

A Syntactic-Semantic-Pragmatic layered data quality framework for reliable, flexible, and traceable maritime data quality assessment

Ke Wang^{1,*}, Xiaocai Zhang^{1,*}, Ong Qi Hao Tristan², Zhe Xiao¹, Xiaoyang Wei¹, Xiuju Fu^{1,#}, Zheng Qin¹

¹Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Singapore 138632, Republic of Singapore

²School of Computing, Singapore Polytechnic, 500 Dover Road, Singapore 139651, Republic of Singapore

*the authors contribute equally to the work

#Corresponding author, fuxj@ihpc.a-star.edu.sg

The advent of Shipping 4.0 is driving a surge in the generation and accumulation of extensive maritime shipping operation data. These data, integral to maritime digitalization, underpin critical transformations and applications, including maritime big data analytics and AI-based solutions, for smart shipping, autonomous shipping, green shipping, and the development of next-generation ports.

Despite these advancements, effectively harnessing the substantial volume of maritime data presents challenges, primarily attributable to variations in data quality across diverse maritime systems. The absence of thorough and consistent data quality examination and monitoring, especially the failure to detect issues resulting in data mis-matching, along with the necessity for high data quality in the early stages of AI application development, can significantly impact the efficient development and confidence in artificial intelligence (AI) applications within maritime systems. This challenge is exacerbated by the inherent complexity of data quality issues.

To address these challenges, we developed a maritime-specific data quality assessment framework featuring reliability, flexibility, and traceability. This framework integrates a Syntactic-Semantic-Pragmatic (SSP) layered structure and multi-dimensional data feature representation. Within this framework, data quality is stratified into three layers: the syntactic (conforming to metadata and maritime data quality standards), semantic (degree of representing real-world entities considering context information), and pragmatic layer (fulfilling specific application needs). In each layer, data quality is assessed and described within five dimensions, namely precision, completeness, uniqueness, and accuracy, with the consideration of exemplar datasets in maritime systems. The inclusion of maritime domain knowledge, such as maritime data standards, maritime databases, and maritime applications, enhances the reliability to identify maritime-specific issues. The overall mutually-exclusive and collectively-exhaustive framework provides adaptable integration capabilities, allowing for customization to suit varying data and organizational needs. Moreover, the framework provides quantitative evaluation of data quality, enables tracking throughout its lifecycle and promotes transparency and accountability in data processing processes.

The effectiveness of the proposed data quality assessment framework is demonstrated through its application to AIS data, which stands as a critical dataset within the maritime sector. Data entities from

different sources were analyzed, evaluated, and compared with this framework. By offering practical insights into data quality realities and trends across different regions, the framework contributes significantly to enhancing data quality awareness.

As one of the first maritime data quality assessment frameworks, our method seamlessly integrates syntactic, semantic, and pragmatic considerations. This approach facilitates robust diagnosis of data quality issues in the maritime context and allows adaptable evaluation of data assets as well as continuous monitoring of data quality quantitatively. Ultimately, our framework empowers the development of reliable maritime AI applications, which is essential for the maritime digital transformation process.

Keywords— data quality, data standards, maritime digitalization, AI applications, AIS data, data analytics.