Data Management Plan V1
Work Package 8 Task 8.3 Deliverable 8.3

Authors

Nikos Giatrakos, Antonios Deligiannakis
Athena Research & Innovation Center

Distribution list:

Groups:	Others:
WP Leader: Athena Task Leader: Athena	Internal Reviewer Partner: Barcelona Supercomputing Center (BSC) INFORE Management Board INFORE Project Officer

Document history:

Revision	Date	Section	Page	Modification
0.6	30/05/2019	All	All	Creation
0.7	31/05/2019	All	All	Self-review
0.8	03/06/2019	-	-	Submitted for internal review
0.9	17/06/2019	All	All	Internal review comments incorporated
1.0	28/06/2019	All	All	Self review, Final version

Approvals:

First Author: Nikos Giatrakos (Athena) Date: 03/06/2019

Internal Reviewer: Arnau Montagud (BSC) Date: 12/06/2019

Coordinator: Antonios Deligiannakis (Athena) on behalf of the Management Board Date: 28/06/2019

Table of contents:

1 Executive Summary

This deliverable provides an initial version of the Data Management Plan in INFORE at Month 6 of the project. The deliverable will be amended on Month 18 and receive its final form in Month 36. To build the initial Data Management Plan we utilized the respective template of the European Research Council (ERC): ERC DMP[1] provided by the Digital Curation Center. We also followed the guidelines for FAIR (findable, accessible, interoperable, reusable) data management in H2020[2].

Following this template, we provide a summary of the datasets that are input to or output of the project and describe our initial provisions for:

- making data findable
- making data openly accessible
- making data interoperable
- increasing data reusability

We further provide responses to questions related to allocation of resources, data security and ethical issues where and when they arise in the scope of INFORE.

[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

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 machine learning and forecasting algorithms optimized over multiple Big Data platforms and HPC infrastructures which, in turn, produce output datasets per domain.

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 a variety of market data, including stock market and crypto-currencies market data, arriving in correlated, high-velocity streams. The aim is to forecast price swings of stocks, currencies, commodities, 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 AIS (Automatic Identification System) data, as well as quantities sensed by autonomous unmanned vehicles navigating at sea. The aim of incorporating both AIS and sensed data from autonomous vehicles is to correlate related data from these data sources towards the identification and forecasting of activities of "dark targets" that (intentionally) hide from AIS monitoring systems. There is an additional source of data for each use case that is related to requirement collection and expert user feedback.

We first summarize these datasets that are available in the project in Section 3, while in Section 4 we elaborate on FAIR (findable, accessible, interoperable, reusable) provisions for data shared in the scope INFORE. Section 5 includes conclusive remarks and mentions future steps with respect to data management.

3 Data Summary

This section outlines the data that will be used or generated in the scope of the INFORE project. There are three use case scenarios where the INFORE approach will be applied and where relevant data are collected and/or generated. Moreover, there is an additional source of data related to requirement collection and expert user feedback on INFORE's demos and prototypes.

3.1 Life Science Use Case Datasets

In the Life Science Use Case INFORE performs simulations on cancer evolution studying the effect of the application of certain drug combinations on destroying (leading to necrosis or apoptosis) cells of various types of cancer. In particular, the goal is to develop a virtual laboratory to conduct simulations that: i) integrate heterogeneous sources of experimental data as well as biological knowledge; ii) generate hypotheses (by analysing simulation outputs) about underlying mechanisms of biological processes determining tumour growth, drug resistance and drug synergies on cells; and iii) in-silico design, model calibration and test of treatments based on combination of drugs. The above spans all tasks within the scope of WP1 of the project. For these purposes, we integrate two simulation frameworks, an agent-based modelling framework and a Boolean modelling one, into a software called PhysiBoSS[3], which will act as the main data generator for this use case and may be extended where needed.

So, roughly speaking in order to give the general picture behind this use case, what is being simulated is populations of cells on which various drugs are applied. The input to the population and, thus, to each cell is a combination of drugs which leads a number of internal cell states, represented by a signalling network, to become active or inactive. The concatenation of state activation/deactivation in this network affects the cell cycle (mitosis, DNA replication, etc) leading to cell survival or determines cell's commitment to different cells deaths such as necrosis, apoptosis, etc.

Simulation input and output files are included in a single simulation folder. Currently there is no specific naming convention for this folder, but for the simulation results that will be generated in the scope of the project, the main folder naming will be of the form (subject to revision if needed, <> denotes lists of properties, parameters):

<Cellpopulationproperties>_<Cellpopulationmajorparameters>_<drugsapplied>_<drugmajorparameters>

For instance, spheroid_null_TNF_pulse150 denotes that a spheroid of cells (with PhysiBoSS default parameters) is simulated and TNF, i.e., Tumour Necrosis Factor (TNF), is applied every 150 milliseconds. In this case TNF is produced by neighbouring cells, but the same naming convention is used whenever a pharmaceutical drug that suppresses the physiologic response of a given protein is simulated.

The main input file is in XML format, as proposed in the MultiCellDS standardization initiative[4], and describes the various simulation parameters related to:

- <simulation> parameters that refer to global properties of the simulation, e.g. the numerical time step, are included in the simulation parameters xml element.
- <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.
- <network> properties related to the Boolean network computation describing the cell's signalling states. The majority of the parameters concerning the Boolean network structure are defined separately in MaBoSS[5] network files, with the MaBoSS conventions in CFG and BND ASCII files stored in a BN sub-folder inside the main 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

[3] https://github.com/sysbio-curie/PhysiBoSS/
[4] http://multicellds.org/
[5] https://maboss.curie.fr/

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.

- An <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). With PhysiBoSS code, an executable PhysiBoSS_CreateInitTxtFile can be given, which allows the user to create a .txt file containing this information for given parameters and chosen modes. This .txt file is referenced in the <initial_configuration> element. The output of a simulation as described below can also be used as the input initial file to another one.

An 'output' and 'microutput' sub-directories are created in the main directory of the simulation. A report.txt file is created inside the main 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. In the sub-folder 'output', .txt files named 'cells_' followed by the time value 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). Vocabularies for such variables relevant to INFORE will be created in the project's lifespan, especially for data sharing and re-usability purposes.

Example of output file for cells at time 0:
Time;ID;x;y;z;radius;volume_total;radius_nuclear;contact_ECM;freezer;polarized_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
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 .txt files are built for the different densities (free-roaming molecules on the extracellular space, such as oxygen, signalling molecules, microenvironment density, etc) in the micro-output sub-folder. These are named after the density they represent and the time they correspond to. For instance, oxygen_t00000.txt.

An exemplary output file for micro-environment density at time 0 is given below. Currently there is no header row which will be added in the files generated for the purposes of INFORE. The first three columns correspond to spatial coordinates and the fourth to the value of the density:
-417.5;-492.5;-492.5;0.0630239
-357.5;-492.5;-492.5;0.0630185

The size of the produced data depends on the size of the simulated tumour and the simulation parameters such as the simulation timestep or granularity (frequency) at which output or micro-output files are generated. Indicatively, we note that simulating tumour of realistic sizes can produce data at a rate of approximately 100 GB/min.

Currently simulations can only be analysed in an offline manner. That is, users have to wait for the simulation to complete before they analyse the summary report, output and micro-output. For the needs of the project, apart from storing useful files as those analysed above, a running simulation will provide equivalently structured data streams that will be analysed online.

Within the scope of INFORE, a large number of simulations will be conducted over High Performance Computing (HPC) infrastructure. This process will be optimized by the project optimization modules and will be assisted by machine learning and forecasting techniques in order to train algorithms that will automatically decide when a running simulation cannot provide a useful outcome with respect to harnessing tumour evolution. In that, computational resources will be saved and devoted to new simulation instances with different parameters, cutting down the time needed to reach conclusions on effective drug combinations. Furthermore, utilized simulation models, 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 using the drug experiments corresponding to that particular line. The starting point will be to incorporate experimental evidence on drug synergies, collected from public available resources such as the Drug

Combination DataBase (DCDB)[6]. Given the output of this process, the simulation input and output results will be enhanced with calibration parameters and will get tagged based on their "usefulness". The above 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.

3.2 Financial Use Case Datasets

The goal of the Financial Use Case (WP2 in INFORE) is to analyse historical and real-time stock market data to train machine learning models that extract valid rules 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.

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 that might be also incorporated in INFORE are international interest rates.

Each asset's data comes in a separate stream and is obtained via an API[7] that has been developed by the SpringTechno consortium participant. The streaming data is available in the following format: "Date, Time, Price, Volume" for each data tick of an asset. "Volume" is the number of shares or contracts traded. For instance:
01/08/2019,00:00:01,1288.68,1
01/08/2019,00:00:01,1288.76,1
01/08/2019,00:00:01,1288.71,1
01/08/2019,00:00:01,1288.7,1
01/08/2019,00:00:02,1288.72,1

The historical data of INFORE have a similar format and characteristics as the real-time data and cover most market players. That is, we have a separate file for each asset and files of historical data are stored in comma separated .txt files where the naming convention that is used is:

Exchange·Symbol[.Index]·ExpiryDate.txt

Exchange is the exchange on which the commodity is traded, Symbol is the root symbol for the commodity and 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 (embraced in []). 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.

This historical data is also provided in time-based-compressed form, where ticks are condensed to specific time frames. These files are of .txt 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.

These files are named after their time frame, exchange, symbol and expiry date:

TimeFrame_Exchange·Symbol[.Index]·ExpiryDate.txt

[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4275564/
[7] http://www.springtechno.com/J-Data_API_Description.pdf

For instance, 201901_Stocks USA·XRAY.NASDAQ·NoExpiry says that the time frame of the condensed view refers to January 2019, the Exchange involves USA Stocks, the Symbol refers to the stock of a particular company traded in NASDAQ, while there is no expiry date for the contract.

INFORE Financial Use Case 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.

3.3 Maritime Use Case Datasets

For the purposes of the Maritime Use Case in INFORE's WP3, a number of historical and real-time datasets will be used/produced and fused in order to provide enhanced Maritime Situational Awareness (MSA), i.e. the ability to perceive and reason with activities, events and threats at sea. This use case incorporates various categories of datasets including AIS data, thermal camera data, acoustic sensor data, Copernicus (Sentinel-1, Sentinel-2) data and camera data from sensorized RHIB/UAV. 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". This is due to the fact that, by definition, dark target hide their identity and position and thus respective information is inevitably missing from the MSA analysis algorithms. Therefore, we aim at fusing 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 camera sensors thereby collecting nearby vessel data related to vessel type, size, etc. The idea is to combine the information of these sources to related AIS data to increase the value of information each separate data source can provide as well as the validity of the extracted MSA-related outcomes.

Below we outline the available datasets along with their format, collection purpose and volume.

AIS Data: These data are formatted using the standard AIS format. They are available both in historical form with approximately 1 TB 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 itself[8]. AIS data will be used to generate vessel trajectory density maps, infer preferred routes at sea and for vessel monitoring and anti-collision purposes.

Kafka Streams: Derived data streams out of AIS streaming data are processed in Kafka 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 (ETA). These streams are updated at a sub-second rate.

Vessel Statistics: Historical, structured (e.g., .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.

Patterns of Life: "Patterns of Life" 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.

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

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

level, next waypoint) required to supervise and control its operation. Approximately, 100 MB of data are collected for every hour of operation.

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 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.

Visible/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.

Copernicus Data[9] and Labelled Ship Target Training Datasets: will also be used 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.

For all of the above, we still need to define common naming convention for portions of the various datasets that are being fused for the same MSA purpose.

3.4 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 4.6. 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 prepared and submitted on Month 3 of the project. Similar data are to be included in future deliverables of WP1, WP2, WP3 and WP4. 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.

[9] https://scihub.copernicus.eu/

4 FAIR Data

4.1 Data findability, including provisions for metadata

Standard EUDAT B2SHARE[10] services will be examined and where appropriate used in making sure that appropriate metadata[11] are created and portions of all Life Science, Financial, Maritime historical data that will be shared is discoverable through B2FIND. Keywords are part of the metadata and will be extracted following appropriate community standards where applicable, while automatic keyword extraction tools will be used for data linked to research publications.

Relevant portions of the data will be shared through the INFORE page at zenodo, which is implementing the FAIR principles. Zenodo uses standard dataset identification mechanisms including Digital Object Identifiers (DOIs). In case new versions of shared datasets exists, they will receive their own DOI and their zenodo description will include the link to the older version(s).

Current naming conventions were described in Section 3, but they are subject to change if need be for data sharing purposes. In any case, DOIs, metadata and keywords override naming conventions with respect to making data findable.

4.2 Making data openly accessible

Portions of historical datasets from all three use case will be made available by INFORE partners for public use, based on the INFORE consortium assessment on their usefulness in training relevant machine learning models, which extract rules that can later on be used to forecast events of interest such as price swings in the Financial Use Case, anomalies at sea in the Maritime Use Case or promising simulations in the Life Science Use Case. For data gathered from surveillance activities in the maritime domain, such as thermal cameras or acoustic sensor data provided by NATO CMRE, restrictions may apply on the time of production of the historical data.

Relevant portions of the data will be shared through the INFORE page at zenodo[12] and, as the dataset summary in Section 3 demonstrates, the Life Science and the Financial data are accessible without requiring advanced software or tools. This is still the case after dataset collection/generation in INFORE. For instance, in order to produce data related to the Life Science Use Case one needs to be familiar with specific frameworks, like the open source code of PhysiBoSS, but the data that will be shared by the project will be the result of simulations using such frameworks. Thus, access to advanced software tools for accessing the Life Science and the Financial data datasets will not be necessary.

On the other hand, different software tools are used for manipulating the datasets used in the Maritime Use Case. The particular code/software/platform used for each of the datasets of WP3 is mentioned below:

AIS Data: SQL server, PostGIS, Python, Spark Human Operator/Combat Management System

Kafka Streams: Akka[13] + Scala, Python

Vessel Statistics: SQL Server and Python

Patterns of Life: SQL Server, Akka + Scala

Vehicle Status Data: Matlab, R, Python, C++

Acoustic Data: Matlab, R, Python, C++

[10] https://b2share.eudat.eu/
[11] http://rd-alliance.github.io/metadata-directory/
[12] https://zenodo.org/communities/infore-project/?page=1&size=20
[13] https://akka.io

Visible/Thermal Camera Data: Matlab, R, Python, C++

Copernicus Data[14] and Labelled Ship Target Training Datasets: Matlab, R, Python, C++

Whenever the code and tools are not proprietary, the consortium will consider providing R or Python/C++ code or data reader APIs along with the data being shared.

There will be no further restriction on the shared data access and therefore there is no need for the establishment of access control mechanisms or data access committee in the scope of the project.

4.3 Making Data Interoperable

All INFORE data and metadata are designed for interoperability and follow community standards. For data of the Life Science Use Case there exist standards developed by major international initiatives such as IHEC[15], ICGC[16], IHMC[17] and MME[18]. For data of the Financial Use Case, European and US Stock market data standards will be used as well as statistical and scientific standards of data harmonization[19]. For data of the Maritime Use Case AIS Data standards[20] will be used. Finally, for the acoustic data from autonomous vehicles, camera data from sensorized RHIB/UAV and thermal camera data of the pilot conducted in WP3 appropriate standards should be further investigated. In case such standards do not exist or are unavoidable to follow, we will gather efforts in creating appropriate mappings to more commonly used ones.

Vocabularies for all data types present in these datasets will be created to allow inter-disciplinary interoperability.

4.4 Increase Data Re-use

Wherever possible the data will be shared right after production (e.g. "useful" simulation data in the Life Science Use Case) following the Creative Commons 4.0 License. Data that are used/produced in the experimental evaluation of research papers within the scope of INFORE will be shared only after relevant publications have been accepted in the intended publication venues. The CC Creative Commons 4.0 License is suggested as it guarantees maximum re-use (and redistribution) while maintaining the traceability of the use and credit to the data providers and their sponsors.

The data that will be shared at zenodo will remain there along with the research publications (if any) they are linked to. They will also be updated subject to their use in follow up research by the corresponding partners or projects extending INFORE's outcomes.

[14] https://scihub.copernicus.eu/
[15] http://ihec-epigenomes.org/
[16] https://dcc.icgc.org/
[17] http://www.human-microbiome.org/
[18] http://www.matchmakerexchange.org
[19] The major stock and commodity exchanges are highly regulated. As examples, to name just a few standards, there are listing rules (e.g. https://www.londonstockexchange.com/home/guide-to-listing.pdf), defined market opening times and identification symbols for each market player (https://eresearch.fidelity.com/eresearch/markets_sectors/global/marketHours.jhtml) and certain market regulating rules like e.g. the "down-tick/ up-tick rule" (http://www.investopedia.com/ask/answers/165.asp). In parallel to the described strict regulation of the market, the standards for providing market data are set to high quality by the stock exchanges. Usually real time data comes as a price plus trading volume at a given time. As historical data, it is usually compounded as an "open-high-low-close- volume" information for a given time frame.
[20] such as those defined in the ITU Recommendation 1371-4, "Technical characteristics for an Automatic Identification System using time-division multiple access in the VHF maritime mobile band," ITU, Tech. Rep. Recommendation, 2001

Data quality control will take place on par with the studies on the operations of the algorithms developed or used in INFORE, using general-purpose, open source data quality software[21].

4.5 Allocation of resources and data security

For data stemming from the Financial Use Case, the corresponding partner, SpringTechno, has claimed costs for servers as "other direct costs" in the project budget. For data stemming from the Maritime and Life Science Use Cases, 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 Project Coordinator is responsible for data management issues throughout the project's lifespan.

In this project, the only data at risk considered are the Expert User Requirements/Feedback data (see Section 3.4) which are discussed separately in Section 4.6.

4.6 Ethics

As already stated, the Expert User Requirements/Feedback data (Section 3.4) 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 envelops 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.

[21] https://datacleaner.github.io/

5 Summary and Future Steps

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 initial Data Management Plan summarizes the aforementioned data and describes initial provisions for the way they will be handled, shared and rendered findable, easily accessible, interoperable and reusable for relevant scientific, research or other communities. The Data Management Plan will be enhanced and amended, where needed, in Deliverable D8.4 on Month 18, while it will receive its final form in Deliverable D8.6 on Month 36 of the project.