OCEAN DATA MANAGEMENT: INFORMATION SYSTEM WITH FLEXIBLE STRUCTURE FOR MANAGEMENT AND SUPPORT FOR INTERNATIONAL DATA STANDARDS AND QUALITY ASSURANCE

Fernández Hermida, Xulio(1)
Lamas Pose, Sonia(2)
Lago Reguera, Manuel D.(3)
Lodeiros Vázquez, Darío(4)
(1)Signal Theory and Communications Department. University of Vigo. xulio@tsc.uvigo.es
(2)HCTECH, R&D department, sonialamas@herculescontrol.com
(3)HCTECH, R&D department, mdlago@herculescontrol.com
(4)Aloxa CEO, dario@aloxa.eu

Abstract- *The oceanographic and weather databases, whether historical or new, have multiple causes for statistical weaknesses: interruptions, errors of measurements, gaps and discontinuities. The interest of the scientific community focuses on time series that have the longest and the best quality possible, so they seek to unite databases from different places and times and that were gathered by different methods of measurement. All these issues further complicate the representativeness and validity of the data. Therefore, there is an international agreement involving major scientific reference entities to establish standards that ensure the quality of ocean-meteorological data. Our group made a proposal that leverages new information technology systems to facilitate management of the data generated by our stations, to ensure their quality according to international standards, and to facilitate post-processing and integration with other databases.*

Keywords- *ITC, ocean data, data standards, no-sql database, ocean monitoring*

I. INTRODUCTION

In January 2008, the "International Oceanographic Data and Information Exchange" (IODE) of the "Intergovernmental Oceanographic Commission" (IOC) of UNESCO and the Joint Commission for Oceanographic and Marine Meteorology (JCOMM) held a forum on Ocenanographic Data Management and Exchange Standards generating the project Ocean data Standards (ODS) (http://www.oceandatastandards.org/) whose objective is to reach a broad agreement and commitment to adopt a set of standards related to the management and exchange of oceanographic data. There are numerous institutions, research centers and international organizations that generate, maintain and manage oceanographic, marine and meteorological extensive databases. Each of these entities has its own systems to determine the quality of data stored, and assigned a code ("flag") that identifies each data or data sets as "verified by quality control", "not verified", "absent", "interpolated" and so on. This initiative aims to establish a system for identifying unique data quality, for sharing databases, add new information sources, and increase the reliability of the data sets that are available.

As providers of ocean-meteorological information, we implemented a project to strengthen and streamline the storage and information management to facilitate the validation of our data internationally.

In this paper we describe the current situation of international agreements for generating reliable and intercomparable ocean data sets, the qualitiy control tests for the validation of data, and our proposal for an information system architecture that will automatically manage and transform raw metocean data from our stations into validated data sets acordingly to those international agreements.

II. CURRENT SITUATION

Data quality control, or data validation, is a stage in data management which is essential whenever data are used by any individual or group other than the originators of the data. With the recent growth in large scale collaborative oceanographic research programmes both in Europe and globally, quality control of data is essential. Without it data from different sources cannot be combined or re-used to gain the advantages of integration, synthesis, and the development of long time series [1].

The identification system for the quality of the data has two levels of information: a primary level, which refers to the identification of quality level itself, and a secondary level, which reports on the justification of this classification (which is why that data have a certain level of quality) [2].The quality of the data depends mainly on two issues: sampling protocols, which basically depend on the proper functioning of equipment and measuring instruments and their proper use, and random errors or deviations in measurements by environmental or unexpected causes (signal drift caused by fouling effect on a sensor, for example).

Quality control tests, or QC-tests, are necessary on the functioning of the system (Gap-test: evaluating whether the measure has come off or not, Sintax-text: that evaluates whether the message comes complete with the information chain intact) and tests on the measurements, like basic statics assesing if they are inside or outside the measuring range of the sensor or instrument, and statistical deviations from average monthly, quarterly, annual, multivariate... in a given dataset or compared to previous series). In any case, the goal is to achieve a system that labels each data in real-time according to the results of these tests.

The four major aspects of metocean data validation are:

a) Instrumentation checks and calibrations which

include calibration /checks of sensor response; tests on instrument or system electronics; and checks on data processing and recording equipment.

b) The documentation of deployment parameters which includes definition of the location and duration of the measurements; method of deployment of the instrumentation; and sampling scheme used for the measurements.

c) Automatic quality control of data which comprises a series of tests on the data to identify erroneous and anomalous values in order to establish whether the data have been corrupted in any way, either during initial measurement, or in copying or transmission to a user.

d) Oceanographic and meteorological assessment which includes an assessment of the results of conditions a) to c); and an assessment of the oceanographic and meteorological 'reasonableness' of the data, comprising checks on expected patterns or trends and comparisons, with other data sources. [1]

Leading the management and processing of marine and oceanographic data in Europe is the British Oceanographic Data Center (BODC). Among others, they are developing a project called NETMAR, that aims to develop a pilot European Marine Information System (EMIS) for searching, downloading and integrating satellite, in situ and model data from ocean and coastal areas. It will use a semantic framework coupled with ontologies for identifying and accessing distributed data, such as near-real time, model forecast and historical data. The focus of this project, as for de IODE standadization efforts, relies on standardising data and metadata formats, as well as exchange protocols as the first steps to bridge existing marine data systems [3]. The technology used for storing data is an ORACLE sql data warehouse. Regarding the quality control tests to be performed, the MyOcean GMES FP7 project is standardizing near real time quality control procedures for operational oceanography purposes, and they achieved a list of recommendations for real-time QC procedures [4].

Our systems for monitoring water quality, Hidroboya, capture and transmit the information gathered in real time, without any prior processing. The final user is who performs the statistical analyses according to the objectives he looks for. We store all met-ocean data collected and maintain a database that can be further processed. If we label our data with the "flags" that are being agreed by IODE/BODC after a standardized quality processing, we will have a more valuable product, and high quality environmental information.

The project that we are undertaking intends to migrate our information management systems from SQL to a new database structure that will facilitate the possibilities of computing, information processing, filtering and verification of the quality of each data with speed, flexibility and reliability. The proposal is based on open-source solutions that handle non-relational databases and data-mining software for information processing.

III.THE PROJECT

We work on implementing a system that will speed up the realization of the QC test automatically. This system will allow the application of automated QC tests, inmediately, and it will offer accessible data with the corresponding quality verifiers. It will also be flexible to adapt comunication protocols for conecting to other data warehouses or web services.

The strategy for designing our new data management system approaches the key issues from three levels: 1-infraestructure, 2-data warehouse and data mining and 3-information management.

The infraestructure will be based on OpenStack open source tools. The technology is a cloud operating system that consists of a series of interrelated projects that control pools of processing, storage, and networking resources throughout a datacenter, all managed through a dashboard that gives administrators control while empowering its users to provision resources through a web interface, for a cloud infrastructure solution [5].

The data warehouse is built under a non-relational database (non-sql) structure. It gives the flexibility to adapt the stored data schema to each particular case, to each client, or to any worldwide standard that is adopted now or in the future. This storage scheme system ensures the traceability of data: its origin, time of capture and measurement, measuring instruments, calibration of these instruments any other circumstances afecting the sampling itself. It also allows us to incorporate two important safety features in information systems management: replication and high availability. The labeling for data storage, instead of the rigid structures in relational database tables, provides ease of use of Data Cluster and, GridFS (storing files as graphics). This tools increase the speed and efficiency of data processing compared to relational databases. Combining this data architecture and software development to extract, filter, combine and process information, we will implement the functionalities needed for giving each measurement its quality flag in real time.

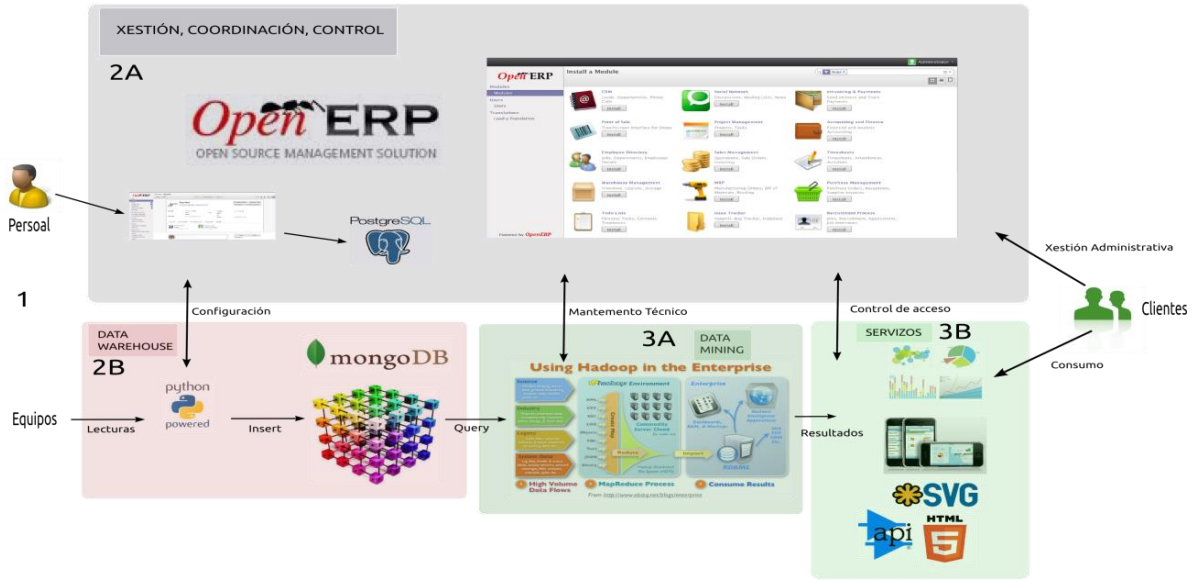

Fig. 1: Information system architecture scheme.

There are open-source software tools, with huge computational possibilities, like MongoDB that facilitate the management of high volumes of data from oceanographic stations. We will then proceed to integrate the data warehouse with an ERP system that will monitor the proper functioning of instruments. In combination with a datamining tool such as Apache Hadoop it will allow us to manage large amounts of information, both their own readings, as monitoring logs, security and integrity of communications, or equipment states. This system architecture level will give us the tools for monitoring and configuring the instrumens on a network of moored buoys or hidrological stations and for automatically label each data with its corresponding QC test results regarding the data validation aspects a) to c) described in section I.

This data mining tool will also assess statistical analysis, for detecting and describing patterns, trends, and for providing high value tools and services like:

- Facilitate the implementation of quality control of data sets prior to being made available to the client, through statistical analysis, detection and / or elimination of outliers, interpolation to complete series, etc. ...
- Apply filtering statistical analysis (QC test) requested by the customer.
- Label each data according to their "level" of quality according to the standards of IODE, or according to any system of labeling required by the customer.
- Provide the customer a personalized treatment that acquires data using an API: the final user will process data with services developed by himself. This system allows a particular user to utilize its own services (web, processing) on the raw data stored (MongoDB) or treated with Apache Hadoop.

REFERENCES

[1] IOC, "Manual of quality control procedures for validation of oceanographic data". Intergovernmental Oceanographic Comission. Commision of the European Comunities. Manual 26. SC-93/WS-19 , UNESCO 1993.

[2]Paris. Intergovernmental Oceanographic Commission of UNESCO. 2013.Ocean Data Standards, Vol.3: Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data. (IOC Manuals and Guides, 54, Vol. 3.) 12 pp. (English.)(IOC/2013/MG/54-3) .

[3]http://netmar.nersc.no, NETMAR project web page.

[4]Pouliquen, Sylvie, Data-MEQ working group. "Recommendations for in-situ data Real Time Quality Control ", EuroGOOS-RTQC pdf file.

[5]http://www.openstack.org/, Openstack official site.