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

Sonia Papili[1]

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

Sonia Papili. [1]Royal 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.

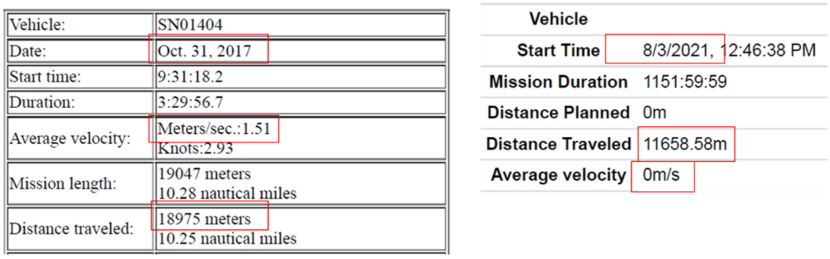
Hereafter, the different topics are illustrated separately. A section on heterogeneous data is illustrated, subsequently a section on historical data and then a section on subjective and descriptive data are discussed. Finally, an overview on the full model is given.

2. 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 scientific domain such as physics, chemistry, geology, acoustics, computing and others; the second level specifies the scientific fields such as sedimentology, bio-communities, hydrodynamics, sonar properties, vehicle properties; and the last level describes the parameters and the metadata important from an environmental perspective.

The extraction of information from scientific literature and free accessible databases presented major difficulties. Lack of standardization results in discharging of information, and duality in wording creates substantial confusion. Common example of lack of standardization is often encountered among software's or among updated versions of the same software giving different formats for the same entity (Fig.1). Duality in wording is often encountered among different disciplines where different meanings are attributed to the same word. It is often a semantic problem that can create confusion and subsequently errors. 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.



Vehicle:	SN01404
Date:	Oct. 31, 2017
Start time:	9:31:18.2
Duration:	3:29:56.7
Average velocity:	Meters/sec.:1.51 Knots:2.93
Mission length:	19047 meters 10.28 nautical miles
Distance traveled:	18975 meters 10.25 nautical miles

Vehicle	
Start Time	8/3/2021, 12:46:38 PM
Mission Duration	1151:59:59
Distance Planned	0m
Distance Traveled	11658.58m
Average velocity	0m/s

Fig.1.Example of lack of standardization. Different formats are here illustrated for the same software in a time range of four years. Differences in entities like "date", "distance travelled" and "average velocity" are evident.

It would be good policy to use always the same name convention and format convention for data collection and analysis.

The application of Rules and Style conventions of the International System of Units (SI) if an example of it. Other examples could be the Climate and Forecast (CF) Conventions, CF standard Names, ACDD Conventions. Furthermore, important is to record which applied method is used and to note which analysis and ancillary information were done.

3. 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 and the digitalization of the samples collection of prof. G. Gilson. Sixty-two sandy samples from the collection were re-analyzed in laboratory for grain size analysis. The resulting measurements were confronted with those made by prof. Gilson and a comparison scale was made.

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 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.

A good example on how to attribute a quality rank is given by Kint et al. in 2020 when under the umbrella of the TILES project (TILES consortium 2018a) they have introduced data uncertainty for horizontal positioning, sampling methods and age of information. Quality rank ranging from 1 (very uncertain) to 5 (very certain) called quality flags were associated to historical measurements. Those quality ranks were converted into corresponding uncertainty percentages. Those percentages were associated to the historical data to weigh the precision of the information.

4. 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 questionary 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. Among all the result of the analysis, interesting is to highlight that a conversational style using first personal pronoun is used when the subject is not confident with his/her observation (I see, I guess, I think, I have no idea). Different semantic is used to describe the same entity: “broken shells, shell detritus, shell fragments”. Differences in styles: structured or lose. This analysis together with the previous ones highlights the important of using dedicated style conventions also for descriptive information.

5. MODEL

This model (Fig. 2) highlights several aspects related to data management. It clarifies how to collect and store heterogeneous data. It also identifies tools and methodologies necessary to have proper interrogations of measured, historical and subjective-descriptive data related to marine environments and gives insights on how human and machine skills can contribute to achieve the best from heterogeneous collection of information.

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.

When referring to collecting and storing 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, the introduction of policies for the conservation of historical data would allow a smooth integration into future datasets. Annotation on used technology, used software and used formats are essential for this integration.

Dedicated template would facilitate the process to collect descriptive information. Annotation of time, encounter problems and used techniques would be valuable ancillary information.

Considering what is necessary to have to be able to interrogable information, the attention in the model is posed on the availability of structured data, conformity and quality rank coefficient for historical data and dedicated vocabularies for descriptive data.

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

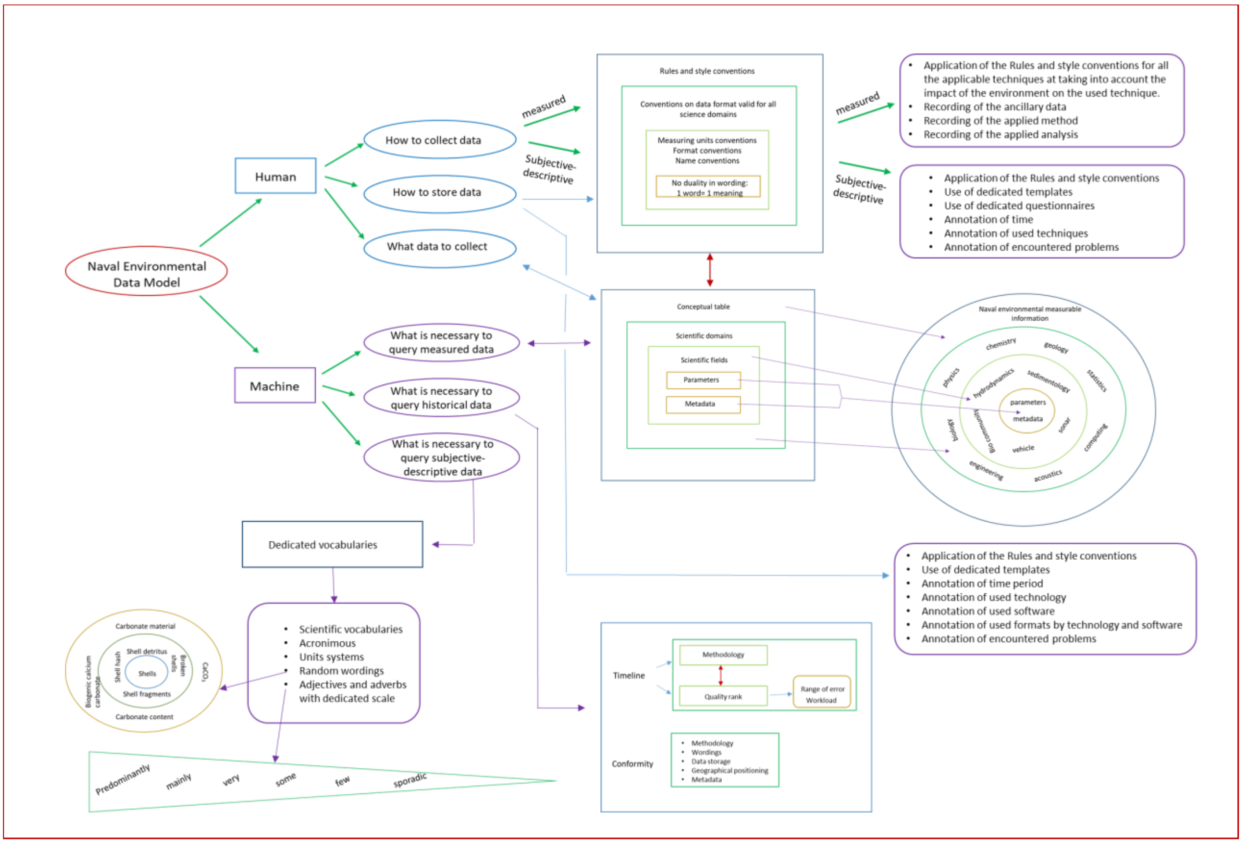

Naval Environmental Data Model
Human
Machine
How to collect data
How to store data
What data to collect
What is necessary to query measured data
What is necessary to query historical data
What is necessary to query subjective-descriptive data
measured
Subjective-descriptive
Rules and style conventions
Conventions on data format valid for all science domains
Measuring units conventions
Format conventions
Name conventions
No duality in wording: 1 word= 1 meaning
measured
Subjective-descriptive
Application of the Rules and style conventions for all the applicable techniques at taking into account the impact of the environment on the used technique.
Recording of the ancillary data
Recording of the applied method
Recording of the applied analysis
Application of the Rules and style conventions
Use of dedicated templates
Use of dedicated questionnaires
Annotation of time
Annotation of used techniques
Annotation of encountered problems
Conceptual table
Scientific domains
Scientific fields
Parameters
Metadata
Naval environmental measurable information
chemistry
geology
physics
statistics
hydrodynamics
sedimentology
parameters
metadata
biology
Bio community
sonar
computing
vehicle
engineering
acoustics
Application of the Rules and style conventions
Use of dedicated templates
Annotation of time period
Annotation of used technology
Annotation of used software
Annotation of used formats by technology and software
Annotation of encountered problems
Dedicated vocabularies
Scientific vocabularies
Acronimous
Units systems
Random wordings
Adjectives and adverbs with dedicated scale
Carbonate material
Shell detritus
Shell hash
Broken shells
Shells
Shell fragments
Biogenic calcium carbonate
$CaCO_3$
Carbonate content
Predominantly
mainly
very
some
few
sporadic
Timeline
Methodology
Quality rank
Range of error
Workload
Conformity
Methodology
Wordings
Data storage
Geographical positioning
Metadata

Fig.2. Conceptual model. A framework wherein measured, subjective and historical data are harmonized together

6. CONCLUSION

The model results on a list of insights and recommendation for future application in deep learning. A unique approach is presented that has the potentiality to be a key concept into the marine data management. The insights presented can be used at different levels and for different purposes in the maritime world.

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 through virtual reality, to a sensor performance tools to risk analysis. If dedicated vocabularies would be developed, automatic understanding of subjective information could be evaluated during the experimentation of new methodologies at sea or new technologies.

7. 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 Journa*l. 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).
- **[7] Kint L., Hademenos V., De Mol R., Stafleu J., van Heteren S., Van Lancker V**. Uncertaintly assessment applied to marine subsurface datasets. *Quarterly Journal of Engineering Geology and Hydrogeology. 54*. (2020). https://doi.org/10.1144/qjegh2020-028