ENVISIA – A SCALABLE ARCHIVING SYSTEM FOR AIS DATA STORAGE

Yann Guichoux, Romain Gallen
F. Bajard-Jacobs, P. Doare, G. Dufil
CENTRE D'ÉTUDES TECHNIQUES MARITIMES ET FLUVIALES
2, boulevard Gambetta – BP 60039- 60321 Compiègne cedex - France
Tel : +33(0)3 44 92 60 30

Abstract

Within the context of an increasing demand for sharing maritime informations, the CETMEF has designed a scalable archiving system, called ENVISIA, which is able to store a large amount of AIS data for a long period. Putting the accent on the end-user needs and the ease of use of the provided data set, ENVISIA rely on an architecture that is following the example of similar initiatives relatives to the monitoring of Environment, like the Europe's GMES programme.
This paper describes the main characteristics of the CETMEF's system, and provides some technical elements for discussion, related to sharing non-real-time maritime data.

1 ENVISIA – A scalable archiving system for AIS data storage

In order to store an always increasing amount of *AIS* data, the *CETMEF* has designed a scalable archiving system which provides efficient tools to access the collected informations. **Using as a template** the ***GMES*[1]**/*MERSEA*[2]*&ARGO*[3] (*Global Monitoring for Environment and Security*) European multinational programmes in link with Environmental real-time in-situ measurements at sea, the *CETMEF* system called ***ENVISIA* is able to store statics and dynamics informations as ship trajectories** and is **ready to process embedded data provided by binary messages**, such as meteorological and hydrographic data or new application messages. Such a storage system is suitable to enhance the ease of use of historical *AIS* data, **facilitating the growth of value added services, such as risk analysis and environmental studies**.

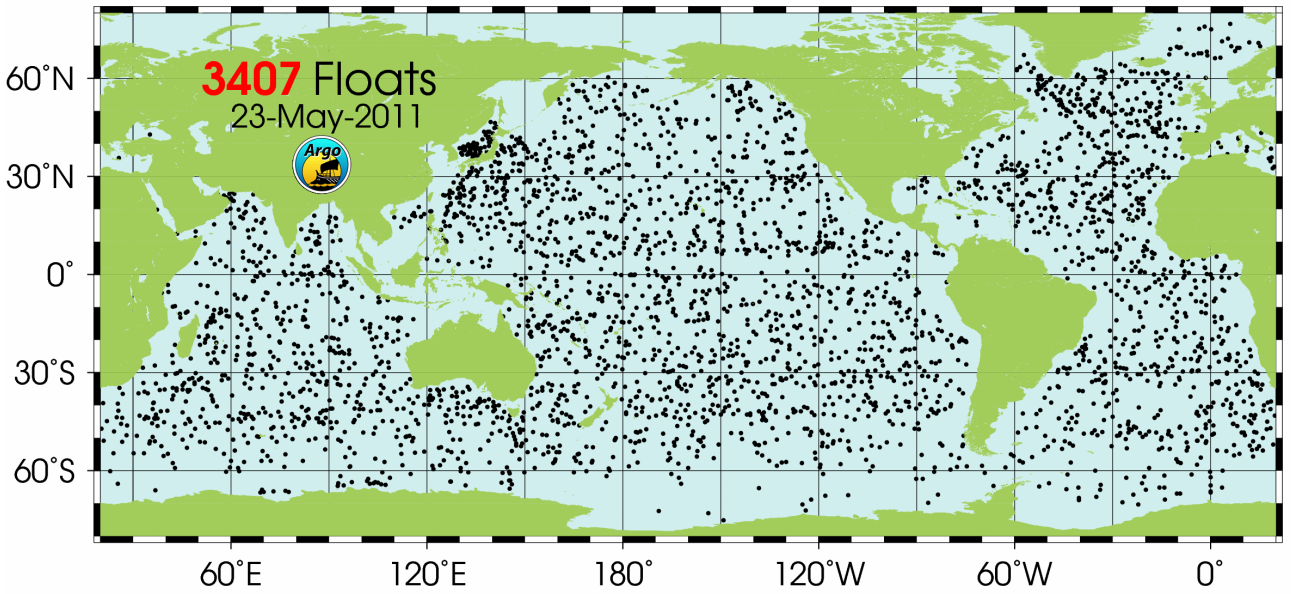

Figure 1: ENVISIA is built to the same design as ARGO or MERSEA projects, parts of the Europe's GMES initiative. Within the framework of Argo, thousands of Floats have been spread over the seas to monitor environmental parameters. These floats have very similar properties with ships (like dynamics spatial informations). (Credits : www.argo.net)

2 What does scalability stand for in ENVISIA ?

One major **principle of *ENVISIA* is the orientation towards information flow**, i.e. the thinking in information sources, information storage and retrieval. Working on the assumption that the end-user is in the best position to design the tools he needs to analyse the *AIS* data, *ENVISIA* only provides an enhanced *AIS* data set, easy to extract and to work with.

AIS data are collected from several providers (which can be *PSS* or *CLSS* (*Physical*/*Central Logical Shore Stations*), using multiple *TCP/IP* connexions to transfer the data in the usual *IEC/NMEA* format (*IVEF* [4] format is planned to be implemented in the system within the next months).

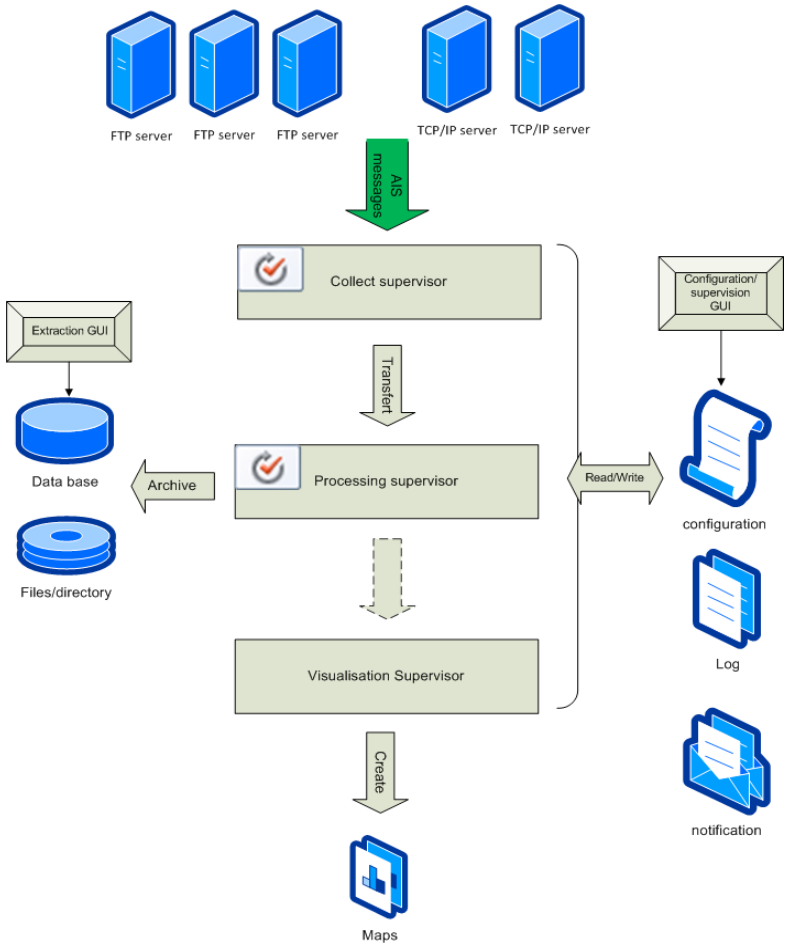

Figure 2: ENVISIA main architecture. The accent has been put on modularity and simplicity to facilitate the future evolution of the system (like processing binary messages embedded data).

These data are then filtered, processed, enhanced and stored in a repository allowing an easy access to the searched data.**« Filtered » is the main key word** in the previous sentence. The implementation of spatial and temporal filters allows the real-time classification of the received *AIS* messages in a hierarchical directory tree. **Each branch of the tree deals with its own storage characteristics** (see figure 3).

These characteristics fit the main end-users' needs and requirements :

- a full sampling rate of received *AIS* data in areas where accurate information can be requested. For instance, this storage directory branch could relate to a small geographical area with short-life storage (like a focus on a traffic separation scheme),
- a low sampling rate branch can relate to larger areas where you only needs macro statistics . This branch can include already stored data included in smaller areas, for an extended period of time. Stored ships trajectories can be reduced using specific algorithms, preserving there main characteristics (like speed or course change),

Furthermore, to prevent from having tied hands in case of a particular request about *AIS* messages which are not stored in the *NetCDF* processed files described below, *ENVISIA* can **keep the *raw* data**, after compression, with a specific life duration.

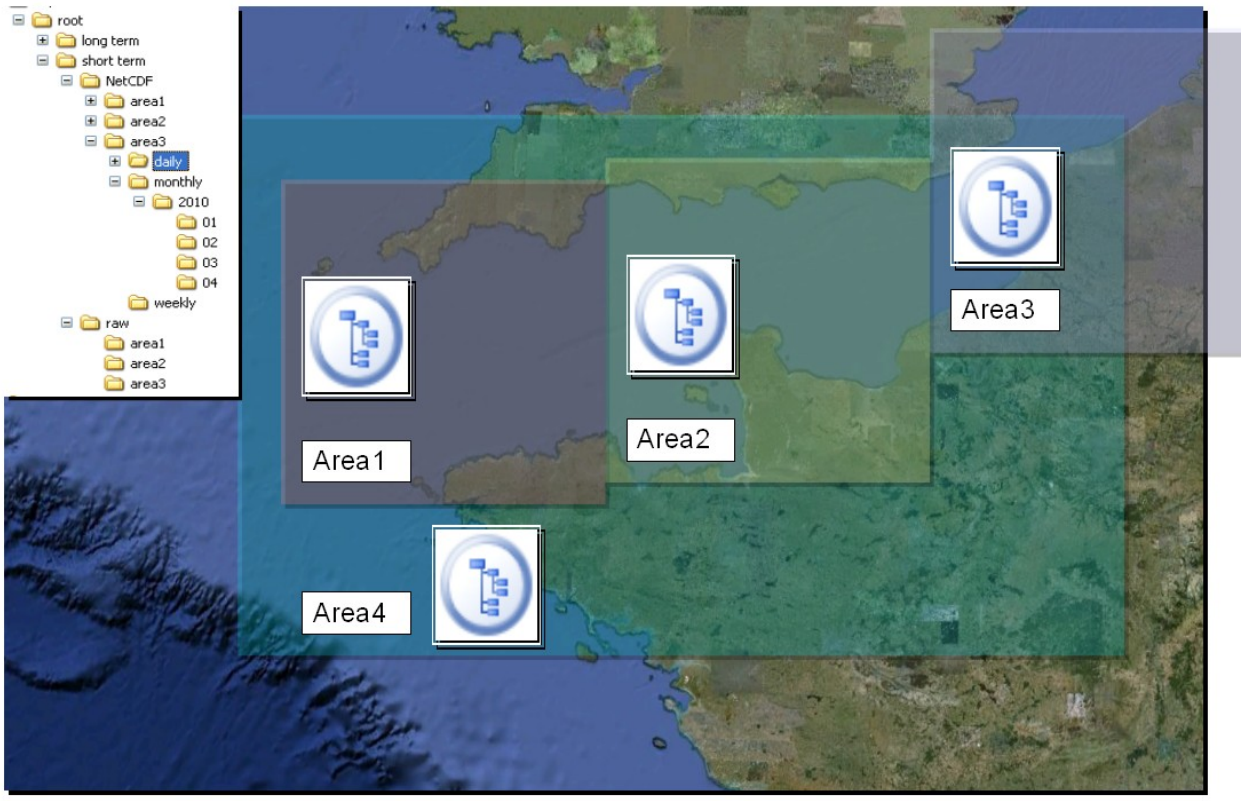

Figure 3: Each geographical area of interests has its own directory tree (cf. top-left corner), with specific AIS data sampling rates and life durations. As an example, every month, data from files stored in the branch Area1, Area2, and a part of the data stored in Area3 can be undersampled, merged and stored in the branch Area4. Oldest files stored into Area1 and Area2 can be automatically deleted while files stored in Area3 can be kept for a longer period.

3 What about a « classic » database storage ?

Database are time consuming and expensive to maintain. In light of those observations, storing each *AIS* message in a database could be counterproductive, regarding the large amount of data transmitted by ships, and considering that you needs to append informations to the database to make the extraction process more efficient. *ENVISIA* is storing the data in such a manner that you does not need to create database tables to allow the users to access **the more common searched informations, mainly related to geographical areas and time periods**.

However, with an eye to service enhancement, like creating a web portal with a *Graphical User Interface* to improve the ease of consulting the *AIS* stored dataset, one can limit the implementation of the data extraction system to an indexing system easier to set-up and maintain. This indexing system (indexing the stored files and their main metadata, like MMSI, ship types, etc.) could rely on a lighter database (such as *MySQL*, *Oracle*, etc.).

4 ENVISIA data format

The *ENVISIA* data format is relying on ***NetCDF* (*network Common Data Form)*** [5], a **portable and self describing structure format**. As a first step, the CF (Climate and Forecast) [6] metadata convention is implemented by the system. This standard has been adopted by a number of projects

and groups involved in the Environment field, such as *MERSEA* or the *WDCC* (*World Data Center For Climate*).

Technically speaking, « *NetCDF is a set of interfaces for array-oriented data access and a freely-distributed collection of data access libraries for C, Fortran, C++, Java, and other languages. The netCDF libraries support a machine-independent format for representing scientific data. Together, the interfaces, libraries, and format support the creation, access, and sharing of scientific data.*
NetCDF data is:

- *Self-Describing. A netCDF file includes information about the data it contains.*
- *Portable. A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.*
- *Scalable. A small subset of a large dataset may be accessed efficiently.*
- *Appendable. Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.*
- *Sharable. One writer and multiple readers may simultaneously access the same netCDF file.*
- *Archivable. Access to all earlier forms of netCDF data will be supported by current and future versions of the software. »*

5 A bunch of ideas regarding data transport and exchange procedure

From an end-user point of view, the **stored data should be as complete as possible**, meaning that an archiving system designed to store real time data, transient by nature, should minimise the risk to lose data which cannot be transmitted again to avoid gaps in the collected data sets (whatever is the nature of the data sets : ships position reports, meteorological, etc.).

Since *AIS* data transmission is mainly relying on a real-time transfer protocol (at least from the *PSS* to the *LSS*), the risk of data loss in case of communication **network failure should not be underestimated**. The longer the network path between the *AIS* receiver and the storage system is, the more important is the risk of data loss.

It is the reason why a **decentralised** but compatible **architecture**, relying on archiving systems maintained at a national level could minimise this risk. Each stakeholder/country participating to an historical *AIS* data exchange, on a delayed base, could store its real time data and share the files using a repository. As an example, data could be accessed using common files transfer protocols (like *FTP*).

Furthermore, to allow a remote access to the data, the projects mentioned before (*ARGO*, *MERSEA*), selected **the Opendap technology** [7] as a good vector for data transmission.
« ...

- *"OpenDAP" stands for "Open-source Project for a Network Data Access Protocol". OpenDAP is both the name of a non-profit organization and the name of the protocol which the OpenDAP organization has developed (see http://www.opendap.org). It has been widely used, **serving the marine community since 1995** and serves already a broad range of data, including oceanographic, atmospheric, and even astronomical data (cf. the Opendap master data sets list).*
  Open DAP is an operational component of IOOS for access to gridded data. OpenDAP servers are available for download without licensing costs.
  Note: OPeNDAP was formerly called DODS (Distributed Oceanographic Data System).
- *The Opendap technology allows the users, where ever he is, to access whatever data they require in a form they select (image, ASCII, NetCDF, etc.), using client applications they already are familiar with (IDL, Matlab, Ferret, netCDF Operators, spreadsheets operators,*

LAS , MapServer, etc.).Alternatively, a simple internet browser may be used as a client and allows to download data from the server.

- *OpenDAP data server is a middleware (XML native + java servlets) that has been designed to distribute large data sets on Internet (uniform access to scientific data on the Internet - HTTP protocol -) and minimize the barriers to sharing them. It converts transparently from a number of commonly used data formats (NetCDF, GRIB, HDF4, MATLAB®, ASCII), into the format appropriate for the analysis package. OpenDAP also allows a client to request only a subset of a dataset (and aggregation) (selection of parameters at given space and time windows, compose a time series) as well as function evaluation like statistics functions (basic operation between parameters, mean computation, use of thresholds, etc.). At the other end, an OPeNDAP client, integrated into familiar analysis and visualization software packages access the data.*
- *The Opendap technology provides strong support for data stored in NetCDF as well as for users of NetCDF enabled programs.*

... »

This technology, not yet implemented in *ENVISIA*, may be well fitted to provide historical *AIS* data information to the end-users.

6 Conclusion

To achieve the goal of defining a multinational/multi-purpose standard format well-fitted to process efficiently AIS archived data, and reach data harmonization, there is a need for an iteration process with strong coordination between the various actors. ENVISIA has been designed to fulfil the user-needs, always keeping in mind that AIS data is interesting different thematic communities. Because of the development of more and more innovative applications relying on the binary messages capabilities, such as in-situ Environment measurement and diffusion, it is important to find a balance and to merge the practices and experiences of these thematic communities into a coherent set of data that will take into account specific features of the products and provide simultaneously interoperability to these surrounding communities. That is :

- to describe the *AIS* resulting products in such a manner it will help to reach coherency, enhancing the informations provided by adding « discovery metadata »,
- to fulfil a common data format structureand harmonise and standardise its description, which will ease exchange and joint use of AIS data sets and derived products,
- to harmonise data transport and exchange procedure, that is the ability to access the data in an interoperable manner from client applications, relying on a decentralised but compatible system architecture for distribution on a public or private network. [2]

References

[1] GMES - Global Monitoring for Environment and Security – www.gmes.info
[2] MERSEA – Marine EnviRonment ans Security for the European Area – www.mersea.eu.org
[3] ARGO – part of the integrated global observation strategy – www.argo.net
[4] IVEF – Inter VTS Exchange Format – http://openivef.org
[5] NetCDF – Network Common Data Form - www.unidata.ucar.edu/software/netcdf
[6] CF – NetCDF Climate and Forecast Metadata convention - http://cf-pcmdi.llnl.gov
[7] OpenDaP - Open-source Project for a Network Data Access Protocol – www.opendap.org