

The Urgent call on data management: are we capable to store valuable (meta)data for naval application?

Sonia Papili¹

¹Royal Military Academy, Robotic & autonomous Systems. Graaf Jansdijk 1, 8380 Brugge. Belgium. Email: Sonia.papili@mil.be

Sonia Papili. ¹Royal Military Academy, Robotic & autonomous Systems. Graaf Jansdijk 1, 8380 Brugge. Belgium. Email: Sonia.papili@mil.be

Abstract: *Technology is increasing rapidly and new concepts as robotics, artificial intelligence (AI) or augmented reality are entering the daily life. Nevertheless, the maritime world is not immune to the disrupting advancing in technology and information production especially regarding shipping industry, military, and monitoring methodologies. The basic will is automatization, efficiencies, and safety at affordable costs. Therefore, in parallel, other aspects are evolving, mostly related to interoperability and data management. An enormous amount of information is produced and the necessity to store in an appropriate way is demanding. In this contest, an interdisciplinary environmental data model for mine countermeasure application is proposed where heterogeneous environmental information is integrated as a base structure for a future data management system. The attention was posed not only on measurable data related to various disciplines such as physics, chemistry, geology, biology, and acoustics, but also on historical data and subjective interpretation such as descriptions and lesson learned. In the model a distinction was made between “humans” and “machines”. The humans are seen as entities that must collect and store data, meanwhile the machines are seen as entities that query information. In this way the model shows what to take into consideration when to collect and store data, how to integrate data following rules and style conventions, how to build a conceptual table with heterogeneous information, what*

are the tools to allow the interrogation of measured, historical, and subjective-descriptive data.

The possibility to automatically access heterogeneous data volume opens a broad spectrum of possibilities, from a characterization of a marine environment to the representation of a marine environment by virtual reality, to a sensor performance prediction, to risk analysis.

Keywords: *data model, historical information, descriptive data, descriptive information.*

1. INTRODUCTION

Although large amount of information is available within different disciplines in marine science, collecting the right information for naval purposes is a demanding and tedious process. As highlighted by Yan et al. in 2014 the difficulties encountered are mainly due to the heterogeneity of marine data being multi-scale, multi-temporal and multi-semantic. Several tentative are on-going to tackle this matter following different strategies.

The United Nations are working on the Global Ocean Observing System (GOOS) with the idea to achieve integrated observations and data sharing.

Yan et al. in 2014 used machine learning model to disentangle the difficulties refining the correspondence relationship between data.

A new International Hydrographic Organization (IHO) is attempting a series of standards developments based on a new universal hydrographic data model, S-100 (Ponce, 2018),

In this framework and considering the new challenges in maritime monitoring, an interdisciplinary data model is hereby presented.

The model clarifies how to collect and store heterogeneous data from various disciplines such as physics, chemistry, geology, biology, and acoustics.

It also identifies tools and methodologies necessary to have proper interrogations of measured, historical and subjective-descriptive data related to marine environments. Moreover, it highlights how personal and subjective perspective can affect descriptive information and metadata collection.

Hereafter, the different topics are illustrated separately. A section on heterogeneous data is illustrated, subsequently a section on historical data and afterwards a section on subjective and descriptive data are discussed. Finally, an overview on the full model is given.

HETEROGENEOUS DATA

To investigate the feasibility and the modalities for constructing an interdisciplinary data model for environmental characterization, information on different domains was extracted from databases and scientific literature and subsequently collected into a comprehensive table. This table is subdivided using a pyramidal structure containing different level of information. The first level describes the scientific domain, the second level specifies the scientific fields, and the last level describes the parameters listed in the table.

The extraction of information from scientific literature presented major difficulties. Lack of standardization results in discharging of information, duality in wording creates substantial confusion. As example it is reported the use of the word “resolution” in relation to sonar images. Some authors refer to the word resolution as “pixel resolution” whilst others refer to it as the “theoretical resolution” of the sonar system.

Therefore, not only a descriptive analysis on what data to collect, but also a detailed analysis on how to collect and store the data is considered in the model.

HISTORICAL DATA

Historical data brings a high potential amount of information.

The possibility to have historical data coupled with recent data will allow to have a broader dataset for deep learning and automatic processing. The term ‘historical’ does not necessarily refer to information made in the past, but eventually to data observed with technologies or software retained obsolete and not more in use. In that case, some recommendations are listed to facilitate a straightforward integration of data.

Conformity between the data is a primary requirement for a good integration of data. Within the context of this model, conformity is related to three main domains: methodology to collect and rework the gathered information, geographical positioning systems and collection of metadata.

Conformity in methodology

To reclassify historical laboratory results to modern one, a good method is to perform modern laboratory analysis to a set of historical samples. The new results would give a scale of comparison between new and ancient way of classification.

An interesting example was made by Houziaux et al. in 2008 during the review of the Gilson samples collection. Sixty-two sandy samples from the collection were re-analyzed in laboratory for grain size analysis.

Conformity in wordings

To verify a conformity in ancient and modern scientific language, a good method is to perform a new description on already described historical data and compare the two descriptions.

Conformity in geographical positioning

When the position accuracy is not satisfying the modern techniques, it is necessary to build a re-sampling strategy. A reference area is re-sampled and subsequently used to verify the historical data with current data.

An interesting approach to estimate the value of ancient maps is proposed by Leriche et al. in 2004. She calculated a Reliability Index (RI). This index weighs three parameters: the scale of the map, the data acquisition method and the geographical positioning system method. Using this method, she could estimate the reliability of the ancient maps.

Next to conformity, historical data need to be labelled by a *quality rank coefficient* expressing the reliability of the information and to assign a coefficient of uncertainty giving a weight to the information self.

Furthermore, considering the fast evolution of new technologies and software application, it is advised to introduce policies for the conservation of historical data to allow a fast and smooth integration into current or future datasets.

2. SUBJECTIVE DESCRIPTION AND LESSON LEARNED

With the idea to develop a method to feed subjective information into the model, an experiment was performed to weigh how individual background, and subjective perspective may affect a scientific description. Voluntaries holding different background as engineers, ecologists, military officers and schoolteachers, gently offered their contribution to this semantic experiment. They were required to make an environmental description watching two video fragments showing the seafloor of the Belgian Continental

Shelf. The simulation was built in two phases. A free description of the videos was requested during the first phase whilst a questionnaire was provided during the second phase. The resulting descriptions were analysed following different criteria. Differences and analogies were highlighted. Attention was posed on the use of the vocabulary and on the syntax. The results highlighted useful insights for the model. To complete the understanding on descriptive information, technical documents reporting lesson learned from maritime surveys were analysed.

MODEL

This model highlights several aspects related to data management.

Considering the insights collected during the different phases of the construction of the model, a substantial choice was made.

The model is separated in two sections that eventually can cross each other: the first section focusses on “humans”, the second on “machines”.

The humans are seen as entities that must collect and store data, meanwhile the machines are seen as entities that query information.

On the section dealing with the collection and storage of information, attention is posed to integrate data following rules and style conventions. Mostly, conventions are related to data format, measuring units and avoiding duality in wording. When historical information is stored, annotation on used technology, used software and used formats are essential. Dedicated template would facilitate this process. The introduction of policies for the conservation of historical data would allow a smooth integration into future datasets.

On the section dealing on what is necessary to have interrogable information, attention is posed on the availability of structured data, conformity and quality rank coefficient for historical data and dedicated vocabularies considering subjective and descriptive data.

Particularly, scientific vocabularies, classification of random wordings, use of the adjectives and adverbs and synonyms needs to be integrated in computer capabilities.

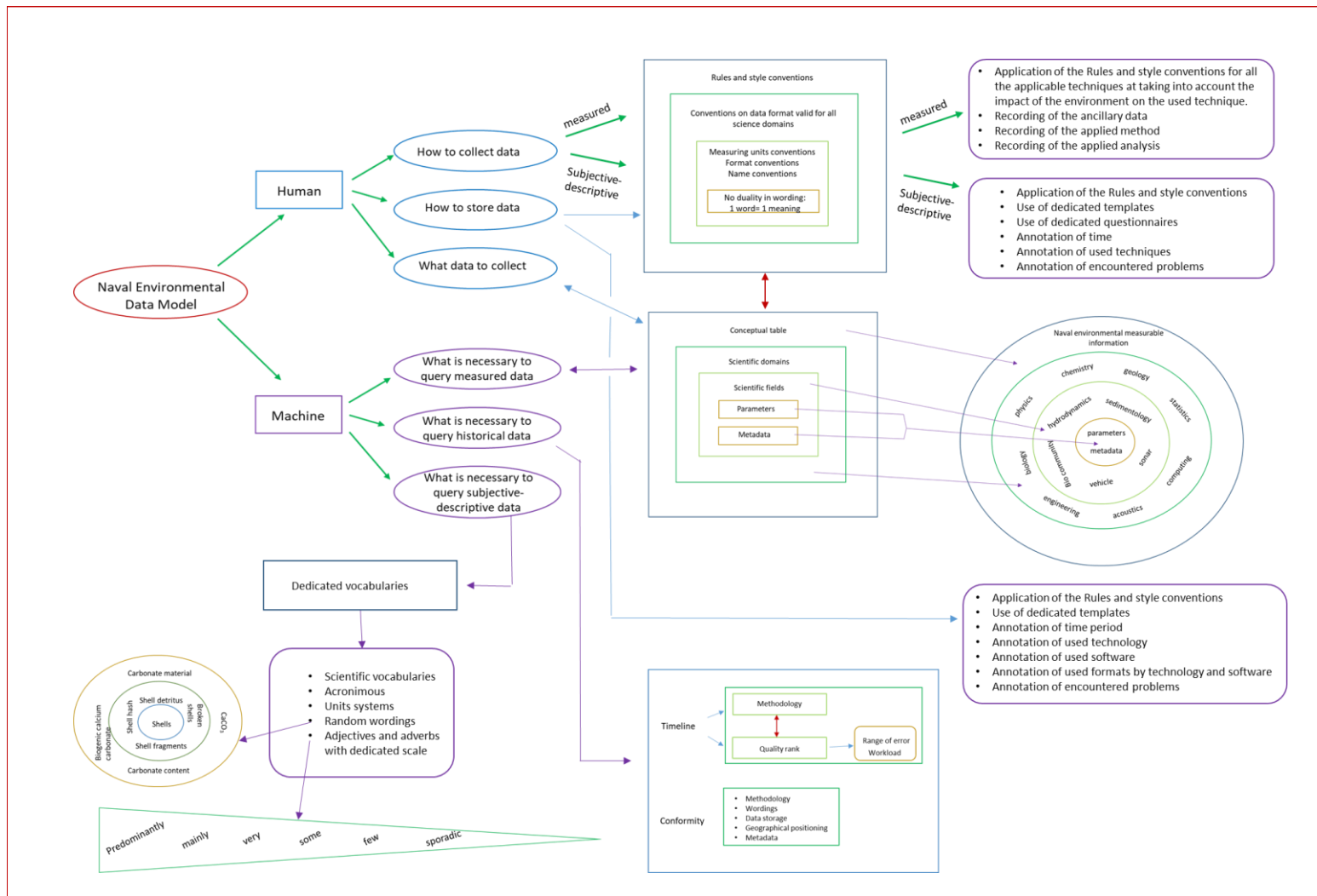


Fig.1. Conceptual model. A framework wherein measured, subjective and historical data are harmonized together.

ACKNOWLEDGEMENTS

The author thanks the volunteers taking part to the semantic investigation, the department of synthesis-evaluation lesson learned for the contribution on extracting information from technical lesson-learned reports.

This paper contributes to the projects DAP19-07 and DAP 22-16 funded by the Belgian Ministry of Defense, Royal Higher Institute for Defense.

REFERENCES

- [1] **Yan,W., Jiajin, L., Yun, Z.** A multianalyzer machine learning model for marine heterogeneous data schema mapping. *Scientific World Journal*. doi: 10.1155/2014/248467. (2014)
- [2] **UNESCO.** United Nations Decade of Ocean Science for Sustainable Development (2021-2030). <https://en.unesco.org/ocean-decade>
- [3] **Ponce, R.** The Maritime World Enters the Fourth Industrial Revolution. *Sea-Technology*. (2018).
- [4] **IHO.** Standards for Hydrographic Surveys.. International *Hydrographic Bureau*. Monaco. Special Publication n° 44 (2008).
- [5] **Houziaux, J.-S., F. Kerckhof, K. Degrendele, M. Roche, A. Norro,.** The Hinder Banks: yet an important area for the Belgian marine biodiversity? *Belgian Science Policy* D/2008/1191/7. (2008)
- [6] **Leriche A., Boudouresque C-F., Bernard G., Bonhomme P., Denis J.** A one-century suite of seagrass bed maps: can we trust ancient maps?. *Estuarine, Coastal and Shelf Science. Elsevier*. 59. 353-362 (2004).