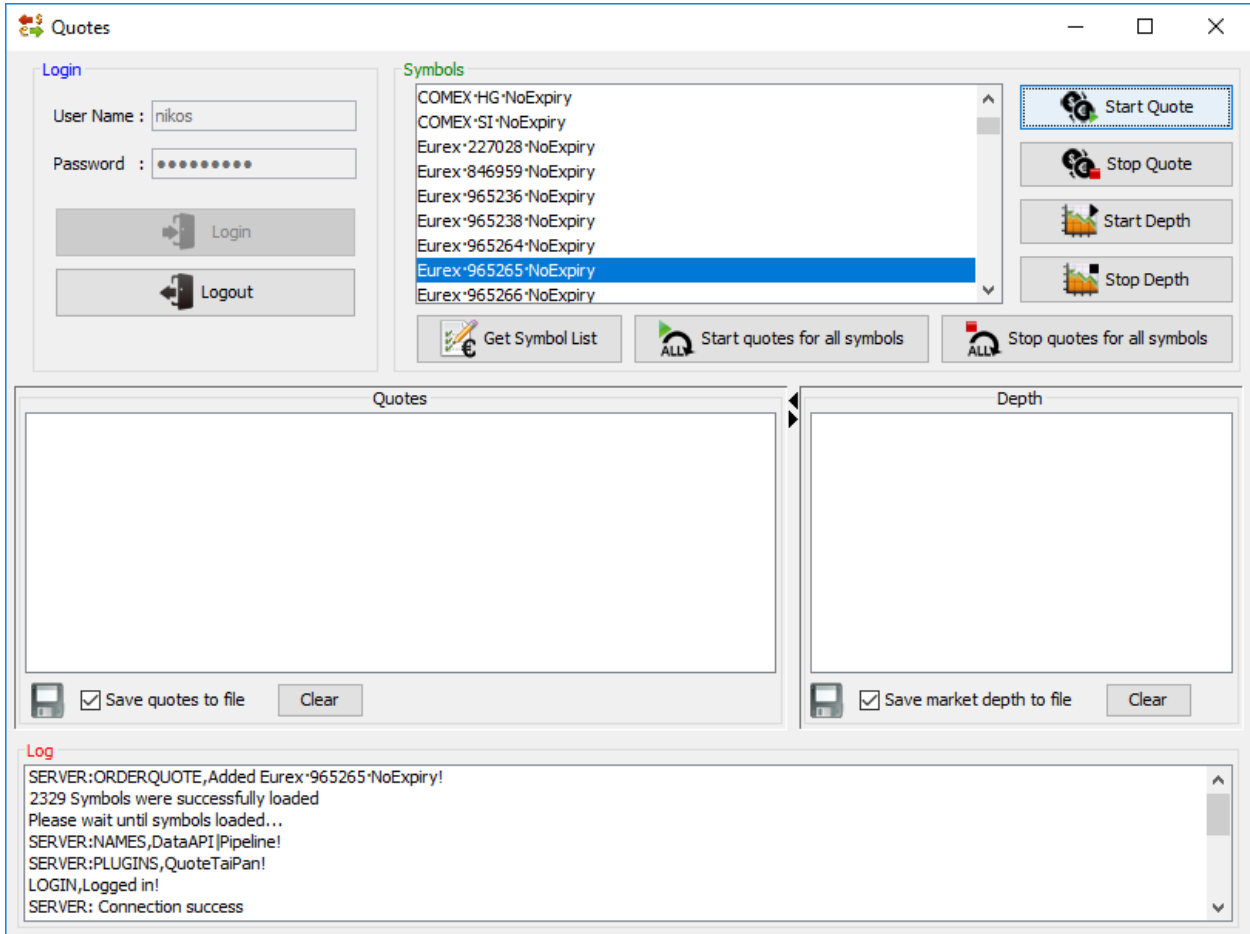


Figure 6: Financial Data Collection API

4.1.2 Types and format of collected data – Reused data – Expected data size

The data of each stock that is being fed using the Data API is stored in separate text files with an integer identifier corresponding to the position of the symbol in the list, shown in the top-middle part of Figure 6. The .txt files for Level 1, Level 2 and Level 3 data are kept in separate folders. These files are automatically assigned a naming at the end of the day and Athena partner has developed code which loads the monitored streams into Kafka³⁴ topics for real time analysis purposes.

³⁴ <https://kafka.apache.org/>

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

Name

- 25.txt
- 31.txt
- 33.txt
- 95.txt

Date (MM/DD/YYYY), Time (HH:MM:SS), Price, Volume
08/23/2019,11:53:24,1.541,660

Figure 7: Level 1 Data from the Data API

The financial data repository, which Spring Techno has made available to the project, includes two root folders named `quotes` and `history` for quote and historical data respectively. All files in these folders are ASCII, text files, though the file extension may differ to `.his` for historical data, `.min` for condensed quote data (explained shortly) and `.txt` for quote tick data as those originating from the Data API.

history folder: Figure 8 shows the organization and file types in the history folder. The data in this folder is organized in a number of `.zip` files, each containing a `.his` file. In each `.his` file a time-based compressed form of quote data exists which are condensed to specific time frames. As shown in Figure 8, file records are of the form “Date, Time, Open, High, Low, Close, Volume”. The meaning of these fields will be explained in Section 3.2.3. Each folder contains data in the granularity denoted in its naming. For instance, 201901 shown in the figure says that the `.his` file includes tuples for January 2019 for a specific exchange and stock.

201901_Forex.CAD.NoExpiry.zip
201901_Forex.CADCHF.NoExpiry.zip
201901_Forex.CDF.NoExpiry.zip

201901_Forex.CAD.NoExpiry.his

```
01/01/2019,00:01:00, 1.3637, 1.3637, 1.3637, 1.3637, 0
01/01/2019,00:02:00, 1.3637, 1.3637, 1.3637, 1.3637, 0
01/01/2019,00:03:00, 1.3637, 1.3637, 1.3637, 1.3637, 0
```

Date (MM/DD/YYYY), Time (HH:MM:SS), Open, High, Low, Close, Volume

Figure 8: history folder and file organization. The reason for the zero volume value is that, for Forex markets no volume is available, as the trading of currencies is de-centralized and therefore not all volumes can be accessed.

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

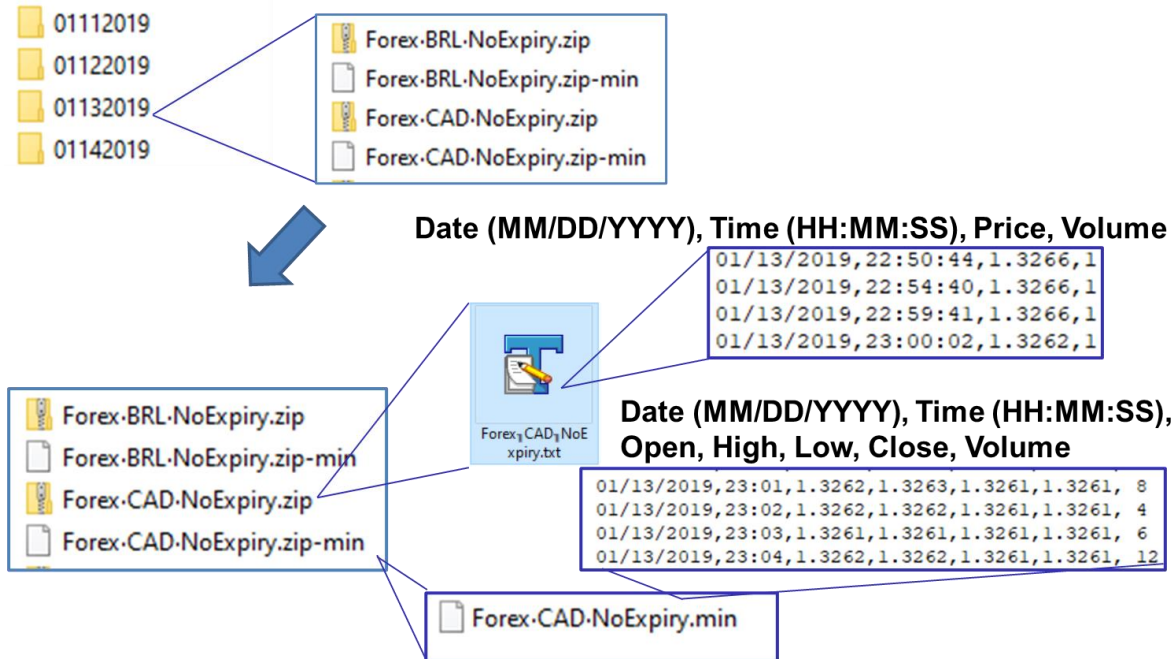


Figure 9: quotes folder and file organization. The reason for the volume value of 1 in the .txt file shown in the figure, is that, for Forex markets no volume is available, as the trading of currencies is de-centralized and therefore not all volumes can be accessed.

quotes folder: The file organization and structure in the quotes folder is illustrated in Figure 9. Inside the quotes folder there is a number of sub-folders, one per recorded day of trades as shown at the top-left part of the figure. For each day and for each stock (symbol) of an exchange (market) there is a couple of .zip and .zip-min files. In each .zip file there is a .txt with Level 1 data composed of “Date”, “Time”, “Price”, and “Volume” of the trade represented by the corresponding tuple. The meaning of these fields is explained in Section 4.2.3 as well. In each .zip-min file there is a .min ASCII file with contents equivalent to those of the .his files as shown at the bottom of Figure 8, but this time at the granularity of minutes of the day (at the bottom of Figure 9).

Finally, related to the size of the accumulated data that need to be analysed in an online fashion: depending on the number of stocks, markets and depth that are being monitored, the collected data may amount to 450GB/day. INFORE must support at least 500 stock market messages per second per studied market player under normal load with growth rates up to a factor of 20 over normal load for stock market streams.

4.1.3 Data utility – to whom it is will be useful

Key stakeholders from the financial domain include Hedge Fund Managers, Investment Companies, Investment Banks, Trading service providers. The financial dataset is also useful to researchers in Computer Science fields related to Big Data management, databases, data mining and machine learning fields.

4.2 FAIR Financial Data

4.2.1 Making data findable including provisions for metadata

4.2.1.1 Standards for metadata creation, discoverability of data (metadata provision), approach towards search keyword

Keywords can be extracted based on the naming conventions used, which follow a regular expression-alike format. Moreover, in Section 4.2.3 we attach, among others, a dictionary that maps stock symbols with their full namings to

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



facilitate not only interoperability, but also keyword search. Spring Techno does not currently create or maintain metadata for either live financial streams or respective data placed at the shared repository.

To our knowledge there is no standard for financial metadata that matches the collected datasets and there is no central regulatory authority that imposes the implementation of a specific framework. There have been efforts in creating ontologies for relevant data, such as [10] , but these still do not match our datasets.

Therefore, we follow the basic approach for keeping metadata as described in Stanford Libraries³⁵. In a nutshell, this means that technical partners create and version metadata for the financial datasets as they work on them for scientific publications [9] and within the scope of the WPs of the project.

The key is to collect all the necessary information (metadata) as we work and then link that metadata to the data files themselves. This will help later in referring back to these files and it will make the structuring of metadata into a future standard easier.


In each folder of the shared repository that the technical partners have worked on, we include a .docx or .pptx file that describes the contents of the files in that folder. Explanations of abbreviations and column headers in all files are also provided within the current document which is itself part of this endeavor. We also include references to publications, such as [9] , that use and describe the data.

Below we provide a description of the metadata which Athena has created for the quotes folder of the shared repository described throughout the current section.

Table 6: Metadata Creation for the quotes folder

Dataset Name	Financial Data – quotes folder
Project	INFORE
Date of Metadata Creation	December 2, 2019
Version	1.0
Creator	Antonis Kontaxakis (ATHENA)
Relevant publications - Experiments	Antonis Kontaxakis, Nikos Giatrakos, & Antonios Deligiannakis. (2020, March 21). A Synopses Data Engine for Interactive Extreme-Scale Analytics. Zenodo. http://doi.org/10.5281/zenodo.3849978
Analysis Purpose	Stock correlation matrix computation, extraction of stock clusters and their real-time evolution, application of distributed data summarization techniques
Number of Folders	223

³⁵ <https://library.stanford.edu/research/data-management-services/data-best-practices/creating-metadata/basic-approach-metadata>

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3 Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



Number of Files	355.308																																																																																																												
Trading Period	11/01/2019 – 31/08/2019																																																																																																												
Number of Monitored Markets	~54																																																																																																												
Number of Monitored Stocks	~5000 (~4871)																																																																																																												
Number of Level 1 Updates (All Stocks)	~2.6 Billion (2589516952)																																																																																																												
Number of Updates per Stock Market /Exchange	<p style="text-align: center;">The number of Updates per Market/Exchange</p> <table border="1"> <caption>Approximate data from the bar chart</caption> <thead> <tr> <th>Market/Exchange</th> <th>Number of Updates (Approximate)</th> </tr> </thead> <tbody> <tr><td>Amsterdam</td><td>~100,000,000</td></tr> <tr><td>Athens</td><td>~100,000,000</td></tr> <tr><td>Bangkok</td><td>~100,000,000</td></tr> <tr><td>Brussels</td><td>~100,000,000</td></tr> <tr><td>Budapest</td><td>~100,000,000</td></tr> <tr><td>CBOT</td><td>~100,000,000</td></tr> <tr><td>CME</td><td>~100,000,000</td></tr> <tr><td>COMEX</td><td>~100,000,000</td></tr> <tr><td>DLB-HSE</td><td>~100,000,000</td></tr> <tr><td>DT-Börse</td><td>~100,000,000</td></tr> <tr><td>Eurex</td><td>~100,000,000</td></tr> <tr><td>Euronext</td><td>~100,000,000</td></tr> <tr><td>Forex</td><td>~1,300,000,000</td></tr> <tr><td>German</td><td>~100,000,000</td></tr> <tr><td>Global</td><td>~100,000,000</td></tr> <tr><td>Hang</td><td>~100,000,000</td></tr> <tr><td>Helsinki</td><td>~100,000,000</td></tr> <tr><td>ICE</td><td>~100,000,000</td></tr> <tr><td>Istanbul</td><td>~100,000,000</td></tr> <tr><td>Johannesburg</td><td>~100,000,000</td></tr> <tr><td>Kopenhagen</td><td>~100,000,000</td></tr> <tr><td>Lissabon</td><td>~100,000,000</td></tr> <tr><td>London</td><td>~100,000,000</td></tr> <tr><td>Madrid</td><td>~100,000,000</td></tr> <tr><td>Mailand</td><td>~100,000,000</td></tr> <tr><td>Manila</td><td>~100,000,000</td></tr> <tr><td>NASDAQ</td><td>~100,000,000</td></tr> <tr><td>NIKKEI</td><td>~100,000,000</td></tr> <tr><td>NYMEX</td><td>~100,000,000</td></tr> <tr><td>NYSE</td><td>~200,000,000</td></tr> <tr><td>Oslo</td><td>~100,000,000</td></tr> <tr><td>OTC</td><td>~100,000,000</td></tr> <tr><td>Paris</td><td>~100,000,000</td></tr> <tr><td>Prag</td><td>~100,000,000</td></tr> <tr><td>Prague</td><td>~100,000,000</td></tr> <tr><td>Russell</td><td>~100,000,000</td></tr> <tr><td>Shanghai</td><td>~100,000,000</td></tr> <tr><td>Singapur</td><td>~100,000,000</td></tr> <tr><td>Stockholm</td><td>~100,000,000</td></tr> <tr><td>Stoxx-Indices</td><td>~100,000,000</td></tr> <tr><td>Swiss</td><td>~100,000,000</td></tr> <tr><td>Sydney</td><td>~100,000,000</td></tr> <tr><td>Taipeh</td><td>~100,000,000</td></tr> <tr><td>Tokio</td><td>~100,000,000</td></tr> <tr><td>Toronto</td><td>~100,000,000</td></tr> <tr><td>US</td><td>~100,000,000</td></tr> <tr><td>Viena</td><td>~100,000,000</td></tr> <tr><td>vwd</td><td>~100,000,000</td></tr> <tr><td>Warsaw</td><td>~100,000,000</td></tr> <tr><td>Warschau</td><td>~100,000,000</td></tr> <tr><td>Wiener</td><td>~100,000,000</td></tr> <tr><td>Xetra</td><td>~100,000,000</td></tr> <tr><td>(blank)</td><td>~100,000,000</td></tr> </tbody> </table>	Market/Exchange	Number of Updates (Approximate)	Amsterdam	~100,000,000	Athens	~100,000,000	Bangkok	~100,000,000	Brussels	~100,000,000	Budapest	~100,000,000	CBOT	~100,000,000	CME	~100,000,000	COMEX	~100,000,000	DLB-HSE	~100,000,000	DT-Börse	~100,000,000	Eurex	~100,000,000	Euronext	~100,000,000	Forex	~1,300,000,000	German	~100,000,000	Global	~100,000,000	Hang	~100,000,000	Helsinki	~100,000,000	ICE	~100,000,000	Istanbul	~100,000,000	Johannesburg	~100,000,000	Kopenhagen	~100,000,000	Lissabon	~100,000,000	London	~100,000,000	Madrid	~100,000,000	Mailand	~100,000,000	Manila	~100,000,000	NASDAQ	~100,000,000	NIKKEI	~100,000,000	NYMEX	~100,000,000	NYSE	~200,000,000	Oslo	~100,000,000	OTC	~100,000,000	Paris	~100,000,000	Prag	~100,000,000	Prague	~100,000,000	Russell	~100,000,000	Shanghai	~100,000,000	Singapur	~100,000,000	Stockholm	~100,000,000	Stoxx-Indices	~100,000,000	Swiss	~100,000,000	Sydney	~100,000,000	Taipeh	~100,000,000	Tokio	~100,000,000	Toronto	~100,000,000	US	~100,000,000	Viena	~100,000,000	vwd	~100,000,000	Warsaw	~100,000,000	Warschau	~100,000,000	Wiener	~100,000,000	Xetra	~100,000,000	(blank)	~100,000,000
Market/Exchange	Number of Updates (Approximate)																																																																																																												
Amsterdam	~100,000,000																																																																																																												
Athens	~100,000,000																																																																																																												
Bangkok	~100,000,000																																																																																																												
Brussels	~100,000,000																																																																																																												
Budapest	~100,000,000																																																																																																												
CBOT	~100,000,000																																																																																																												
CME	~100,000,000																																																																																																												
COMEX	~100,000,000																																																																																																												
DLB-HSE	~100,000,000																																																																																																												
DT-Börse	~100,000,000																																																																																																												
Eurex	~100,000,000																																																																																																												
Euronext	~100,000,000																																																																																																												
Forex	~1,300,000,000																																																																																																												
German	~100,000,000																																																																																																												
Global	~100,000,000																																																																																																												
Hang	~100,000,000																																																																																																												
Helsinki	~100,000,000																																																																																																												
ICE	~100,000,000																																																																																																												
Istanbul	~100,000,000																																																																																																												
Johannesburg	~100,000,000																																																																																																												
Kopenhagen	~100,000,000																																																																																																												
Lissabon	~100,000,000																																																																																																												
London	~100,000,000																																																																																																												
Madrid	~100,000,000																																																																																																												
Mailand	~100,000,000																																																																																																												
Manila	~100,000,000																																																																																																												
NASDAQ	~100,000,000																																																																																																												
NIKKEI	~100,000,000																																																																																																												
NYMEX	~100,000,000																																																																																																												
NYSE	~200,000,000																																																																																																												
Oslo	~100,000,000																																																																																																												
OTC	~100,000,000																																																																																																												
Paris	~100,000,000																																																																																																												
Prag	~100,000,000																																																																																																												
Prague	~100,000,000																																																																																																												
Russell	~100,000,000																																																																																																												
Shanghai	~100,000,000																																																																																																												
Singapur	~100,000,000																																																																																																												
Stockholm	~100,000,000																																																																																																												
Stoxx-Indices	~100,000,000																																																																																																												
Swiss	~100,000,000																																																																																																												
Sydney	~100,000,000																																																																																																												
Taipeh	~100,000,000																																																																																																												
Tokio	~100,000,000																																																																																																												
Toronto	~100,000,000																																																																																																												
US	~100,000,000																																																																																																												
Viena	~100,000,000																																																																																																												
vwd	~100,000,000																																																																																																												
Warsaw	~100,000,000																																																																																																												
Warschau	~100,000,000																																																																																																												
Wiener	~100,000,000																																																																																																												
Xetra	~100,000,000																																																																																																												
(blank)	~100,000,000																																																																																																												

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



<p>Number of Stocks per Stock Market</p>	<p>The number of Stocks per StockMarket</p>
<p>Namings</p>	<pre>INFORE FINANCIAL DATA\quotes\^\d{2}\d{2}\d{4}\$/ INFORE FINANCIAL DATA\quotes\ /\d{2}\d{2}\d{4}\$//<Exchange>\<Symbol>(<\.Index>)?\<ExpiryDate>.zip</pre>
<p>Example Namings</p>	<pre>INFORE FINANCIAL DATA\quotes\01112019 (date) INFORE FINANCIAL DATA\quotes\01112019\Amsterdam.AALB.NoExpiry.zip INFORE FINANCIAL DATA\quotes\01112019\Amsterdam.AALB.NoExpiry.zip-min</pre>

4.2.1.2 Identifiability of data, naming conventions, clear versioning

Financial data that are shared through the INFORE page at Zenodo utilize the FAIR principles implemented by it, i.e., standard dataset identification via DOIs, new versions of shared datasets receive their own DOI and their Zenodo description is included in the link to the older version(s). DOIs, metadata and keywords at Zenodo override naming conventions with respect to making data findable.

The sample data repository that is being made available within the INFORE consortium are uploaded at project code and data platforms as detailed in Deliverable D8.2 Quality Assurance Plan, submitted at Month 6 of the project. Platforms such as Confluence and Bitbucket keep versions of uploaded data.

We now describe the naming conventions used for the Financial datasets in regular expression-alike format. Please also recall our discussion in Section 4.1.2.

quotes data namings (see also Figure 9): Each sub-folder of the quotes folder possesses a name that corresponds to a date in the form MMDDYYYY, because that folder includes all the trades of that particular day and, thus, it is named as `/^\d{2}\d{2}\d{4}$/`. For instance, in a folder named 11132019, we have all stock data for 13/11/2019.

Inside the `/^\d{2}\d{2}\d{4}$/` sub-folder, both `.zip` and `.zip-min` files as well as their respective `.txt` and `.min` file contents are named as `Exchange.Symbol<.Index>.ExpiryDate.txt` and

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	<p>Doc.nr.: WP8 D8.4</p>
		<p>Rev.: 1.0</p>
		<p>Date: 30/06/2020</p>
		<p>Class.: Public</p>



Exchange·Symbol<.Index>·ExpiryDate.min, respectively. The corresponding namings in terms of regular expression-alike naming conventions are /<Exchange>·<Symbol>(<\.Index>)?·<ExpiryDate>(?:\.txt|\.min)/. Not that the .txt and .min files have the same name as the .zip and .zip-min containers. In a nutshell, an exchange is a market where securities, commodities, derivatives and other financial instruments are traded. A stock symbol is a unique series of letters assigned to a security for trading purposes. Symbols are just a shorthand way of describing a company's stock. The expiry date stands for the expiry date of a contract. A contract is a description of the commodity. Every derivative contract, which is based on an underlying security such as a stock, commodity, or a currency, has an expiry date, though the underlying security usually does not have any expiry date. Index, e.g., NASDAQ is optional. For instance: Forex·EURTRY·NoExpiry.txt says that the data involves foreign exchange, which is a decentralized global market where all the world's currencies trade, EURTRY refers to Euro and Turkish lira, while there is no expiry date for the contract.

history data namings (see also Figure 8): the history folder contains .zip files an inside each .zip file there is an ASCII file .his with the same name. These files are named as /^\d{4}\d{2}_\<Exchange>·<Symbol>(<\.Index>)?·<ExpiryDate>.his/. The date at the beginning of the name corresponds to the month of every year for which stock data are provided, in the form of YYYYMM. For instance, 201902_Tokio·7733·NoExpiry.his says that the time frame of the condensed view refers to February 2019, the Exchange involves Tokio, the 7733 symbol refers to Olympus Corp., while there is no expiry date for the contract.

4.2.2 Making data openly accessible

Portions of Financial data have been made openly available at Zenodo and under the INFORE community: <https://doi.org/10.5281/zenodo.3886895> without further access restrictions. This includes quote and historical data and documentation structured as described in the current deliverable.

Recent financial data of similar format, as those described in this section, are of commercial value and part of current stock market analysis solutions provided by Spring Techno, before and beyond the scope of the project. Therefore, shared data include quote and historical data that are older than a month.


Financial data are kept in simple ASCII files and require no special software to get loaded and processed. The Athena partner has developed code in order to stream the tuples of the Data API to Kafka³⁴ topics, so as to later process them in Big Data platforms supported by the INFORE architecture. This code is part of the open sourced software stacks of INFORE and for Deliverable D8.6 (Data Management Plan V3) at Month 36 of the project, Athena will consider detaching that particular part of the code from the rest of the architecture, in order to make it available together with the shared financial datasets, along with appropriate documentation.

The Data API described in Section 4.1.1 and shown in Figure 6 is also a product of Spring Techno that has been developed before and independently of the project. Although regulations related to stock exchange data require them to be openly available on demand to ensure various types of transparency among stakeholders (see Section 3.5 in Deliverable D8.1 Ethics Management Plan), in order to acquire this data in real time or near real time, via the Data API, entails a fee for Spring Techno to respective data providers. Therefore, the Data API itself cannot be provided open source.

Having mentioned the above, we currently make available Level 1 quote and historical data which we judge will be most useful for target audiences described in Section 4.1.3. The maximum allowed size of a dataset at Zenodo is 50 GB, but multiple datasets can be uploaded and there is no specific limit for communities²⁵. INFORE partners are still experimenting with the utility of Level 2 and Level 3 data for the relevant workflows in the scope of the project, for instance see [9], and we will examine the possibility to make available useful portions of such market depth data together with Deliverable D8.6 (Data Management Plan V3) at Month 36 of the project.

4.2.3 Making data interoperable

In this section we first describe the practical utility of Level 1-3 data of the Financial dataset and then provide data dictionaries for each Level and of the historical data which provide a condensed, historical view of stocks' activity.

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



Level 1 is a type of trading screen used in stock trading that displays the best bid-offer-volume quotes in real time. Level 2 provides market depth and momentum statistics to traders. It is intended to provide a bird's eye-view of market action. About Level 3 data, this data has exactly the same format as the Level 1 and 2 data (Stock prices and market depth), but Level 3 means that stock data comes from different exchanges/liquidity providers for the same symbol.

³³All three levels of quotes build on top of each other. Level 1 quotes provide investors with the highest bid and the lowest ask prices for an individual stock. These types of quotes are the most common and is what private investors see when they request information from their financial services company. Level 2 quotes provide the same bid and ask information but also show the bid and ask prices for each individual market maker. This allows investors to identify the market maker with the lowest bid/ask spread, which is important for larger investors who conduct high volume and high frequency trades. Level 3 quotes provide all the information and services of Level 1 and Level 2 quotes. In addition, Level 3 quotes also grant an investor the ability to enter or change quotes, execute orders, and send out confirmations of trades. These types of quotes are reserved for registered brokers and financial institutions. Market makers, for example, participate in level 3 quotes, which allows them to execute customer orders³³.

This historical data is also provided in time-based-compressed form, where ticks are condensed to specific time frames. These files are of ASCII format and come in the form “Date, Time, Open, High, Low, Close, Volume”. “Date” and “Time” stands for the end of a specified condensed time frame. “Open” represents the first price occurred in this time frame, “High” stands for the highest price, “Low” stands for the lowest price and “Close” for the last prices of the specified time frame.

Table 7: Data Dictionary for Level 1 (tick) stock data

Field Name	Type	Description
Date	MM/DD/YYYY	Trade date.
Time	HH:MM:SS	Trade time. Granularity to the second HH:MM:SS.
Price	Number (14,8)	Trade price per contract. Up to seven (8) decimal places.
Volume	Integer (9)	Number of contracts traded.

Example of Level 1 data tuple (no header line in the corresponding files):

```
01/15/2019,09:57:41,29.21,152
01/15/2019,09:58:45,29.2,190
01/15/2019,09:58:45,29.2,177
```

Table 8: Data Dictionary for Level 2 (Quote) stock data

Field Name	Type	Description
Date	MM/DD/YYYY	Quote date.
Time	HH:MM:SS	Quote time.
Bid Price	Number (14,8)	Bid price per contract. Up to seven (7) decimal places.
Bid Volume	Integer (9)	Bid size.
Ask Price	Number (14,8)	Ask price per contract. Up to seven (7) decimal places.
Ask Volume	Integer (9)	Ask size.


Example of Level 2 data tuple (no header line in the corresponding files):

```
01/15/2019,19:47:41,2724.5:4:2719:3
01/15/2019,19:56:46,11671:1:11672:3
01/15/2019,19:57:45,2725.5: 3: 2718:8
```

The data dictionaries for Level 1 (Table 7) and Level 2 (Table 8) data also cover Level 3 which come in the form:

Providername, Symbol, Date, Time, Price, Volume

Example of Level 3 data tuple (no header line in the corresponding files):

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

Optiver, AAPL, 05/29/2020,11:34:23, 812.23,1500
 Virtu, AAPL, 05/29/2020,11:34:23, 812.24,1200

Table 9: Data Dictionary for Historical stock data

Field Name	Type	Description
Date	MM/DD/YYYY	End date of the condensed time frame.
Time	HH:MM:SS	End time of the condensed time frame.
Open	Number (14,8)	First price occurred in this time frame.
High	Number (14,8)	Highest price occurred in this time frame.
Low	Number (14,8)	Lowest price occurred in this time frame.
Close	Number (14,8)	Last price occurred in this time frame.
Volume	Integer (9)	Number of contracts traded.

Example of historical data tuple (no header line in the corresponding files):

```
01/02/2019,15:35:00,36.98,37.07,36.88,37.01,58
01/02/2019,15:40:00,37.02,37.06,36.82,36.82,81
01/02/2019,15:45:00,36.81,36.88,36.66,36.87,67
```

To further facilitate interoperability, we also provide a dictionary of, approximately 2500, symbols frequently used in the Financial datasets. The first column holds the exchange/market, the second the symbol (abbreviation) of the stock, the third column is the full stock name and the fourth the currency at which the symbol is traded. The full list of mappings is included in Attached File 1.



syms.txt

Attached File 1: Dictionary for Symbol to Full Stock Name Mapping

Example of Symbol to Full Stock Name mapping (no header line in the corresponding files):

```
Amsterdam,AALB,Aalberts NV, EUR
Amsterdam,AGN,AEGON N.V., EUR
Amsterdam,AKZA,Akzo Nobel, EUR
Amsterdam,APAM,Aperam S.A., EUR
```

4.2.4 Data re-use (through clarifying licenses)

The Financial data space is made publicly available under Creative Commons 4.0 licensing as described in Deliverable D7.4 Initial Exploitation and Business Plan submitted on Month 18 of the project.

Based on Spring Techno’s extensive experience the financial data are highly regulated and of high quality. There may occasionally be gaps due to connectivity issues with the live stream sources, but these gaps are estimated to only 0.1% of the tuples in the shared repository.

The data that have already been uploaded at Zenodo will be made incrementally updated with new contributions by the end of the project. Data out of which novel scientific results are reached, will be made available when the corresponding papers get accepted for publication or a technical report is made publicly available to repositories such as Zenodo and ArXiv. There is no restriction to the use of the Financial data, shared through Zenodo, by third parties even after the end of the INFORE project.

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public




4.2.5 Data security

Quote data are backed up on a daily basis and are kept in multiple replicas within Spring Techno’s premises. Historical data are extracted, backed up and replicated accordingly on a per month fashion. All procedures of data recovery, secure storage and transfer of data abide by good everyday business practices followed by Spring Techno. There are no sensitive data involved in the Financial dataset. In contrast, regulations related to stock exchange data require them to be openly available on demand to ensure various types of transparency (see Section 3.5 in Deliverable D8.1 Ethics Management Plan) among stakeholders.

4.2.6 Allocation of resources

Spring Techno claims costs for servers as “other direct costs” in the project budget. After the end of the project, Spring Techno will continue to preserve the data on its own expense.

 European Commission Horizon 2020 European Union Funding for Research & Innovation	Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.: WP8 D8.4
			Rev.: 1.0
			Date: 30/06/2020
			Class.: Public

5 Maritime Data and Management Procedures

5.1 Maritime Datasets – Collection Purpose – Relation to project objectives

For the purposes of the Maritime use case in INFOR's WP3, a number of historical and real-time datasets are being collected/re-used and fused in order to provide enhanced Maritime Situational Awareness (MSA), i.e. the ability to perceive and reason about activities, events and threats at sea. This use case incorporates various categories of data, analyzed later on in this section, including: AIS Raw data and AIS-derived Kafka streams as well as Patterns of Life (observable activities described as patterns), thermal camera data, acoustic sensor data and image - Copernicus (Sentinel-1, Sentinel-2) data. AIS data provide vessel identification and positioning information. However, vessels provide such information on a voluntary basis and basing MSA solely on AIS data poses barriers in case of uncooperative targets (vessels), termed as "dark targets". Dark targets hide their identity and position and, thus, respective information is inevitably missing from the MSA analysis algorithms.

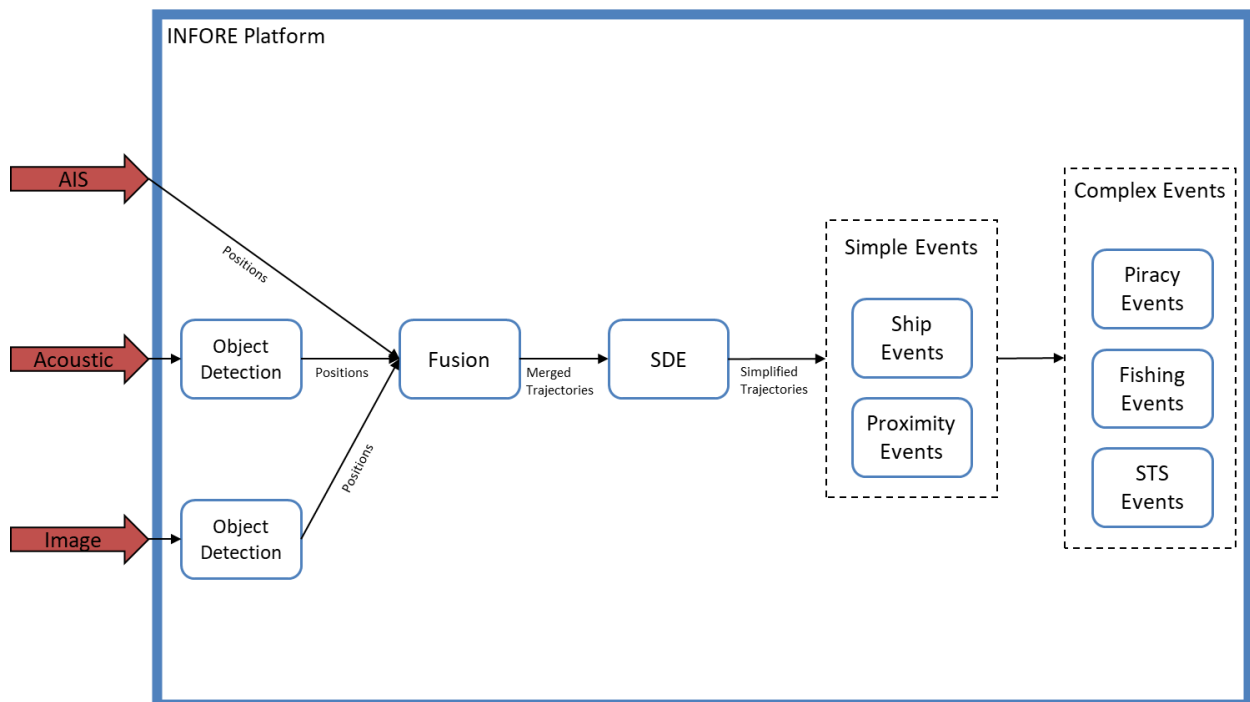


Figure 10: Maritime Data Analysis Workflow – High Level (Partial) Picture. SDE stands for Synopses Data Engine (SDE). STS stands for Ship-to-Ship Transfer Events.

Because of the above, we aim at fusing historical and online AIS data with data coming from sensors placed on autonomous vehicles that navigate at sea and collect nearby vessel information. These vehicles do not provide vessel identification information, but are equipped with acoustic and, potentially, camera sensors collecting nearby vessel data related to vessel type, size, etc. To achieve more complete and accurate MSA, one has to combine the various data sources and perform data analysis tasks in order to derive activities and events at sea. Using large-scale data fusion techniques and real-time analytics, we can also predict such events, as well as possible risks and threats and enable stakeholders to take actions to prevent them.

Figure 10 shows a high-level workflow designed to serve enhanced MSA purposes. The concept illustrated in Figure 10 is to: (a) fuse continuous/real-time and historic, heterogeneous data sources for specific, geographic areas of interest, combining AIS, acoustic and image data, (b) use the synopsis facilities of the SDE Component of the INFOR architecture to construct compact data summaries that will bring important data properties to the front and harness the volume, velocity and complexity of datasets, and then (c) detect or forecast various types of MSA events (in a two level hierarchy in Figure 10).

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



Such workflows will be graphically specified using the Graphical Editor Component of the INFOR architecture and get optimized by the Optimizer Component. The Machine Learning Component of INFOR can be used for target classification purposes, while the Complex Event Forecasting Component is to be employed to detect and/or forecast MSA events.

We further comment on the purpose each maritime dataset serves later on in this section.

5.2 Maritime Data utility – to whom it is will be useful

Maritime data are of value to maritime authorities, ship operators, brokers, maritime insurance companies, supply chain managers, data-intensive maritime businesses, maritime intelligence companies, developers working in related fields. Moreover, the target audiences regarding the maritime datasets include researchers in Computer Science fields related to Big Data management, databases, data mining and machine learning fields.

5.3 Datasets and Dataset Summary: Origin of Data – Types and format of collected/generated data – Reused data – Expected data size

5.3.1 AIS Raw Data

These data are formatted using standard Automatic Identification System (AIS) format³⁶ gathered by MarineTraffic’s proprietary, vessel monitoring network. They are available both in historical form with approximately 100 GB of data collected per day, as well as in a streaming fashion. In the streaming version of the data, each of the hundreds of thousands of vessels being monitored corresponds to a separate stream and the position of each vessel gets updated based on the class of the transponder and the moving status of the vessel itself³⁷. AIS data are being used to generate vessel trajectory density maps, infer preferred routes at sea, for vessel monitoring and anti-collision purposes among others.


MarineTraffic has set up, in its premises, a computer sub-cluster devoted to INFOR. The rest of the INFOR architecture can access streams of AIS Raw and derived data by issuing `http` requests via a RESTful API. Both requests and responses stemming from the rest of the INFOR architecture are provided in `json` snippets. The exact schema, so that interoperability with the rest of the INFOR architectural components is ensured, is still to be finalized. Details on the schema are to be provided in Deliverable D8.6 (Data Management Plan V3) at Month 36 of the project. Requests may involve individual vessels identified by a unique id each (e.g. Maritime Mobile Service Identity (MMSI)), or spatial/spatiotemporal windows.

Moreover, MarineTraffic has provided a repository of historical AIS data in order to serve research, experimentation and testing purposes in the project. MarineTraffic stores AIS Raw data in a database supporting spatial and spatiotemporal data types. Portions of these data, in the form of historical AIS positions (and port calls if required) along with vessel characteristics as described in 5.4.1.1, are exported and made available within the consortium, on demand, in `.csv` format. The source of data may be terrestrial AIS receivers or Satellite AIS receivers.

Figure 11 provides an illustration of single ground based AIS receiver vessel tracking dataset, while Figure 12 shows an equivalent illustration where individual positions have been used to construct geometries of vessel trajectories, in particular, linestrings i.e., one-dimensional objects representing a sequence of points and the line segments connecting them.

³⁶ <https://help.marinetraffic.com/hc/en-us/articles/204581828-What-is-the-Automatic-Identification-System-AIS->

³⁷ <https://help.marinetraffic.com/hc/en-us/articles/217631867-How-often-do-the-positions-of-the-vessels-get-updated-on-MarineTraffic->

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

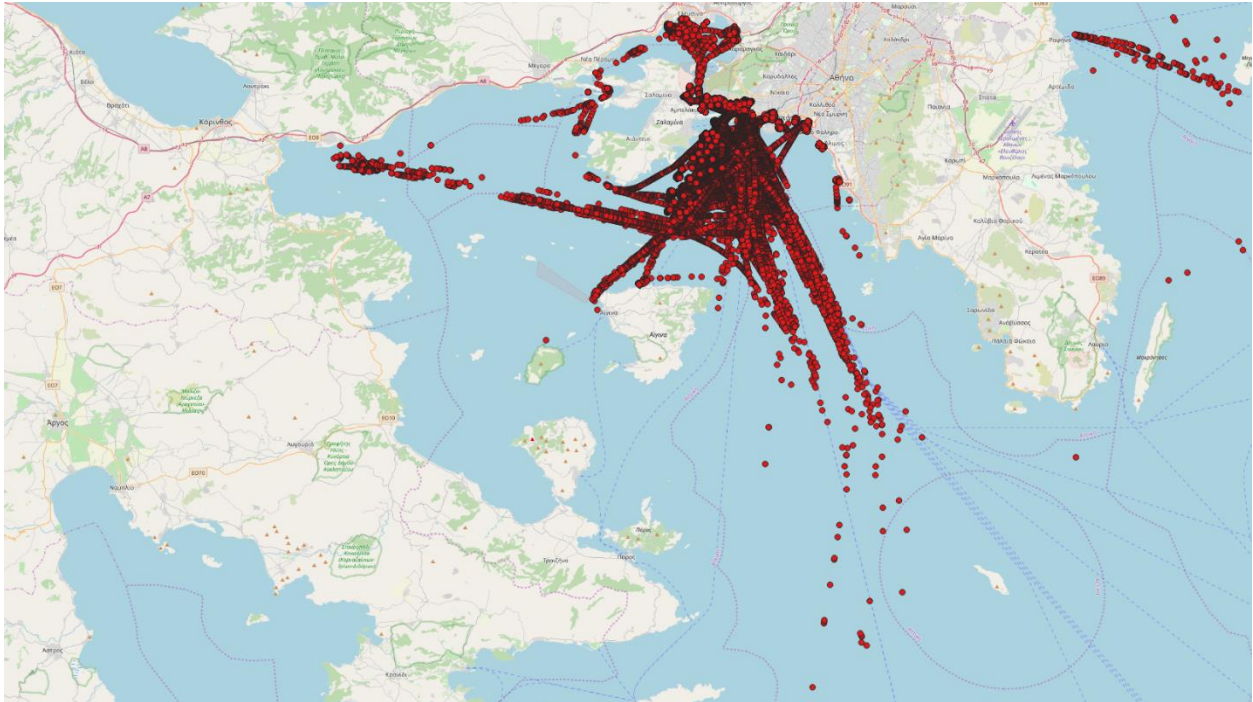


Figure 11: Illustration of AIS position signals from dataset shared at Zenodo (<https://doi.org/10.5281/zenodo.3754481>)

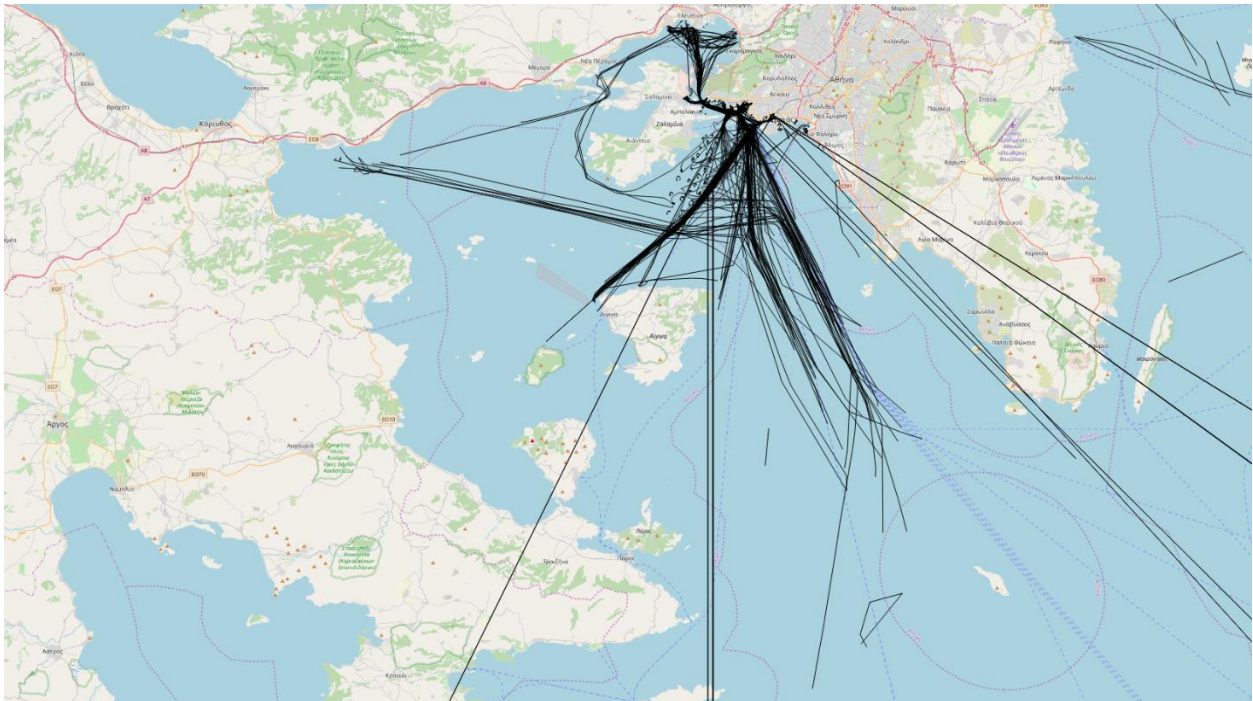



Figure 12: Illustration of linestrings (trajectories) constructed by AIS position signals (Figure 11) from dataset shared at Zenodo (<https://doi.org/10.5281/zenodo.3754481>)

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3 Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

5.3.2 Kafka Streams (AIS Derived Data Streams)

Derived data streams out of AIS streaming data are processed in Kafka³⁴ in real-time for the generation of vessel-related metrics. For instance, accurate calculation of metrics in the context of real-time detection of anomalies in a vessel's trajectory, such as the deviations in the arrival time of a vessel at a port (Estimated Time of Arrival (ETA)). These streams are updated at a sub-second rate.

Historical, structured (.csv) data of vessel statistics are produced daily. Their total volume is approximately 300 GB. These data are being used for vessel static correction purposes.

The access to these streams is technically provided in a manner similar to the one used for AIS raw data.

5.3.3 Patterns of Life

Patterns of Life (PoL) are observable human activities that can be described as patterns in the maritime domain, related to a specific action (e.g. fishing) taking place at a specified time and place. The spatial element (geometry, such as polygon) describes recognised areas where maritime activity takes place; ports, fishing grounds, offshore energy infrastructure and others, while the temporal element (timestamp or interval) often holds additional information for categorising these activities. Patterns of Life are extracted in an offline fashion (therefore only historical data are available in an annual basis) and are/can be used for anomaly detection purposes in the scope of MSA. These data are stored in a database table and their volume amounts to approximately 5 GB. Details on PoL extraction have been published in [11].

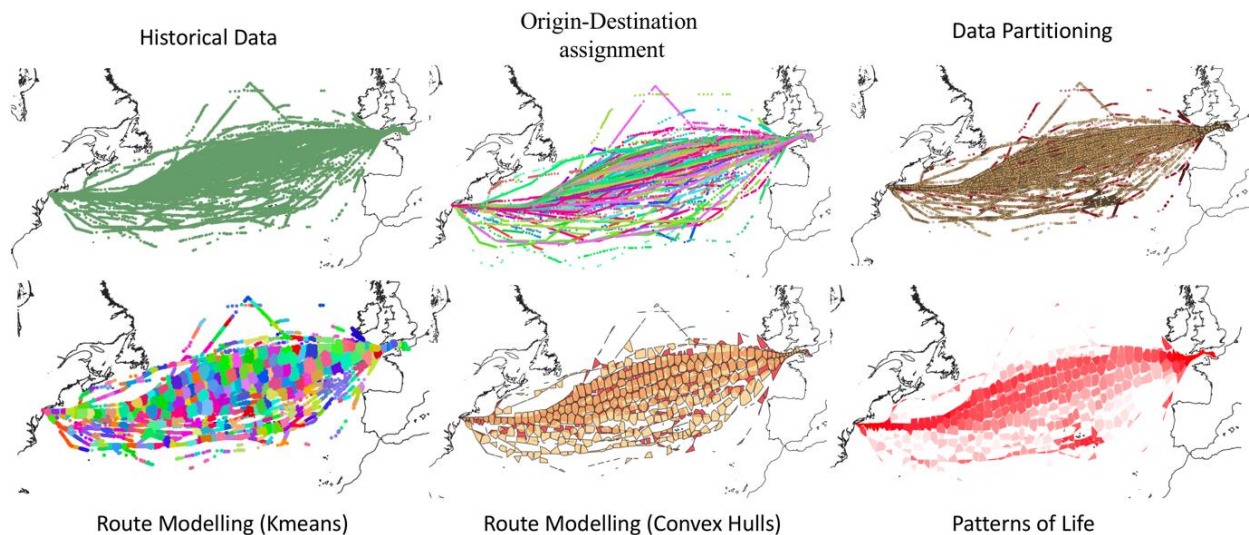


Figure 13: From Raw Historical AIS Data to Patterns of Life [11]

The collection of acoustic (upcoming Section 5.3.4), thermal camera (upcoming Section 5.3.7) and vehicle status datasets (upcoming Section 5.3.6), in INFOR was planned to take place in a maritime pilot conducted within the scope of WP3. The pilot was planned to begin in July 2020 in northern Italy (Palmaria island). Due to COVID-19 outbreak and especially due to the extent to which the broader area was affected, the initial schedule will be reformed and a report on relevant data management issues will follow in Deliverable D8.6 (Data Management Plan V3) at Month 36 of the project. For now, we work with historic acoustic data provided by CMRE which we report below along with the rest of the data that are being used or reused. For the rest of the datasets (thermal camera and vehicle status datasets) we describe the specifications according to which they will be acquired.

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

5.3.4 Acoustic Data

Acoustic data stem from hydrophones towed by autonomous vehicles. The data also include asset angles for the acoustic sensor, vehicle speed and position/heading, acoustic sensor position/heading (depth), as well as information relative to target classification. These data stream in binary format for acoustic data from hydrophones and XML format for the remainder of the data. The size of the binary data amounts to approximately 10 MB/sec, and that of XML data to 100 MB/hour of operation. These data will be used to complement AIS data for target detection, localization and activity classification purposes.

We are currently working with a historical acoustic data distribution provided by CMRE which contains acoustic data of three trials, named as: `poma13`, `collab-ngas14` and `collab13` [14] [15], respectively.

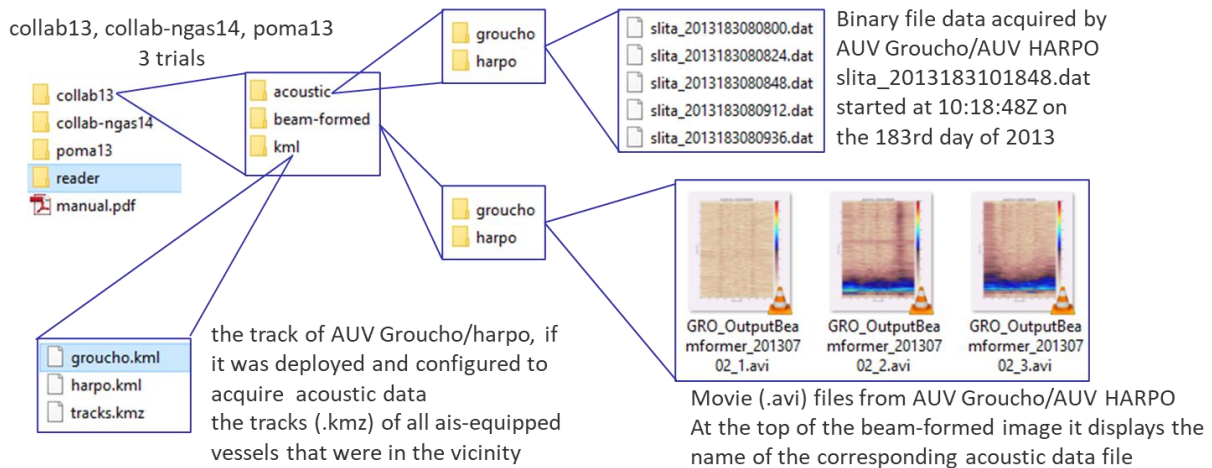


Figure 14: Acoustic data structure and description

As shown in Figure 14, each trial directory contains an `acoustic` directory. In the `acoustic` directory there may be a `groucho` directory, with acoustic data acquired by AUV Groucho, and a `harpo` directory, with acoustic data acquired by AUV Harpo. The `groucho` and `harpo` directories contain acoustic data files. The name of an acoustic data file contains the time stamp, in UTC, or Z(ULU), of the start of the acquisition. For example, the acquisition of the file `slita_2013183101848.dat` started at 10:18:48Z on the 183rd day of the year 2013. The 183rd day is July 2nd. Slita stands for SLim Towed Array (for AUV applications)³⁸.

Each trial directory also contains a `beam-formed` directory. In the `beam-formed` directory there may be a `groucho` directory, with a movie (`.avi`) of beam-formed data acquired by AUV Groucho, and a `harpo` directory, with a movie of beam-formed data acquired by AUV Harpo. The `.avi` files can be played with any suitable movie player. At the top of the beam-formed image, it displays the name of the corresponding acoustic data file, which contains a time stamp as described above.

Each trial directory contains a `kml` directory, with up to three files:

- `tracks.kmz` - the tracks of all AIS-equipped vessels that were in the vicinity.
- `groucho.kml` - the track of AUV Groucho, if it was deployed and configured to acquire acoustic data.
- `harpo.kml` - the track of AUV Harpo, if it was deployed and configured to acquire acoustic data.

The `reader` directory in Figure 14 contains Matlab/octave source code files to read the acoustic data files.

³⁸ <https://openlibrary.cmre.nato.int/bitstream/handle/20.500.12489/651/NURC-PR-2009-004.pdf>

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

5.3.5 Satellite Image Data

Copernicus data³⁹ and labelled ship target training datasets are (re)used here for target detection and classification purposes. Image (Copernicus) data come in GeoTiff, JP2 and Tiff formats. The volume of Sentinel-1, Synthetic Aperture Radar (SAR) and Sentinel-2, Multi-Spectral Instrument (MSI) data amounts between 600 MB to 2 GB per image. In addition, derived data sets will be considered, including labeled ship target data sets used to train machine learning algorithms for target classification. Labelled ship metadata are added in XML, csv, json formats with respective data volumes amounting between 50 MB to 500 MB depending on the number of targets. Further details are also provided in Section 5 of Deliverable D3.1. The total size of these data is estimated to approximately 350 GB/day.

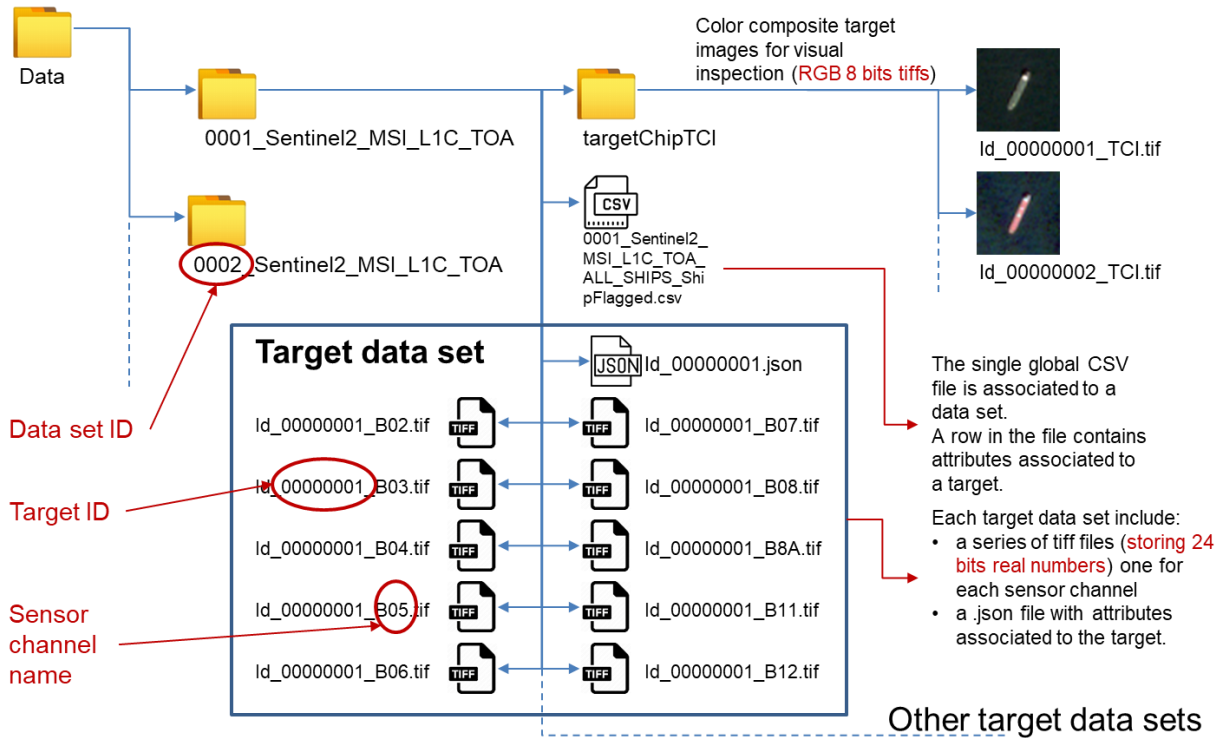


Figure 15: Satellite dataset overview and organization

The data set contains ship target images and associated metadata to be used to train and test machine learning classification algorithms. Each data set is organized in a folder that is structured like in Figure 15. The root data folder contains data set folders with a folder name following the convention:

```
/<Data set ID>_<Satellite name>_<Sensor name>_<ESA product processing level>_<type of data>/
```

For the two data folders displayed in Figure 15 (with data set ID 0001 and 0002), the Satellite name is “Sentinel2”, the sensor is the “MSI” (Multi Spectral Instrument), the processing level is “L1C” and the type of data is “TOA” (Top Of the Atmosphere reflectance).

Each data set folder contains:

- N_i target data sets, with $i = 1, \dots, M$,
- a “targetChipTCI” folder storing N_i 8 bit RGB TIFF images, one for each target for visual inspection of the data set,

³⁹ <https://scihub.copernicus.eu/>

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

- and a single `.csv` file with N_i rows, each row storing attributes metadata of a target.

Each target data set, identified by a target ID, includes:

- L 24 bits `.TIFF` images of the target, where L is the number of sensor channels,
- a `.json` file storing additional metadata of the target.

The fields of the `.csv` global metadata file are summarized in Table 12 in Section 5.4.1.3, along with all relevant metadata descriptors.

5.3.6 Vehicle Status Data

This dataset involves autonomous vehicles that are used in the scope of this use case for collecting acoustic and (thermal, sensorized RHIB/UAV) camera data which complement AIS data as explained at the beginning of this section. In particular, autonomous vehicles exploit the wave energy to move, equipped with acoustic passive sensors, and other sensors such as optical/thermal cameras for vessel data collection purposes. The format of the data is in XML streaming information relative to a vehicle’s status (position, heading, speed, battery level, next waypoint) required to supervise and control its operation. Approximately, 100 MB of data are collected for every hour of operation.

5.3.7 Thermal Camera Data

Required for target detection and classification purposes. The camera is able to provide a continuous composite video PAL or NTSC stream which can be converted in a digital stream (e.g. MPEG4 or AVI) for further processing. The volume of camera data that can stream in INFORE are estimated to approximately 108 GB/day.

5.3.8 Composite Maritime Event Data

The dataset published⁴⁰ by project partners from NCSR and CMRE [13] includes approximately 4M composite maritime events (e.g., anchored vessels, loitering, ship-to-ship transfer, etc.), recognised by RTEC⁴¹ on semantically annotated AIS position signals, over a period of six months, from approximately 5K vessels sailing around the port of Brest, France.

The dataset of AIS position signals and the semantic annotation of the AIS signals are openly available^{42,43}. Information on the specification of the composite maritime events is available in [13]. The size of the data is approximately 200MB. Below we describe the dataset with information that comes from a readme file provided together with data [16].

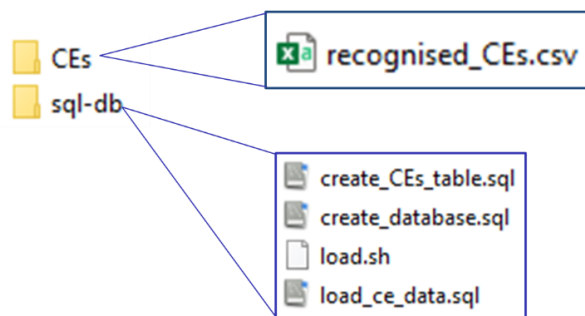



Figure 16: Structure and organization of the Composite Maritime Event Data

⁴⁰ <https://doi.org/10.5281/zenodo.2557290>

⁴¹ <https://github.com/aartikis/RTEC>

⁴² <https://doi.org/10.5281/zenodo.1167595>

⁴³ <https://doi.org/10.5281/zenodo.2563256>

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

As shown in Figure 16, there is a .csv file in `CEs/recognised_CEs.csv` which includes recognised activities in a text format, with '|' field separator. The contents of this file are described in Section 5.4.1.5. Another folder `sql-db` contains a series of scripts to load the dataset in a PostgreSQL database. The database name is by default `maritime_ces` and has a table called `ces` where the complex events detected are kept.

In order to load the recognised activities into the database, first the `<PATH-TO-DATA>` in file `load_ce_data.sql` has to get substituted with the actual path to the data file. The `load.sh` file is to be run to execute the database creation script.

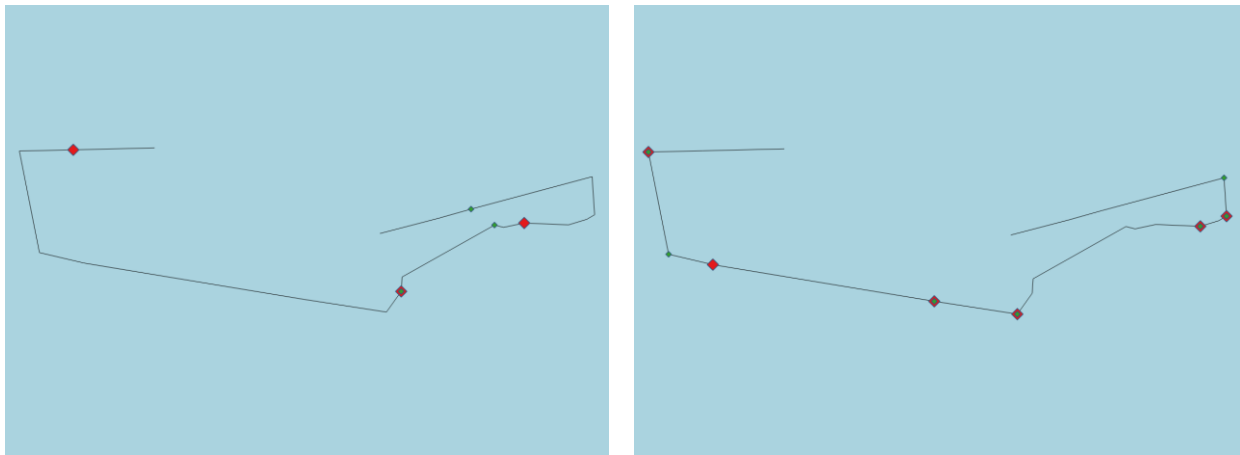


Figure 17: Illustrations of simple Turn (left) and Acceleration (right) events

5.4 FAIR Maritime Data

5.4.1 Making data findable including provisions for metadata

Portions of historical maritime datasets are shared through the INFOR page at Zenodo which implements FAIR principles including standard dataset identification via DOIs, new versions of shared datasets receive their own DOI and their Zenodo description is included in the link to the older version(s). DOIs, metadata and keywords at Zenodo override naming conventions with respect to making data findable.

Streams and samples of datasets are provided within the INFOR consortium via the sub-cluster set up by MarineTraffic for the purposes of the project. Sample data are also made available within the INFOR consortium, uploaded at project code and data platforms as detailed in Deliverable D8.2 Quality Assurance Plan, submitted at Month 6 of the project. Platforms such as Confluence and Bitbucket used in INFOR keep versions of uploaded data.

For naming conventions, were applicable, used for each dataset, please see the dataset description in Section 5.3.


5.4.1.1 AIS Raw, Kafka Streams and Patterns of Life data

For AIS Raw, Kafka Streams and Patterns of Life data that are stored in a database, we utilize the metadata creation capabilities provided by both commercial (such as SQL Server) and open-source (such as PostgreSQL/PostGIS) database managements systems. PostgreSQL/PostGIS databases⁴⁴ have their own, built-in, OGC-compliant⁴⁵ `geometry_columns` view, which can be automatically maintained by PostGIS. Microsoft SQL Server spatial databases do not have built-in OGC compliant metadata tables, but this can be implemented manually⁴⁶. In both

⁴⁴ https://postgis.net/docs/using_postgis_dbmanagement.html

⁴⁵ <http://www.ogc.org/docs/is/>

⁴⁶ <https://stackoverflow.com/questions/58673397/implementing-geometry-columns-view-in-ms-sql-server>

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3 Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

cases we can automatically create and manage the contents of the OGC standard metadata table geometry_columns in such database management systems.

AIS data are fully structured according to Recommendation M.1371 [12]. Besides geometry metadata and the timestamp assigned to each AIS message, AIS data are searchable and findable using the fields prescribed by the recommendation as shown in Table 10.

Table 10: AIS data and searchable fields description [11] [12]

Field name	Description	Range
Maritime Mobile Service Identity	Identification number for the vessel.	-
Rate of turn	Right or left turn angle of vessel.	0 to 720 degrees with minute resolution.
Speed over ground	Ship's speed measured in knots.	0 to 102 knots with 0.1 knot resolution.
Position coordinates	Vessel's latitude and longitude.	Latitude ranges from -90 to 90 and longitude from -180 to 180. Both with up to 0.0001 minutes accuracy.
Course over ground	Vessel's motion direction relative to the magnetic north pole.	0 to 359 degrees with 0.1 minute resolution.
Heading	Vessel's heading direction relative to the magnetic north pole.	0 to 359 degrees.
International Maritime Organization Number	9-digit number that is assigned by HIS Maritime (Information Handling Services) when a commercial vessel is constructed.	-
Destination	The vessel's destination that is manually inserted by crew members.	Free text up to 20 characters.
Type	The vessel type id.	0-255 code that is mapped to tis type (e.g. tanker, passenger, etc.)
Dimensions	Dimensions of ship in meters.	Four integers indicating dimension to bow, dimension to stern, dimension to port (i.e., left side of the vessel when facing the bow), and dimension to starboard (i.e., right side of the vessel when facing the bow)
Name	The vessel's name that is manually inserted by crew members.	Free text up to 20 characters.

5.4.1.2 Acoustic Data

Acoustic data use .kml descriptors which follow OGC KML schemata⁴⁷. Moreover, Table 11 describes the Slita header specification which can also be used for documentation, searchability and metadata extraction purposes.

Table 11: Acoustic data - Slita header specification

Field Name	Data Type	Description
headersize	int	Size of header in bytes
dataFormat	int	0 means 2's complement; 1 means Offset Binary
fs	float	Sampling Frequency [Hz]

⁴⁷ <http://schemas.opengis.net/kml/2.2.0/ogckml22.xsd>

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3 Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public




Field Name	Data Type	Description
inputRange	int	Flag to determine the max voltage. {3=10V, 2=5V, else 2.5V}
gainHydPreamp	float	Preamplifier gain
gainA2dAmp	float	A/D gain
dataWidth	int	Flag to determine the number of bit per sample {3=24bit, 2=20bit, 1=18bit, else 15bit}
acqLength	float	Size (in seconds) on each block of data returned by A/D
octave	int	Determine the array spacing: {1=0.21m, 2=0.42, 3=0.84, 4=1.05m}
pc_day	int	Day from PC time
pc_month	int	Month from PC time
pc_year	int	Year from PC time
pc_hr	int	Hour from PC time
pc_min	int	Minute from PC time
pc_sec	int	Seconds from PC time
gps_month	int	Month from GPS
gps_day	int	Day from GPS
gps_year	int	Year from GPS
oex_hr	int	Hour from Frontseat PC
oex_min	int	Minute from Frontseat PC
oex_sec	double	Seconds from Frontseat PC
lat_deg	int	Latitude [degrees]
lat_min	double	Latitude [minutes]
lon_deg	int	Longitude [degrees]
lon_min	double	Longitude [minutes]
heading	float	Size of header in bytes
cog	float	0 means 2's complement; 1 means Offset Binary
depth	float	Sampling Frequency [Hz]
altitude	double	Flag to determine the max voltage. {3=10V, 2=5V, else 2.5V}
sog(dm/s)	int	Preamplifier gain
sow(dm/s)	int	A/D gain
track_stat	int	Flag to determine the number of bit per sample {3=24bit, 2=20bit, 1=18bit, else 15bit}
fix_type	int	Size (in seconds) on each block of data returned by A/D
pps_output	int	Determine the array spacing: {1=0.21m, 2=0.42, 3=0.84, 4=1.05m}

5.4.1.3 Copernicus Data

Labelled ship metadata are added in XML, csv, json formats. Recall (Section 5.3.5) that each data set folder contains:

- N_i target data sets, with $i=1, \dots, M$,
- a “targetChipTCI” folder storing N_i 8 bit RGB TIFF images, one for each target for visual inspection of the data set,

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h2>Deliverable D8.4</h2>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



- and a single .CSV file with Ni rows, each row storing attributes metadata of a target.

Each target data set, identified by a target ID, includes:

- L 24 bits .TIFF images of the target, where L is the number of sensor channels,
- a .json file storing additional metadata of the target.

The fields of the .csv global metadata file are summarized in Table 12, along with all relevant metadata descriptors.

Table 12: Fields of the global metadata .csv file

Field name	Description
GlobalID	Target ID
MMSI_AIS	MMSI code from associated AIS contact
ShipType_AIS	Ship type from associated AIS contact text fields
Lat_Interpolated_AIS__deg_	Latitude of the associated AIS contact at the closest time of the sensor acquisition time [deg]
Lon_Interpolated_AIS__deg_	Longitude of the associated AIS contact at the closest time of the sensor acquisition time [deg]
TimeInterpolated_AIS	AIS time of the associated AIS contact [YYYYMMDDTh:mm:ss.sssZ]
SOG_AIS__m_s_	Ship speed over ground from associated AIS contact [m/s]
COG_AIS__deg_	Ship course over ground from associated AIS contact [deg]
Length_AIS__m_	Ship length from associated AIS contact [m]
Width_AIS__m_	Ship width from associated AIS contact [m]
Draught_AIS__m_	Ship draught from associated AIS contact [m]
NavStatus_AIS	Ship navigational status from associated AIS contact (it follows AIS conventions)
Time_SENSOR	Sensor acquisition time [YYYYMMDDTh:mm:ss.sssZ]
Lat_SENSOR__deg_	Latitude of the target centre of mass estimated from sensor data [deg]
Lon_SENSOR__deg_	Longitude of the target centre of mass estimated from sensor data [deg]
PRODUCT_URI	Name of the input image from which the target has been detected (ESA convention)
MetaDataFileName	Target .json metadata file name
ShipFlag	Binary flag. If 1 the target is a ship, if 0 the target is not a ship (obtained after visual inspection of the data set). If the flag is 0 an AIS contact has been associated to a non-ship target and so the remaining fields are not significant.

The target metadata in the .json file (one for each target) partially replicate the fields in the .csv file. The .json file is organized in four group of variables:

- sensor
- target
- bands
- projection

The metadata fields for each group are described in the tables below (from Table 13 to Table 16).


 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

Table 13: Fields of the target metadata file (“sensor” group).

Field name	Description
SpaceCraftName	Satellite name
OrbitDirection	Orbit direction (ascending or descending)
OrbitNumber	Number of the acquisition orbit
ProductType	Product type (ESA conventions)
ProductURI	Name of the image data set from which the target has been detected (ESA conventions)
ProcessingLevel	Processing level of the input image product from which the target has been detected (ESA conventions)
MeanSunAzimuth	Mean sun azimuth [deg]
MeanSunZenith	Mean sun zenith [deg]
MeanAzimuth	Sensor mean azimuth at the target position for each sensor channel [deg]
MeanZenith	Sensor mean zenith at the target position for each sensor channel [deg]
PhiSatellite	Satellite direction [deg]

Table 14: Fields of the target metadata file (“target” group).

Field name	Description
MMSI_AIS	MMSI code from associated AIS contact
ShipType_AIS	Ship type from associated AIS contact text fields
Lat_Interpolated_AIS	Latitude of the associated AIS contact at the closest time of the sensor acquisition time [deg]
Lon_Interpolated_AIS	Longitude of the associated AIS contact at the closest time of the sensor acquisition time [deg]
X_Interpolated_AIS	Projected X position of the associated AIS contact interpolated at the sensor time [m]
Y_Interpolated_AIS	Projected Y position of the associated AIS contact interpolated at the sensor time [m]
TimeInterpolated_AIS	AIS time of the associated AIS contact [YYYYMMDDThh:mm:ss.sssZ]
SOG_AIS	Ship speed over ground from associated AIS contact [m/s]
COG_AIS	Ship course over ground from associated AIS contact [deg]
Length_AIS	Ship length from associated AIS contact [m]
Width_AIS	Ship width from associated AIS contact [m]
Draught_AIS	Ship draught from associated AIS contact [m]
NavStatus_AIS	Ship navigational status from associated AIS contact (it follows AIS conventions)
Time_SENSOR	Sensor acquisition time [YYYYMMDDThh:mm:ss.sssZ]
Lat_SENSOR	Latitude of the target centre of mass estimated from sensor data [deg]
Lon_SENSOR	Longitude of the target centre of mass estimated from sensor data [deg]
X_SENSOR	Projected X position of the target centre of mass estimated from sensor data [m]
Y_SENSOR	Projected Y position of the target centre of mass estimated from sensor data [m]
TileObjectStruct_TileID	Reserved
TileObjectStruct_TargetID	Reserved
TileObjectStruct_FileName	Reserved


 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public

Table 15: Fields of the target metadata file (“bands” group, one set of fields for each sensor channel).

Field name	Description
BandName	Sensor channel name
Resolution	Image resolution [m]
QuantificationValue	Calibration factor
TargetImageFileName	File name of the target image for the channel

Table 16: Fields of the target metadata file (“projection” group).

Field name	Description
MapProjection	Image map projection (“utm”)
Zone	UTM zone
Geoid	Name of the projection geoid (“World Geodetic System 1984”)

5.4.1.4 Thermal Camera and Vehicle Status data

Details will be included in Deliverable D8.6 (Data Management Plan V3) at Month 36 of the project.

5.4.1.5 Composite Maritime Event Data

This dataset comes along with database schema creation and data insertion statements for PostgreSQL. Therefore, it can be handled within database management systems according to our discussion in Section 5.3.1. Moreover, it is accompanied by documentation on the meaning of the complex events extracted by RTEC⁴¹. In Section 5.4.3 we comment on our provisions regarding interoperability related to complex event data, together with the rest of the maritime datasets.

According to the documentation of [16], in the recognised_CEs.csv file, each row includes FluentName|MMSI|Argument|Value|T_start|T_end “|” separated fields where [T_start, T_end) is the interval of a fluent (recognised activity in RTEC terminology) and fluent_name (MMSI, Argument)=Value, (Argument is optional). In case a fluent does not have an argument, the corresponding field has a space ‘ ’. Recognised activities (fluents) involve:

- withinArea (Vessel, AreaType)=true: A vessel is inside an area of type AreaType. AreaType={fishing, natura, nearCoast, nearPorts}.
- gap (Vessel)=Status: A vessel has a communication gap near ports or far from ports. Status={nearPorts, farFromPorts}.
- stopped (Vessel)=Status: A vessel is stopped near port or far from ports. Status={nearPorts, farFromPorts}.
- lowSpeed (Vessel)=true: A vessel moves with low speed.
- changingSpeed (Vessel)=true: A vessel changes speed.
- movingSpeed (Vessel)=SpeedStatus: A vessel is moving with speed relative to its type. SpeedStatus={below, normal, above}.
- underway (Vessel)=true: A vessel is under way.
- highSpeedNC (Vessel)=true: A vessel has speed greater than 5 knots within 300 meters from coast.
- anchoredOrMoored (Vessel)=true: A vessel is anchored or moored.
- loitering (Vessel)=true: A vessel is loitering.
- pilotBoarding (Vessel1, Vessel2)=true: A pilot boarding operation takes place between Vessel1 and Vessel2.
- sarMovement (Vessel)=true: A sar Vessel moves with characteristics of a sar operation.
- sarSpeed (Vessel)=true: A sar Vessel has speed coinciding with speed required for a sar operation.
- sar (Vessel)=true: A sar vessel is engaged in a sar operation.

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



- `rendezVous (Vessel1, Vessel2)=true`: Vessels Vessel1, Vessel2 are engaged in a rendez-vous (i.e., transshipment).
- `trawlingMovement (Vessel)=true`: A fishing Vessel moves with characteristics of a vessel engaged trawling.
- `trawlSpeed (Vessel)=true`: A fishing vessel has speed coinciding with speed required for trawling.
- `trawling (Vessel)=true`: A fishing vessel is trawling.
- `tuggingSpeed (Vessel)=true`: A vessel has speed coinciding with speed of a vessel involved in a tugging operation.
- `tugging (Vessel1, Vessel2)=true`: Vessels Vessel1, Vessel2 are involved in a tugging operation.

5.4.2 Making data openly accessible

AIS Raw Data portion available at: <http://doi.org/10.5281/zenodo.3754481>

Acoustic Data portion available at: to be uploaded by Month 19.

Composite Maritime Event Data available at: <https://doi.org/10.5281/zenodo.2557290>

AIS Raw, Kafka Streams and Patterns of Life data are part of MarineTraffic’s commercial activities⁴⁸ before and beyond the scope of the project. Therefore, only selected historical datasets are made publicly available for non-commercial (scientific) use.

All datasets are available within the INFOR consortium as soon as they get collected and in real-time for streams of data.

With respect to software tools that are needed to access the shared data that are made available, AIS Raw data and Composite maritime event data require no special software to get processed as they are provided in simple `.csv` files. Open source spatial and spatiotemporal libraries^{49,50} exist so that one can convert raw data to geometries. They can further be loaded to open source database management systems such as PostgreSQL/PostGIS.


For the Composite Maritime Event Data, the `sql-db` folder contains a series of scripts to load the dataset in a PostgreSQL database. The database name is by default `maritime_ces` and has a table called `ces` where the complex events detected are kept. In order to load the recognised activities into the database, first the `<PATH-TO-DATA>` in file `load_ce_data.sql` has to get substituted with the actual path to the data file. The `load.sh` file is to be run to execute the database creation script.

For the acoustic data, we also provide the Matlab code of the reader to help accessing the data. To see the tracks in Google Earth, one has to open all three (`tracks.kmz`, `groucho.kml`, `harpo.kml`) files in Google Earth, then navigate to the trial area (e.g. see [14] for `collab13`), recognisable by the large number of little green arrows and yellow and orange text labels. When zooming in white lines will appear, those are tracks. Initially all tracks are shown, for the entire duration of the day. The time slider in the top left corner of the Google Earth window allows for the selection of a time window, restricting the visible tracks to those that fall within the time window.

⁴⁸ <https://www.marinetraffic.com/en/p/ais-historical-data>

⁴⁹ <https://www.osgeo.org/projects/geotools/>

⁵⁰ <https://sis.apache.org/>

 <p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3 Deliverable D8.4</h2>	Doc.nr.: WP8 D8.4
		Rev.: 1.0
		Date: 30/06/2020
		Class.: Public

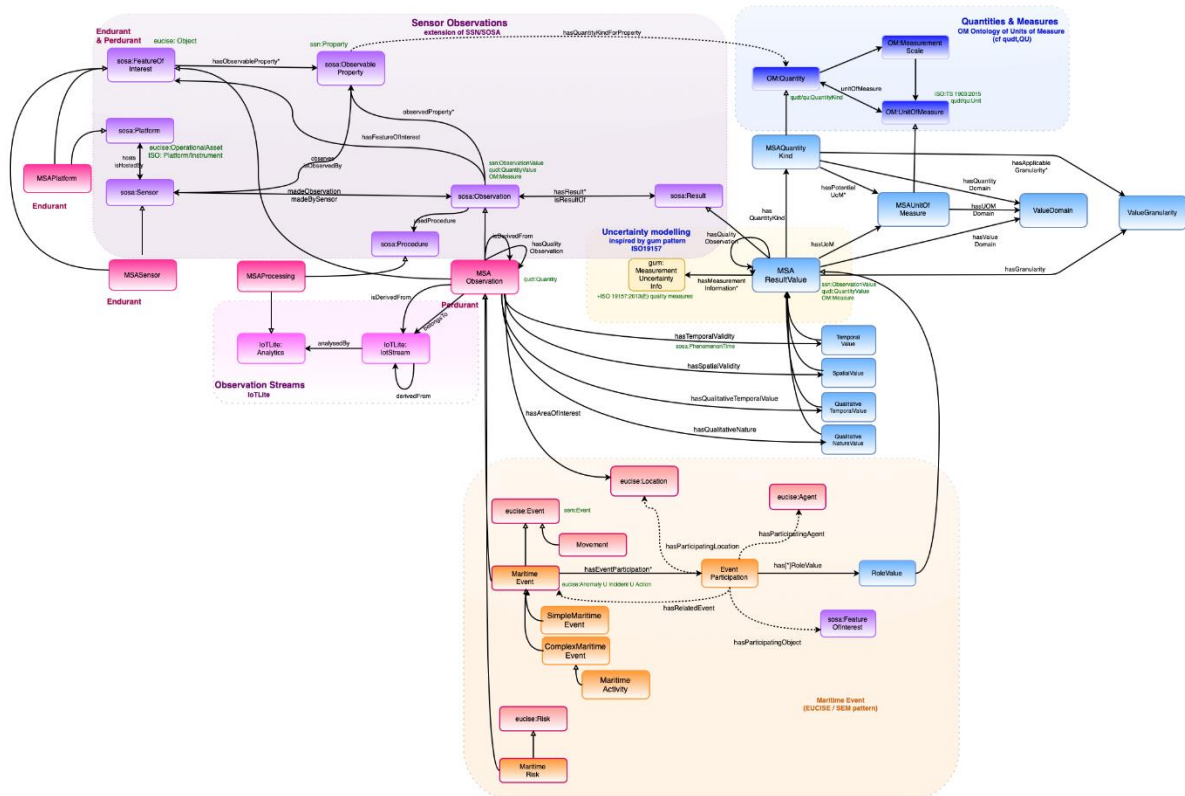


Figure 18: INFORE Maritime Ontology (IMO) - Overview

5.4.3 Making data interoperable

An INFORE Maritime Ontology (IMO) is being built in the scope of the project in order to encompass the variety of heterogeneous maritime data sources employed and formalize descriptors of generated data. IMO aims at modelling and annotating INFORE maritime data and processing results including sensor measurements (e.g., AIS positions, speed, Estimated Time of Arrival, destination), simple maritime events (such as those in Figure 10), complex maritime events and activities (such as those in Figure 10), vessel type classification from satellite data, acoustic data processing results.

IMO is designed as an extendible and customizable model for maritime events (facts, indicators, anomalies). In addition, IMO will enable uncertainty annotations for sensors, their measurements and outcomes of processing algorithms and machine learning models.

The IMO Data model is aligned with reference models for sensors, measurements, uncertainty, and (maritime) events. In particular:

- Sensor patterns defined in SSN/SOSA (Semantic Sensor Network/Sensors, Observations, Samples, and Actuators).
- Quantities and units as for measurement ontologies: qudt (Quantity, Units, Dimensions and Data types), QU (quantity, Unit) and OM (Ontology of Units of Measure and Related Concepts).
- Uncertainty of measurements and models as for GUM (Guide to the expression of Uncertainty modelling in Measurement) and ISO 19157:2003 (Geographic information - data quality).
- Maritime Situational Facts and Maritime Activities are modelled extending the SEM (Simple Event Model) and the EUCISE model/ontology (EUCISE-OWL).

<p>Project supported by the European Commission Contract no. 825070</p>	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



5.4.4 Data re-use (through clarifying licenses)

The maritime datasets are provided under Creative Commons Attribution Non Commercial No Derivatives 4.0 International license. Anyone may request access to the files of the shared data, provided that she fulfils the conditions of the license. The decision whether to grant/deny access is solely under the responsibility of the data provider.

The datasets that are shared at the INFORE community at Zenodo will remain available for reuse under the specified license scheme, even after the end of the project.

For AIS Raw, Kafka Streams and Patterns of Life data MarineTraffic follows specific procedures to ensure data quality. These procedures are detailed here⁵¹. Such procedures are yet to be formalized for the rest of the maritime datasets upon the maritime pilot details are determined and pilot execution is performed.

5.4.5 Data security

Maritime datasets are preserved, and curated following database/stream management system best practices as prescribed by the corresponding vendors and open-source communities.


There is no sensitive data used in the project (see also Section 8).

5.4.6 Allocation of resources

For Maritime data and respective streams, costs for servers used throughout the duration of the project for scientific research are covered as “other costs” in the overall project budget. These include the costs of data curation and preservation prior or after data sharing. The CMRE partner has further claimed costs for IT equipment required for collecting and preserving data from the planned pilot within the scope of WP3.

The Project Coordinator is responsible for data management issues throughout the project’s lifespan.

⁵¹ <https://www.marinetraffic.com/blog/four-ways-marinetraffic-ensures-ais-data-accuracy/>

 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.: WP8 D8.4
		Rev.: 1.0
		Date: 30/06/2020
		Class.: Public



6 Scientific Publication Data and Zenodo Repository Status

We now present the status of the INFORE repository¹⁵ at Zenodo. By the time of submission of this deliverable, the status of our Zenodo repository is as shown in Figure 19. Besides the datasets described in the previous sections, our Zenodo repository also includes publication file and data.

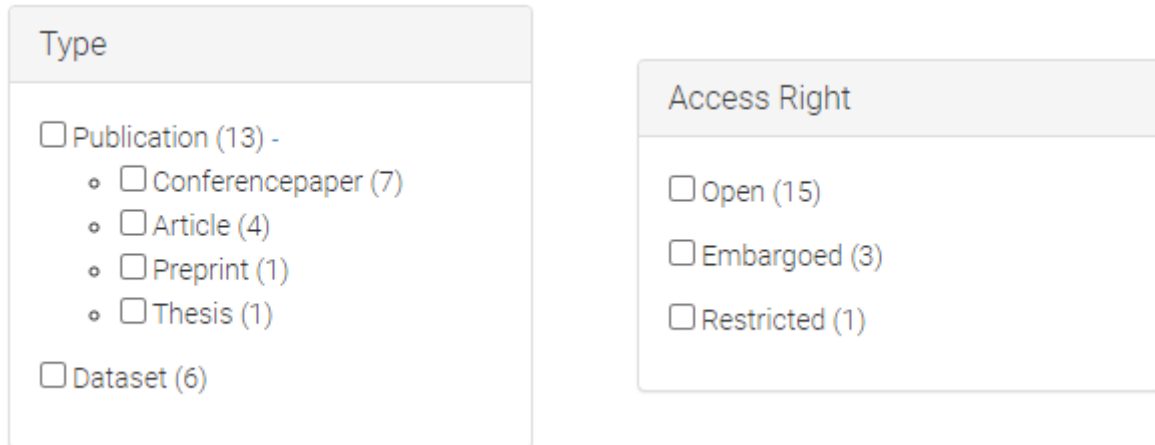



Figure 19: Status of INFORE's Zenodo Repository

As Figure 19 illustrates, we have been sharing 13 publications acknowledging INFORE and 6 datasets (on the left-hand side of Figure 19). Out of these items the vast majority of 15 items are provided with open access rights (on the right of Figure 19). There is 1 item with restricted access rights, which corresponds to the AIS Raw dataset. The restriction however only applies to ensure the non-commercial, no-derivatives license rights prior to allowing a download as mentioned in Section 5.4.2. There are also 3 embargoed publication items with embargo periods expiring on July 2020, June 2021 and October 2021, i.e., all publication items will be available open access within the project's duration.

Focusing on publication data, there have been uploaded 7 scientific publications involving conference papers, 4 journal articles, 1 preprint (arXiv) items and 1 master thesis.


Concentrating on dataset uploads, the total size of uncompressed data that are shared by the project reaches 300GB, while we estimate 350GB of uploaded data will be reached by Month 19 when Acoustic Maritime Data are to be uploaded.

 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



7 Expert User Requirements/Feedback Data

In order to achieve its goals and evaluate the success of the developed technologies, INFORE engages expert users in key phases of its workplan including both the requirement analysis and scenario definition of the INFORE use cases and the evaluation of the demos and prototypes that are being developed. Use case and technical partners, first identify candidate expert users to be interviewed. Then, expert user engagement comes after receiving all the necessary information about the project, its purposes and the aim of the interview/questionnaire and having provided their explicit consent on a voluntary basis. Requirements and feedback, respectively, are collected using questionnaires based on which users are interviewed. The format of the questionnaire, interview, feedback changes depending on the phase of the project. Respective datasets are built incorporating expert user responses as described in Section 7. These datasets are anonymized, and only anonymized and aggregated data are included in project reports. Examples of such anonymized data are currently included in Deliverables D1.1, D2.1 and D1.3 submitted on Month 3 of the project and respective gathered feedback data are discussed in Deliverables D4.1 submitted on Month 12 of the project and Deliverables D1.3, D2.2, D3.2 submitted on Month 18 of the project. Similar data are to be included in future deliverables of WP1, WP2, WP3 and WP4 during the final evaluation of project results by expert users of relevant application fields. The aggregated results are used so that INFORE captures both functional and non-functional requirements of application fields fostering its approach, as well as in applying corrective actions regarding the technological components it develops. The ethical issues that arise and the way ethics are managed during this process have been detailed in Deliverable D8.1 submitted on Month 3 of the project.

 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.: WP8 D8.4
		Rev.: 1.0
		Date: 30/06/2020
		Class.: Public



8 Ethical aspects


As already stated, the Expert User Requirements/Feedback data (Section 7) undergo anonymization procedures and only aggregated results of expert users responses on questionnaires and interviews are included in INFORE's reports. Expert users participate in the collection of such data on a voluntary basis after having obtained a complete description about the project, its objectives and vision. All participants need to first provide their explicit consent by signing respective consent forms. All such procedures and compliance with existing regulatory frameworks were detailed in Deliverable D8.1 Ethics Management Plan, prepared and submitted on Month 3 of the project.

Hard or electronic copies of gathered data are collected in person by a responsible INFORE researcher (also mentioned on the signed consent form) and are safely kept in shielded envelopes or password protected files. It is the responsibility of the INFORE partner, affiliated with the corresponding researcher, to ensure abidance by the relevant regulatory frameworks until data comes at the possession of the data controller partner, which is the Project Coordinator (Athena). The Coordinator receives these copies in person in the first project plenary meeting after data collection has been completed. In case this is not possible, it is the responsibility of the respective INFORE researcher to create password protected files of electronic copies of all data that remain at their possession and communicate them via secure partner-specific institutional repositories until these data come at the possession of the data controller. As soon as the data come at the possession of the data controller, the INFORE researcher erases all collected data.

The data then moves to the premises of the data controller, where only electronic copies are created and are kept at a server without internet connection. The data controller has secure access to these data granted by Athena. Each set of data items (consent form, questionnaires, recorded interview or responses to surveys) for a single participant is assigned a code to identify their data after anonymization procedures. This is necessary, for instance, so that an expert user's data can be withdrawn and erased upon their request. This code is made known to the corresponding questionnaire participant at the time of data collection. The key-file containing identity information is kept separately from the de-identified, pseudonymized parts of the data on the same server. The data containing information about the participants' identity will be stored in a secure file to which only the data controller will have access. The encryption will use the AES 256 algorithm. Any data acquisition or communication will be performed using asymmetric algorithms using keys of at least 2048 bits.

Gathered data are de-identified i.e., all possible personal details of the participant are removed, and anonymized, i.e., even data that can implicitly reveal a person's identity are eliminated to extract aggregated results used in the scope of the project. This process is performed only by the project Coordinator. The Coordinator has appointed a Data Protection Officer (DPO – see deliverable D8.1) who approves (or not) the de-identified, anonymized, aggregated data inclusion in project reports, prior to publication, according to the relevant regulatory frameworks.

Responses to questionnaires/interview/surveys conducted during INFORE will be stored by the data controller in its premises until 31/12/2021; then they will be deleted.


 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public



9 Conclusions

INFORE architectural components receive input in the form of voluminous, high velocity streams and historical data from three different application fields including Life Sciences, Financial and Maritime domains. The output data consists of results of calibrated models describing cancer evolution under combinational drug therapies, market data forecasts and enhanced situational awareness in the maritime domain with the identification and tagging of activities even in the case of “dark targets”.


This deliverable describes the datasets and steps followed for relevant data management purposes, following the FAIR principles, within the scope of the project. The current document constitutes the second version of the Data Management Plan and provides an analysis of all the data sets and sources utilized in INFORE that are being used during the project as identified by the project consortium partners and the way the project results are being shared. The DMP will be updated by Month 36 of the project in Deliverable D8.6 with additions outlined in the current DMP version, as well as any other significant changes should they arise.

 Project supported by the European Commission Contract no. 825070	WP8 T8.3 Deliverable D8.4	Doc.nr.: WP8 D8.4
		Rev.: 1.0
		Date: 30/06/2020
		Class.: Public



10 References

- [1] A. Ghaffarizadeh, R. Heiland, S.H. Friedman, S.M. Mumenthaler, and P. Macklin, PhysiCell: an open source physics-based cell simulator for 3-D multicellular systems, *PLoS Comput. Biol.* 14(2): e1005991, 2018. DOI: 10.1371/journal.pcbi.1005991.
- [2] <http://physicell.org/> and <https://github.com/MathCancer/PhysiCell>
- [3] Stoll G, Caron B, Viara E, Dugourd A, Zinovyev A, Naldi A, Kroemer G, Barillot E, Calzone L. MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics* btx123. 2017 Mar. DOI: <https://doi.org/10.1093/bioinformatics/btx123>
- [4] <https://maboss.curie.fr/> and <https://github.com/sysbio-curie/MaBoSS-env-2.0>
- [5] Gaëlle Letort, Arnau Montagud, Gautier Stoll, Randy W. Heiland, Emmanuel Barillot, Paul Macklin, Andrei Yu. Zinovyev, Laurence Calzone: PhysiBoSS: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling. *Bioinform.* 35(7): 1188-1196 (2019)
- [6] <https://github.com/gletort/PhysiBoSS> and <https://github.com/sysbio-curie/PhysiBoSS>
- [7] Nikos Giatrakos, Nikos Katzouris, Antonios Deligiannakis, Alexander Artikis, Minos N. Garofalakis, George Paliouras, Holger Arndt, Raffaele Grasso, Ralf Klinkenberg, Miguel Ponce de Leon, Gian Gaetano Tartaglia, Alfonso Valencia, Dimitrios Zissis: Interactive Extreme: Scale Analytics Towards Battling Cancer. *IEEE Technol. Soc. Mag.* 38(2): 54-61 (2019)
- [8] Effrosyni Anesti, "Forecasting promising biological simulations at PhysiBoSS", Diploma Work, School of Electrical and Computer Engineering, Technical University of Crete, Chania, Greece, 2020. <https://doi.org/10.26233/heallink.tuc.84787>
- [9] Antonis Kontaxakis, Nikos Giatrakos, & Antonios Deligiannakis. (2020, March 21). A Synopses Data Engine for Interactive Extreme-Scale Analytics. Zenodo. <http://doi.org/10.5281/zenodo.3849978>
- [10] Rafael Valencia-García, Francisco García-Sánchez, Dagoberto Castellanos Nieves, Jesualdo Tomás Fernández-Breis: OWLPath: An OWL Ontology-Guided Query Editor. *IEEE Trans. Syst. Man Cybern. Part A* 41(1): 121-136 (2011)
- [11] Dimitris Zissis, Konstantinos Chatzikokolakis, Giannis Spiliopoulos, Marios Vodas: A Distributed Spatial Method for Modeling Maritime Routes. *IEEE Access* 8: 47556-47568 (2020)
- [12] M.1371: Technical Characteristics for an Automatic Identification System Using Time-Division Multiple Access in the VHF Maritime Mobile Band. Accessed: May 22, 2020. [Online]. Available: <https://www.itu.int/rec/R-REC-M.1371/en>
- [13] Pitsikalis M., Artikis A., Dreo R., Ray C., Camossi E., and Joussetme A. Composite Event Recognition for Maritime Monitoring. *International Conference on Distributed and Event-Based Systems (DEBS)*, 2019
- [14] Goldhahn, R. & Braca, Paolo & Ferri, Gabriele & Munafò, Andrea & Lepage, Kevin. (2014). Adaptive Bayesian behaviors for AUV surveillance networks. *2nd International Conference and Exhibition on Underwater Acoustic (UAC)*
- [15] Munafò, Andrea & Canepa, Gaetano & Lepage, Kevin. (2018). Continuous Active Sonars for Littoral Undersea Surveillance. *IEEE Journal of Oceanic Engineering*. PP. 1-15. 10.1109/JOE.2018.2850578.
- [16] Manolis Pitsikalis, & Alexander Artikis. (2019). Composite Maritime Events (Version 0.1) [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.2557290>

 Project supported by the European Commission Contract no. 825070	<h2>WP8 T8.3</h2> <h3>Deliverable D8.4</h3>	Doc.nr.:	WP8 D8.4
		Rev.:	1.0
		Date:	30/06/2020
		Class.:	Public