Platform for exploiting spatio-temporal maritime data

Cyril Carré, Victor Litoux, Anne-Clémence Duverger, Cyril Ray
Naval Academy Research Institute, Brest, France
cyril.carre@ecole-navale.fr, victor.litoux@ecole-navale.fr, anne.duverger@ecole-navale.fr, cyril.ray@ecole-navale.fr

***Abstract*—Maritime data are wide-ranging and originate from multiple sources, including vessel locations, administrative records, and geographic information. Extracting and managing this knowledge can be useful for monitoring fisheries, trade, illegal acts, etc. The paper presents a geospatial data processing architecture dedicated to multisource maritime data. Our aim is to provide a solution to process these versatile data in order to provide analysis and interpretation capabilities of maritime mobilities, behaviours and situations. This paper describes the processing chain, highlighting the technical challenges associated with big data issues, data analysis and visualization. The workflow of data, the design of databases and main data processing and algorithms are presented.**

***Index Terms*—Maritime data management and processing, Geographic information system, Vessel mobility**

I. Introduction

Monitoring, analyzing and understanding maritime mobility constitutes a major challenge for modern society, as it reflects the commercial, societal and geopolitical stakes among nations. In this context, mastery of maritime information is crucial for providing intelligence and ensuring rapid, well-informed operational decision-making [1]. This knowledge relies on exploiting multi-source data collected from four maritime dimensions: the sea surface, the water column, the ocean floor, and space. Handling and analyzing these disparate data sources is complex and requires mobilizing multiple disciplinary fields, including big data management, knowledge engineering, artificial intelligence, geographic information processing, and human sciences. This necessitates the development of robust and efficient analytical tools to monitor fishing activities, commercial exchanges, environmental safety, and detect and prevent illicit activities.

The processing of maritime big data through advanced computational architectures generally follows a step-based methodology that transforms raw Automatic Identification System (AIS) data into actionable maritime geospatial intelligence [2]. For instance, the European project BigDataOcean considered a data value chain composed by 5 steps [3]; data acquisition, data analysis, data curation, data storage and data usage where each step encompasses big data techniques including databases and data lakes. Generally speaking, at a first stage, data acquisition involves collecting high-volume AIS transmissions from multiple sources, utilizing centralized/distributed streaming platforms to capture real-time vessel movements and characteristics. The second phase focuses on data preprocessing and cleaning, where advanced parsing techniques and machine learning algorithms remove noise, standardize formats, and handle incomplete or erroneous entries, ensuring data quality and reliability. In a third phase, geospatial analysis is performed, leveraging distributed computing frameworks to conduct complex spatial computations, tracking vessel trajectories, identifying maritime patterns, and correlating geographic information with temporal data [4]. The fourth phase involves advanced analytics, mining algorithms [5] and predictive modelling based on machine learning and deep learning techniques to extract meaningful insights, such as predicting vessel behaviour, detecting anomalies, and assessing maritime traffic patterns [6]. The final phase encompasses visualization and decision support, transforming processed data into interactive dashboards and geospatial visualizations that enable maritime stakeholders to make informed decisions about navigation, safety, and resource management [7].

This paper aims to present the architecture we designed to support the management and the analysis of worldwide maritime data. The description of platform's functionalities and capabilities will be presented. The objective of the information system is to support geospatial intelligence case studies through analysis and interpretability of:

- Maritime mobilities, e.g.: What are the vessels? What are their characteristics?
- Maritime behaviors, e.g.: Where are vessels moving to? Where do they come from?
- Maritime situations, e.g.: What is the purpose of the vessel's movements? What is the geopolitical context of the area?

The article is structured as follows. The next section provides an overview of the different data sources currently used by the platform. Section III presents the platform's architecture dedicated to the processing of multi-source maritime data. In Section IV, we explore the processing pipeline used to extract and analyze the data and several algorithms are introduced. Section V illustrates several use cases based on this processing workflow. Finally, the last section concludes and points the way to future research.

II. Maritime Data

The integration of Automatic Identification System (AIS) data with Lloyd's Register maritime information offers a high potential for comprehensive maritime situational awareness, including risk assessment. While AIS provides real-time vessel tracking and positional data, Lloyd's Register offers detailed

static information about vessel characteristics, ownership, classification, and historical performance. By combining these two data sources in our methodological approach, we aimed to develop more advanced and robust analyses of maritime activities, enabling more precise investigations of maritime safety, vessel movement patterns, potential maritime risks, and environmental impact assessments. This combined approach allows for a multi-dimensional understanding of maritime ecosystems, transcending the limitations of individual data sources and providing a more holistic view of both maritime transportation networks and long-term vessel identity (including identity fraud). This section describes the main data sources currently integrated into our platform.

A. AIS data

The primary source of navigation data in the world is the Automatic Identification System (AIS) [8]. AIS is a short-range tracking system designed to prevent collisions at sea by broadcasting the position of vessels alongside associated data (e.g., ship identifier, name, speed, etc.). Equipped vessels regularly broadcast dynamic and nominative information. This enables other vessels, coast guards and emergency services to know the vessel's current position. Although AIS was designed to ensure ship safety, the data is particularly well suited to provide global information on the vessel fleet. Messages are transmitted in near-real time and collected by terrestrial or satellite receivers. AIS data offers numerous advantages, including near-global data coverage and near-real-time availability.

Nowadays, around 400,000 vessels worldwide utilize AIS, transmitting static (ship name, flag, dimensions) and dynamic (position, speed, heading) information every 2 seconds to 3 minutes. Consequently, this system generates daily data representative of typical big data issues. The two principal challenges addressed in this work are:

- Volume, which refers to the amount of data, too abundant to be acquired, stored, processed, analyzed and disseminated by standard tools. A day of AIS messages worldwide represents around 9 gigabytes of raw data in NMEA format.
- Veracity which is an intrinsic problem. AIS is an open system, partly filled in manually but also facing errors and malpractices it conveys are numerous [9]. It covers two kinds of errors. The first refers to data quality: simple anomalies due to the nature or acquisition of the data (outliers or missing values, duplicates, etc.). The second concerns their deliberate manipulation: falsification, hacking. For these reasons, depending on the analyses to be carried out, appropriate data handling processes need to be put in place.

For the implementation of our geospatial data processing architecture (cf. Section III) and our experiments, we used one year of worldwide AIS data (year 2019). The data was subsampled at 5-minute intervals. This results in 17.2 billion rows and represents 4 TB of data.

B. Administrative data and expert knowledge

Our second data source consists of administrative data on vessels. This encompasses data from ship registers (e.g. French National Radiofrequency Agency, European Fleet Register) or data from insurance companies. In our studies, we mainly used daily administrative information supplied by Lloyds insurance company. This dataset provides 37 additional fields to the AIS data: e.g. vessel owners, commercial operators, year of construction, tonnage, etc. These fields are jointed to AIS data via the MMSI and IMO, when the latter is available.

In addition, depending on the study, additional data can be integrated, such as the list of black and grey ships established by the Paris Memorandum [10], or the list of known flags of convenience from the International Transport Workers' Federation [11].

In addition, sector-specific data are also integrated. For example, in the case of fishery studies, the identification of fishing situations is based on fishing vessel speed thresholds.

C. World ports and geographical data

The geographical context provided by maritime geographic data, such as Exclusive Economic Zones (EEZs) and global maritime boundaries, is essential for understanding and for extracting meaningful insights from AIS data. These spatial data help to compute and annotate raw vessel tracking information into a rich, contextually-informed data that may reveal complex maritime interactions, regulatory compliance, and spatial patterns of maritime activity. Overlaying AIS trajectories with such geographic information allows the analysis vessel behaviors within different jurisdictional waters, for instance by identifying Illegal, Unreported, and Unregulated (IUU) fishing. Moreover, a detailed geographical understanding of maritime regions—including coastal waters, international straits, and open ocean zones enables more nuanced interpretations of vessel movements, taking into account environmental constraints, maritime infrastructure, geopolitical tensions, and maritime economic zones.

Beyond geographical areas, we also integrated a world port database (around 21,000 entities) which includes port location, name and spatial extension structured in a hexagonal grid derived from the clustering of historical AIS data [12].

III. Platform

This geospatial data processing architecture we designed for maritime mobility analysis integrates multiple technologies and methodological stages (cf. Figure 1).

A. Architecture principles

The system is fed by diverse input data sources, including historical navigation data from AIS and cartographic information. Python serves as the primary preprocessing engine, transforming raw inputs into structured formats suitable for database storage. The architecture relies on two databases. Elasticsearch provides high-performance indexing and rapid search capabilities in historical AIS data, while Neo4j enables

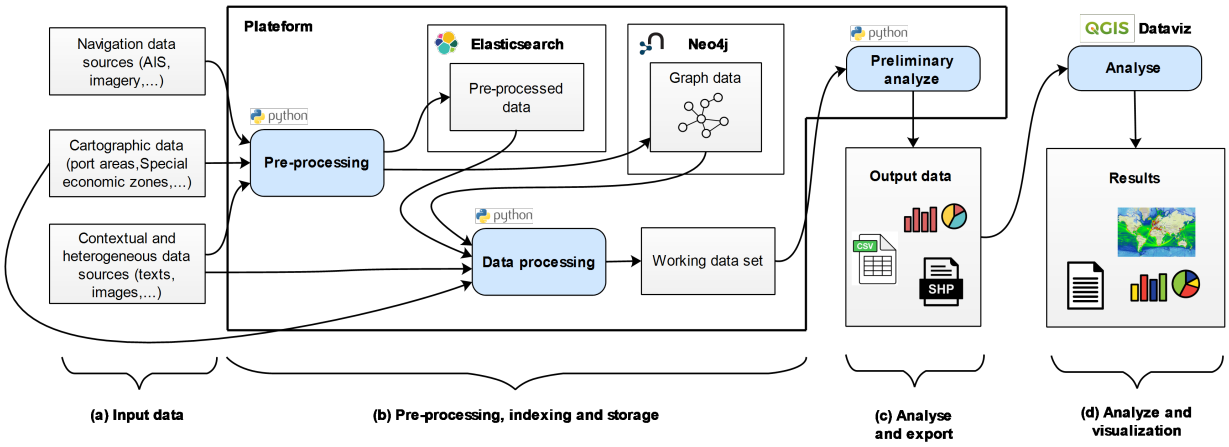

Fig. 1. Schema of the platform.

complex graph-based relationship modeling of maritime trajectories and interactions.

The data processing pipeline transforms and enriches the initial datasets, creating a rich working dataset that captures the dynamics of maritime movements according to use cases (e.g. analysis of fishing activities, international energy transportation). Through Python-driven processing, the system generates working datasets that can be exported in multiple formats including Comma Separated files (CSV), parquets and shapefiles (SHP), allowing for flexible use in programming and geospatial tools. The architecture ends with advanced geospatial visualization using QGIS and specialized data visualization tools, which transform the working dataset into comprehensible maps and charts designed for geospatial use cases. This multi-stage approach ensures robust data handling, from initial collection through sophisticated analysis to final visualization, providing an efficient (mostly automated) approach for the understanding of maritime mobility. The modular design allows for scalability and adaptability in terms of processing steps and databases.

B. Databases

The exponential growth of maritime data volumes poses significant computational challenges in extracting and analyzing historical datasets, constraining users' ability to efficiently retrieve pertinent information. Our platform architecture has been designed to facilitate rapid dataset extraction from large-scale maritime data repositories while simultaneously supporting adaptable prototyping and experimental analytical workflows. To address these complex data management requirements, we implemented a modular infrastructure based on popular technological solutions that optimize computational efficiency, minimize maintenance overhead, and provide scalable data processing capabilities. Our approach relies on three complementary data storage solutions that collectively address the multidimensional challenges of maritime big data management.

- File-based storage: this approach is mainly used for low-volume tabular files (e.g., Parquet, CSV) and geographic files (e.g., Shapefile, GeoPackage).
- ELK Stack platform: this solution is based on the Elasticsearch - Logstash - Kibana suite. Elasticsearch, as a search engine, provides storage and fast retrieval of tabular AIS data. Indeed, Elasticsearch has the advantage of being very fast, thanks to efficient indexing. Kibana offers data visualization and exploration capabilities.
- Neo4j graph database: Neo4j is employed to structure data into Knowledge Graph. We use graphs to represent the maritime network and the operating phases of vessels.

The integration of ELK (Elasticsearch, Logstash, Kibana) and Neo4j databases offers a complementary approach to maritime mobility data analysis. Elasticsearch enables fast and highly scalable storage and searching of raw AIS data, with real-time full-text search and aggregation capabilities. Data ingestion and transformation has been experimented both using Logstash and a mapping in Python. Kibana provides dynamic visualizations and interactive dashboards for data exploration.

In addition, Neo4j brings a dimension of relationship and graph modeling. For maritime mobility data, this means the ability to map complex relationships between vessels, routes, ports, and geographical contexts. Neo4j's graph querying capabilities allow identifying behavioral patterns, tracking complex trajectories and vessel life cycles. The graph has been designed and pre-computed with historical AIS data to represent vessel trajectories on multiple spatial, temporal and semantic scales [13]. This allows deriving mobility models and highlight maritime network structures. At the finest level, the graphs represent individual vessel trajectories. These trajectories are then aggregated at higher levels of abstraction. The model also includes abstraction functions for moving from one level of abstraction to another. These abstraction functions are used to aggregate graph nodes (ports and stationary areas, turning points) and edges to build flow networks. The graph-based maritime network is complemented with a second lightweight

graph model representing the life cycle of a ship's identity [14]. The graph has been designed from static information provided by AIS messages (static and voyage data). This enables fast extraction of identities of interest.

IV. Geospatial data processing

In this section, we specify the main steps of the knowledge discovery process, transforming raw data through iterative stages of selection, cleaning, transformation, mining, and visualization to ultimately generate actionable insights and knowledge .

We based our design and implementations on a GKDD (Geographic Knowledge Discovery in Databases) approach [15]. Indeed, the GKDD method is particularly adapted to large and noisy datasets, which include non-numeric and possibly incomplete data. The following figure (cf. Figure 2) illustrates the various stages of our GKDD approach. The first step is to select the relevant data (according to use case objective). This is followed by a phase of preprocessing, cleaning and organizing the data. The third phase consists in transforming and reducing the volume of data. The fourth phase is data mining itself. The fifth phase is the visualization of results. Finally, the analysis concludes with the evaluation and interpretation of the results, producing new knowledge.

A. Hexagonal indexing

To facilitate both data analysis and the aggregation, visualization, and analysis of large data volumes, navigation data have been indexed within a hexagonal grid derived from the Uber H3 library [16]. This provides several advantages in terms of data structuring and spatial information representation. Indeed, the hexagonal grid will suffer less distortion compared to other grids due to the Earth's curvature [17]. Furthermore, with 16 different hexagon sizes, each covering the entire globe: mesh 0 corresponds to hexagons with an average area of 4,250,546 km², while mesh 15 corresponds to hexagons with an average area of 0.895 m². This varying level of granularity allows for the selection of a mesh suitable for all analysis scales, for all phenomena, and with an appropriate data volume. Each hexagon being subdivided into seven smaller hexagons, multiscale analyses are feasible [18]. Finally, the hexagonal representation enhances the restitution of vessel movements in six equidistant directions (compared to four for a square grid) [19]; this increases the possibilities for understanding vessel behavior.

Consequently, while maritime objects, inherently mobile and evolving within a fluid environment, pose representation challenges, hexagonal indexing offers the ability to model vessel trajectories and better identify and delineate areas of interest [17]. Let us note hexagons have a unique index that unambiguously identifies and delimits each geographical area. Hexagonal cells also play a pivotal role at multiple stages within the data processing pipeline.

B. Data preprocessing

Data preprocessing constitutes the preliminary phase prior to archiving. This process is essential to ensure correct integration of the data into the platform and to optimize its subsequent exploitation. Our approach is to preserve as much information as possible, including anomalies inherent in the source data. This strategy is required for studies focused on detecting of manipulations or irregularities [20], [21]. Consequently, as far as possible, the original data format is preserved wherever feasible. However, this preservation principle encounters limitations, especially concerning AIS data which comes from highly heterogeneous sources. Some AIS providers supply pre-merged and decoded messages, while others deliver raw frames in the NMEA format.

When AIS data is delivered in raw NMEA frame format, a robust parsing process is essential. Our decoder, built upon the Python library PyAIS [22], has been enhanced to ensure detailed logging. We implemented logging functionalities that track the total number of successfully decoded messages, as well as the count of raw NMEA frames that fail to decode. We adopt a conservative approach to handle anomalies. Even when decoded messages contain inconsistencies, such as invalid MMSI numbers (fewer than 9 digits) or aberrant values, they are retained. This ensures that no potentially usable information is lost. Subsequently, these messages can be used in studies relative to AIS tampering, i.e. to determine whether these anomalies are simply due to inerrant AIS errors or to deliberate alterations. Furthermore, for positioning messages, we add a new field corresponding to the transcription of geographic coordinates in H3 indexing. We use resolution 9 (105 m² per hexagon). All decoded messages are archived in the Elasticsearch database.

Geographic data layers typically come in standardized formats such as Shapefile, GeoJSON, or GeoPackage. In these cases, we preserve the original format. Occasionally, some layers contain topological inconsistencies or other errors. When such issues are detected, we archive both the original layer and a corrected version, ensuring data reliability while keeping the unaltered source for reference.

C. AIS data filtering

Analyzing billions of AIS messages is a complex task that requires the preselection of essential data to optimize processing times. Consequently, the filtering operation involves both selecting relevant AIS messages and cleaning the datasets of any inherent errors. These steps correspond to the selection and cleaning stages illustrated in Figure 1. The first step, data selection, is based on criteria such as AIS message type, vessel characteristics, geographical zones (study areas) or time range. The second step focuses on error reduction or elimination to improve data reliability.

Typically, AIS data anomalies can be classified into two main categories [21]: anomalies resulting from errors and intentional anomalies. Unintentional errors arise from various sources, including signal transmission issues, poor synchronization of AIS receivers' internal clocks, malfunctioning

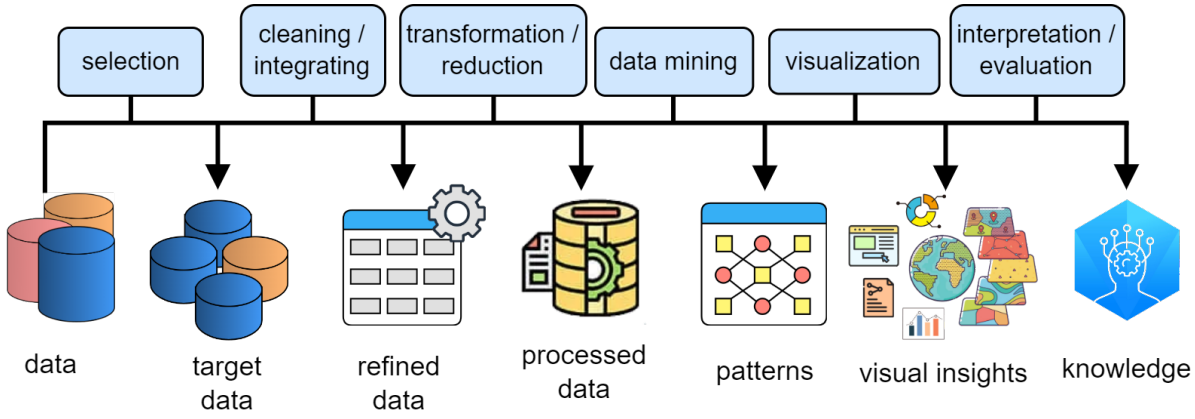

Fig. 2. Knowledge discovery process

equipment producing noisy data, sensor errors, isolated outlier errors, manual data entry mistakes, and the use of the same MMSI number by multiple vessels [23]–[25].

Additionally, inherent errors in the data fusion process, such as duplicates, must also be considered. In contrast, intentional anomalies result from deliberate manipulations, such as data falsification or acts of piracy [26], [27]. The approach to handling these anomalies varies according to the study's objectives. In some cases, researchers seek to eliminate anomalies to ensure clean data, while in others, they want to keep them. For instance, in studies of maritime traffic between ports, we aim to analyze the overall behavior of vessels. In such studies, we seek trends, and it is preferable to remove aberrant data. Conversely, when investigating data falsification, researchers are particularly interested in inaccurate data or detection of missing data. In this context, detecting impossible vessel movements or unexplained stops are valuable indicators.

D. Data cleaning and data integration

During this phase, we correct errors and inconsistencies while simultaneously merging datasets. This process involves harmonizing disparate data sources to resolve conflicts arising from divergences between them. As discussed in the previous section, filtering, cleaning, and integrating AIS data occur simultaneously. Now, let us focus on the processing of administrative and geographical data. Regarding vessel administrative data, the primary objective is to merge information derived from multiple registers. We also verify the consistency between AIS data transmitted by vessels and the registers. This verification process helps detect inconsistencies or anomalies in vessel identification and registration. We achieve this using Python's data analysis library Pandas and the machine learning library Scikit-Learn.

Harmonizing geographical data presents its own challenges, particularly when integrating datasets from different sources. A common issue is that spatial layers often operate at different scales, leading to topological inconsistencies at zone boundaries. To ensure spatial data integrity, we resolve overlaps and gaps between zones. These operations are carried out using Python libraries such as GeoPandas or GIS software like QGIS.

As with AIS data, the need for data homogenization depends on the study's objectives. In certain contexts, such as general maritime traffic analysis, standardization is essential for accuracy. However, in studies focusing on fraudulent activities, preserving irregularities and anomalies is crucial, as these inconsistencies may provide valuable insights into suspicious behaviors.

E. Transformation and reduction

The transformation process is highly dependent on the study case, and the reduction varies between each case. However, some general principles can be established. In most cases, the primary objective is to reduce the volume of data while preserving relevant spatial and temporal information. To achieve this, we use the H3 geospatial indexing system, which can assign any position on Earth to a hexagonal cell with selectable resolution. This ability to adjust resolutions allows us to adapt the grid to different analytical scales and phenomena under study. For instance, this process is particularly effective in generalizing the AIS positions and generalized vessel trajectories. The H3 hexagonal indexing is also employed to aggregate AIS messages into larger hexagons. This facilitates, for example, the creation of small-scale maps, such as fishing zone maps. The adaptive granularity of H3 ensures an optimal balance between spatial precision and computational efficiency. In the case of vessel trajectories, hexagonal aggregation also facilitates temporal aggregation. When multiple consecutive positions of the same vessel fall within the same hexagon, we can retain only the arrival and departure times from the hexagon. So, we convert instantaneous time positions into temporal intervals. This method is particularly effective for calculating the duration of vessel presence in hexagons. Regarding geographical data layers, we transform the geometries to be compatible with the use of H3 hexagons. Finally, for attribute data, we employ the classic technique of dimensionality reduction, selecting only the fields (variables) pertinent to the analysis.

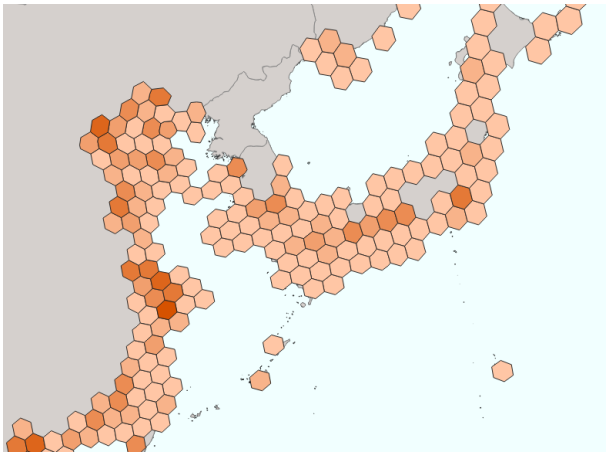

Fig. 3. An example of density map representing the time spent by vessels in stationary situations (speed less than 3 knots and indicated as "*at anchor*" or "*moored*" in the AIS) over the year 2019

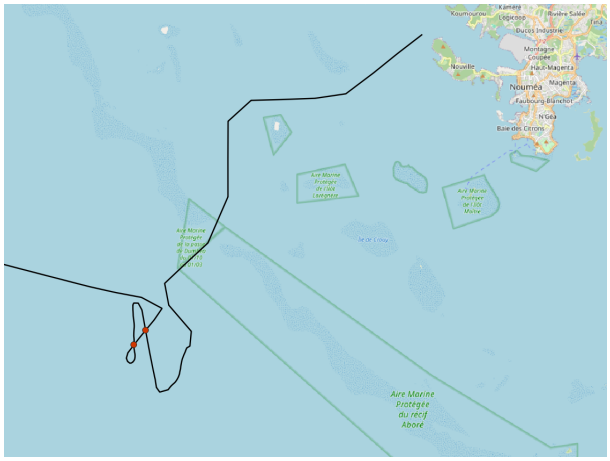

Fig. 4. An example of loops detected by the algorithm

F. Main algorithms

The platform integrates a suite of specialized algorithms. This section provides an overview of the key algorithms implemented. We mainly rely on classical data mining, including segmentation, dependency analysis, deviation and outlier analysis, trend detection, and generalization and characterization.

1) Density map production: Several types of density maps can be produced with varying levels of analytical depth. A density map represents the distribution of a variable over a given surface. To create these maps, the maritime space is divided into hexagonal cells. As previously mentioned, we employ the H3 indexing system, which offers adjustable resolution. The precision and relevance of the calculated statistics depend on several factors, including the size of the data within the area, the hexagonal resolution, and the time interval considered.

A common approach for representing density involves calculating the total duration vessels spend within a given hexagon. By combining vessel positions data with thematic information (such as static data contained in AIS messages or Lloyd's registers), we can produce more sophisticated maps. For example, we have successfully represented gaps in AIS messages from shipping vessels in proximity to reefer ships.

By analyzing vessel navigation patterns, particularly their speeds, we can estimate certain behaviors. This enables the creation of specialized maps related to stationary vessel positions, fishing activities, and more. For instance, 14.6 % of damage to submarine internet cables is caused by vessel anchors [28]. Identifying vessel anchoring zones (excluding port areas) is one of the steps in determining risk areas for submarine cables [29]. The figure illustrates a temporal density map of vessels in anchoring situations (cf. Figure 3).

2) Loop detection: Loops in vessel trajectories can reveal specific behaviors such as search maneuvers, fishing operations, waiting areas, or route changes. They are also indicative of potentially unusual or illegal activities, including unauthorized transshipment, fishing in prohibited zones, or movements suggesting possible illicit maritime activities.

We have developed an algorithm capable of detecting loops in vessel trajectories. To achieve this, we transformed the AIS messages into multi-line string geometries, where each line string represents the path between two consecutive points. By calculating the self-intersection of each line string for a vessel, we successfully identified loops within their trajectories (cf. Figure 4). We experimented the algorithm using all cargo and tanker AIS messages near New Caledonia, focusing on vessels outside ports over a one-year period.

The primary objective of this analysis was to detect instances of illegal ballast water discharges, a critical environmental concern that can severely impact marine ecosystems. By correlating detected loops with draught variation analysis, we aimed to pinpoint suspicious activities that may indicate unauthorized discharge events.

3) Change of identity: Vessels deliberately turning off their AIS and subsequently disappearing represent a critical maritime security challenge, often signaling potential illegal activities such as unauthorized fishing, clandestine transshipments, or embargo bypass. To address this issue, we have developed an algorithm aiming to detect vessel disappearances and reappearances accompanied by an identity change (potentially illicit). To achieve this, when a vessel stops transmitting, we have developed a multi-target hexagonal expansion (multi-threaded from a computer science perspective) with a Gaussian evolution law adapted to the vessel type. The hexagonal coverage expands over time. The appearance of vessels in these expansion areas triggers the search for identity falsification.

4) Terrestrial receiver coverage: Considering terrestrial AIS receivers, one of the obstacles to a perfect understanding of the maritime situation is the presence of shadow zones likely to mask maritime activities and which makes the identification of system misuse difficult [30]. Consequently, we have designed and implemented two algorithms. The first algorithm aggregates historical location information to produce density maps. The result is a set of hexagons (with configurable size), each annotated with statistical information (number of vessels, average speed, etc.). In addition, a learning algorithm

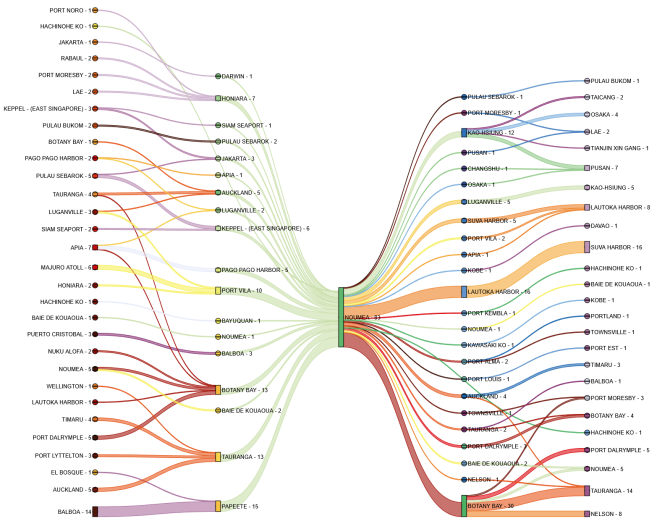

Fig. 5. An example of maritime flow of Noumea port

(based on XGBoost) for predicting the AIS coverage of a receiver has been developed. The algorithm uses historical AIS data coupled with environmental data collected by the European Copernicus program. The data are spatially and temporally aligned using H3 hexagons. The resulting maps (with a configurable time periods) are also hexagon maps predicting receiver coverage under specific weather conditions.

5) Analysis of maritime flows: Understanding maritime flows at different geographic and temporal scales is essential for many strategic, economic, and environmental studies. Maritime flow analysis provides insights into global transportation networks, reveals complex global commerce mechanisms, highlight changes in patterns during pandemics.

To determine daily maritime flows, such as port-to-port, sea-to-sea, or economic zone-to-economic zone, we have designed a generic algorithm capable of quantifying vessel entries and exits in any maritime area. The developed algorithm generates maritime flow dashboards between geographical zones (cf. section II-C). Results are typically presented in tabular format, chord diagrams or as Sankey diagrams (cf. Figure 5), depending on specific requirements.

6) Trajectory prediction: Maritime trajectory prediction represents an important aspect of maritime situational awareness. The development of predictive models that anticipate vessel movements with increasing accuracy has been widely considered in the literature. Many representation models of trajectory data can be considered and processed on a range of AI methodologies. Trajectories can be processed as images for discerning movement patterns or as text-like data for semantically enriched trajectories [6].

In this work, we considered language models for trajectory prediction. We first explored the relevance of machine learning algorithms for trajectory prediction using H3 hexagons where is each vessel trajectory is converted into a sequence of hexagons. We opted for an original approach that transforms the sequence of hexagon indexes in a trajectory into natural language characters (specifically Chinese). We then used the RoBERTa (Robustly Optimized BERT Pretraining Approach) model on these texts for trajectory prediction. The prediction is a sequence of hexagons. We also implemented a similar principle on AIS raw data using GPT2 model for trajectory prediction. We fine-tuned the model on AIS messages of several thousand cargo and tanker ships in the Atlantic Sea, and used it to predict the next message of each ship by looking each precedent messages. On both case, we observed that the prediction work correctly for cargos and tankers on small temporal windows (within a quarter-hour range), far less accurately for the long term.

G. Data exportation

The platform allows data export at each processing stage. Indeed, researches might require to obtain a simple dataset or preprocessed data before applying algorithms. The platform is equipped with a set of export tools. Among simple queries, it is possible to export AIS data according to different criteria, such as vessel name, MMSI, IMO, vessel type, etc. It is also possible to specify a geographical area and a temporal range. By combining queries, we can, for example, answer questions like ”obtain the list of vessels passing through a port on a specific date”. In practice, we can retrieve vessel trajectories over several years. We can perform analyses on vessel origins and destinations. We can also extract navigation patterns, route deviations, etc.

Exports can be done in several formats: tabular files (CSV, Parquet), geographical layers (Shapefile, GeoJSON, GeoPackage). These are mostly analysis results of the data, which serve for visualization and interpretations. Another type of export concerns analysis results on the data. These are generally statistics about the data or indicators.

H. Visualization

The platform relies on three main solutions for data visualization, each serving a specific purpose and complementing the others to provide a comprehensive analytical workflow.

The first solution is Kibana. Kibana is a data visualization plugin that operates with Elasticsearch. Consequently, Kibana is primarily employed to explore AIS data (as a reminder, the Elasticsearch database primarily serves for storing decoded AIS messages). Kibana is accessible via a web browser and allows in-depth dataset exploration through a user-friendly graphical interface. One of Kibana’s key strengths is its ability to generate automated dashboards, incorporating statistical visualizations (histograms, time series, graphs, donut charts, etc.) as well as maps integrating geospatial information. Its simplicity makes it ideal for basic searches and exploratory analysis. However, it shows its limitations, especially when dealing with heterogeneous data sources, which require more advanced processing and integration capabilities.

The second solution involves creating Jupyter Notebooks, which enable flexible and in-depth data analysis during processing and computation phases. These notebooks integrate

data visualization libraries, allowing researchers to immediately visualize results, refine their analyses, and rapidly identify trends or anomalies. This approach facilitates dynamic data exploration, ensuring that insights can be adjusted and refined. Within the platform, we primarily use the following data visualization libraries: Bokeh, Folium, HoloViews, Matplotlib, Plotly, and Seaborn. Generally, the data resulting from these analyses are small and can be exported into files readable by more common tools like Excel.

The third solution is the visualization and analysis of results in QGIS. QGIS is an open-source GIS software that offers a multitude of spatial analysis and geographic data visualization functionalities. In our studies, QGIS is used to visualize geographical layers and to format information (applying graphical semiology rules). QGIS offers the capability to cross-reference heterogeneous datasets: processed AIS data results and external contextual data that varies according to the case studies. Additionally, the numerous geoprocessing capabilities are offered by QGIS: distance calculation, hexagon centroid calculation, buffer zones, layer merging, location-based extraction, etc.

A key feature of the platform's geospatial visualization approach is the use of hexagonal cells. Hexagons are particularly suited for map creation, and their hierarchical structure facilitates data visualization at different scales.

V. On Use Cases

The various algorithms developed and applied to AIS data, coupled with Lloyds data, stored on the platform ELK, have shed light on different use cases related to maritime safety. Cybersecurity has been the subject of reflection through the analysis of maritime activities occurring above submarine internet cables. Indeed, the ability to select vessels based on their types, speeds, and flags, considering the geopolitical and/or physical context (areas of high or low bathymetry) and the location of cables, has enabled the development of a vulnerability index for the Internet submarine cables in Pacific (VISIC) [29]. In this case study, algorithms related to density and vessels navigation time for each hexagonal cell were particularly utilized.

In the domain of environmental safety, the processing chain and various algorithms (e.g. loop detection) have been employed not only to analyze density and presence (to understand the interactions between maritime incidents and navigation areas) but also to examine vessel trajectories to detect illegal ballast water discharge. Additionally, geographic proximity between different vessel types and AIS cut-off durations were calculated to approximate IUU (undeclared and unregulated) fishing activities. The results of these analyses were combined, leading to the creation of a Maritime environmental vulnerability index (MEVI) [31].

Current use case research focuses on securing trade and exchanges, based on the complementarity between port, maritime flows algorithm and identity graphs. This makes it possible to identify suspicious interactions between ships and services.

VI. Conclusion

This paper has presented a platform, coupled with a methodological and technical approach, dedicated to the management and analysis of spatiotemporal maritime data. In response to the challenges posed by the heterogeneous and massive nature of maritime data, our algorithms enable the analysis and the interpretation of mobility dynamics, vessel behaviors, and complex maritime situations.

We have detailed the strategies for data integration, preprocessing, and storage, highlighting the importance of an architecture suited to the constraints of big data. Moreover, the integration of diverse processing algorithms significantly enhances our ability to generate datasets, extract pertinent knowledge, and visualize and interpret the results.

The explored case studies demonstrate our capability to produce analyses for a variety of applications, particularly in maritime security and surveillance of illicit activities. By providing precise insights into vessel mobility, our system contributes to better decision-making in identifying illicit activities and managing environmental risks.

Future work will focus on representing vessel identities through time and improving the process of linking records from different data sources. The goal is to enhance the reliability of vessel tracking, enabling their unique and precise identification.

Acknowledgment

This work belongs to Maritime Geospatial Intelligence project (GEOINT). The authors thank the French Defense Innovation Agency (AID) for their funding and support.

References

[1] C. Ray, R. Dréo, E. Camossi, A.-L. Jousselme, and C. Iphar, "Heterogeneous integrated dataset for maritime intelligence, surveillance, and reconnaissance," *Data in brief*, vol. 25, p. 104141, 2019.

[2] N. K. B. Krismentari, I. M. O. Widyantara, N. I. ER, I. M. D. P. Asana, I. P. N. Hartawan, and I. G. Sudiantara, "Data pipeline framework for ais data processing," in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, 2022, pp. 1–6.

[3] I. Lytra, M.-E. Vidal, F. Orlandi, and J. Attard, "A big data architecture for managing oceans of data and maritime applications," in *2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. IEEE, 2017, pp. 1216–1226.

[4] A. Artikis and D. Zissis, Eds., *Guide to Maritime Informatics*. Springer, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-61852-0

[5] A. Troupiotis-Kapeliaris, C. Kastrisios, and D. Zissis, "Vessel trajectory data mining: A review," *IEEE Access*, vol. 13, pp. 4827–4856, 2025.

[6] A. Graser, A. Jalali, J. Lampert, A. Weißenfeld, and K. Janowicz, "Mobilitydl: a review of deep learning from trajectory data," *GeoInformatica*, vol. 29, no. 1, pp. 115–147, 2025.

[7] H. Liu, X. Chen, Y. Wang, B. Zhang, Y. Chen, Y. Zhao, and F. Zhou, "Visualization and visual analysis of vessel trajectory data: A survey," *Visual Informatics*, vol. 5, no. 4, pp. 1–10, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468502X21000401

[8] M. Series, "Technical characteristics for an automatic identification system using time-division multiple access in the vhf maritime mobile band," *Recommendation ITU: Geneva, Switzerland*, pp. 1371–1375, 2014.

[9] C. Ray, C. Iphar, A. Napoli, R. Gallen, and A. Bouju, "Detection of ais spoofing and resulting risks," in *OCEANS'15 MTS/IEEE*, Gênes, Italie, May 2015, p. 6.

[10] P. MoU, "Performance list - paris memorandum," https://parismou.org, 2025, accessed on March 27, 2025.

[11] I. Seafarers, "Home - itf seafarers," https://www.itfseafarers.org, 2025, accessed on March 27, 2025.

[12] A. Ménard, V. Lambert de Cursay, C. Ray, C. Guenois, M. Maslek Elayam, and M. Dréau, "Construction d'une méta-base de ports à l'échelle mondiale," in *Conférence internationale de Géomatique et d'Analyse Spatiale*. SAGEO'21, May 2021.

[13] M. M. Elayam, C. Ray, and C. Claramunt, "A hierarchical graph-based model for mobility data representation and analysis," *Data & Knowledge Engineering*, vol. 141, p. 102054, 2022.

[14] A. Amouriq, M. Maslek Elayam, and C. Ray, "Vessel's identity graph and analysis," in *Graph Embedding and Mining workshop (GEM@ ECML PKDD)*, 2021, p. 10.

[15] H. J. Miller and J. Han, *Geographic Data Mining and Knowledge Discovery*. USA: Taylor & Francis, Inc., 2001.

[16] I. Brodsky, "H3: Uber's Hexagonal Hierarchical Spatial Index," https//www.uber.com/en-FR/blog/h3/, 2018.

[17] M. M. Elayam, G. Kerhoas, V. L. d. Cursay, C. Ray, and A. Ménard, "On the interest of hexagonal abstraction of maritime information," in *OCEANS 2022, Hampton Roads*, 2022, pp. 1–6.

[18] A.-C. Duverger, C. Carré, C. Ray, and J.-M. Kowalski, "Intelligence géospatiale maritime et sciences de l'espace géographique : regards croisés," in *Mesurer l'espace et après (Géopoint 2024)*, 2024, p. 16.

[19] D. Tsatcha, É. Saux, and C. Claramunt, "A bidirectional path-finding algorithm and data structure for maritime routing," *Int. Journal of Geographical Information Science*, vol. 28, pp. 1355–1377, 2014.

[20] C. V. Ribeiro, A. Paes, and D. de Oliveira, "Ais-based maritime anomaly traffic detection: A review," *Expert Systems with Applications*, vol. 231, p. 120561, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423010631

[21] C. Iphar, A. Napoli, C. Ray, E. Alincourt, and D. Brosset, "Risk analysis of falsified automatic identification system for the improvement of maritime traffic safety," in *ESREL 2016*. Taylor & Francis, 2017, pp. 606–613. [Online]. Available: https://api.semanticscholar.org/CorpusID:53959163

[22] M. R. Leon, "Pyais library," https://pypi.org/project/pyais, 2025, accessed on March 27, 2025.

[23] H. Greidanus, M. Alvarez, T. Eriksen, and V. Gammieri, "Completeness and accuracy of a wide-area maritime situational picture based on automatic ship reporting systems," *The Journal of Navigation*, vol. 69, no. 1, pp. 156–168, 2016.

[24] T. Emmens, C. Amrit, A. Abdi, and M. Ghosh, "The promises and perils of automatic identification system data," *Expert Systems with Applications*, vol. 178, p. 114975, 2021.

[25] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, "Automatic identification system (ais): Data reliability and human error implications," *the Journal of Navigation*, vol. 60, no. 3, pp. 373–389, 2007.

[26] C. Ray, R. Gallen, C. Iphar, A. Napoli, and A. Bouju, "Deais project: Detection of ais spoofing and resulting risks," in *OCEANS 2015-Genova*. IEEE, 2015, pp. 1–6.

[27] C. Iphar, A. Napoli, and C. Ray, "A method for integrity assessment of information in a worldwide maritime localization system," in *19th AGILE International Conference on Geographic Information Science (AGILE 2016)*, 2016.

[28] L. Cater, D. Burnett, S. Drew, G. Marle, L. Hagadorn, D. Bartlett-McNeil, and N. Irvine, "Submarine cables and the oceans-connecting the world. unep-wcmc biodiversity series no. 31," ICPC/UNEP/UNEP-WCMC, Tech. Rep., 2009.

[29] A.-C. Duverger, J.-M. Kowalski, and C. Ray, *L'intelligence géospatiale maritime, une méthode de veille des câbles sous-marins*. Presses universitaires de Rennes, 2025, 20 pages.

[30] L. Salmon, C. Ray, and C. Claramunt, "Continuous detection of black holes for moving objects at sea," in *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoStreaming*, 2016, pp. 1–10.

[31] A.-C. Duverger, L. Tesorio, C. Ray, C. Carré, and V. Litoux, "A gis analysis of environmental security: application to the french territories of the south pacific," in *Oceans 2025 Brest*. IEEE, 2025, p. 8.