A Framework for Computational Reproducibility in Environmental Science with Support for Machine Learning Applications

Zhengyuan Chai[1], Ivana Ivanova[1,2], Fang Yuan[2], Yongze Song[3], Ayalsew Zerihun[4]

Introduction

Many satellite data users face challenges in efficiently accessing, preprocessing, and utilising the growing availability of data from space for decision-making at local, regional, and national levels[1]. Various efforts are underway to produce Analysis Ready Data (ARD) for the effective utilisation of Earth Observation (EO) data. However, these initiatives vary in geographical coverage, durations, and specialised research objectives, often inhibiting reproducibility[2]. There is an urgent need to establish a comprehensive framework compliant with findable, accessible, interoperable, and reproducible (FAIR) principles for computational workflows aimed at maximising the potential of EO products.

Aims

The research aims will be fulfilled by pursuing four distinct objectives:

Framework Design: Create a FAIR compliant workflow designed for generating ARD EO data within a representative environment monitoring.

Implementation: Implement the workflow for generating ARD EO data within a representative environmental context, ensuring the framework's robustness and functionality in environmental scenarios.

Requirement Definition: Define the metadata requirements necessary for achieving FAIR ARD, outlining the essential elements for both metadata and workflows involved in their production.

Testing and evaluation: Evaluate the workflow against the FAIR principles for computational workflows of the ARD framework.

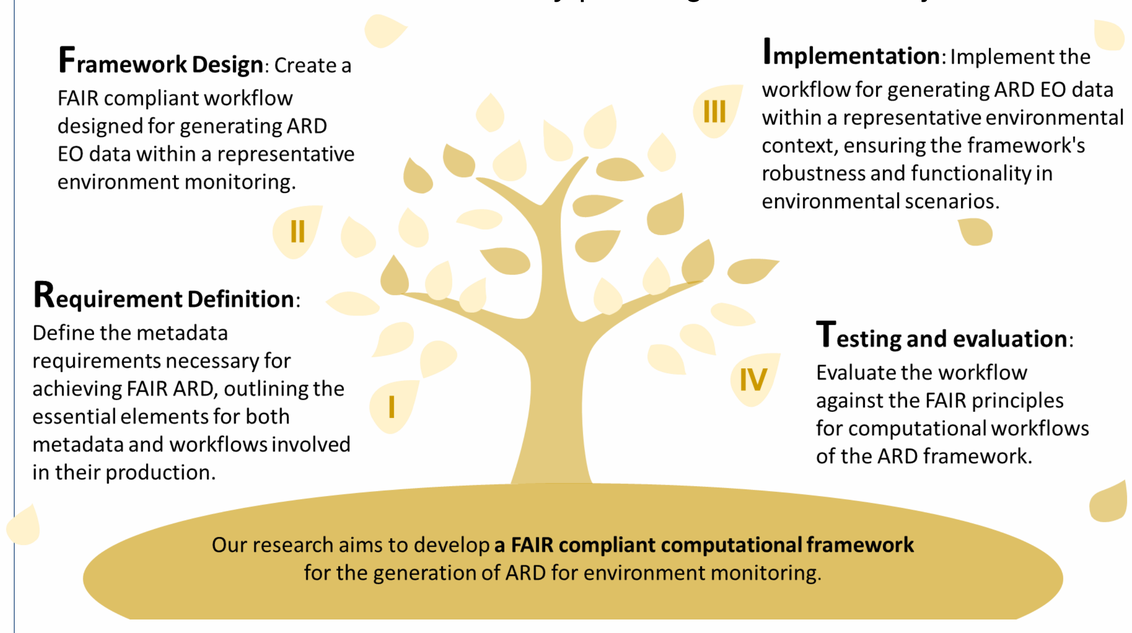

Fig. 1. Overview of the four research objectives

Expected output

A FAIR compliant framework for producing ARD to maximize the use of EO; preprocessing workflows of ARD made transparent to the data consumers, as the workflows are automatically visualised in a Web-based system.

Benefits:

- ✓ Enhanced use of ARD leads to more efficient and effective environmental management strategies.
- ✓ Actively promoting the development and translation of technical specification requirements into the International Organization for Standardization (ISO) or Open Geospatial Consortium (OGC) standards.

Methods

The development of the framework will follow the general software development phases, including requirement analysis, design, development, and testing to produce ARD. These ARDs are grouped into four levels with support for machine learning applications[3], unlocking the potential of EO products in the environmental sector and beyond, as described in Fig. 1.

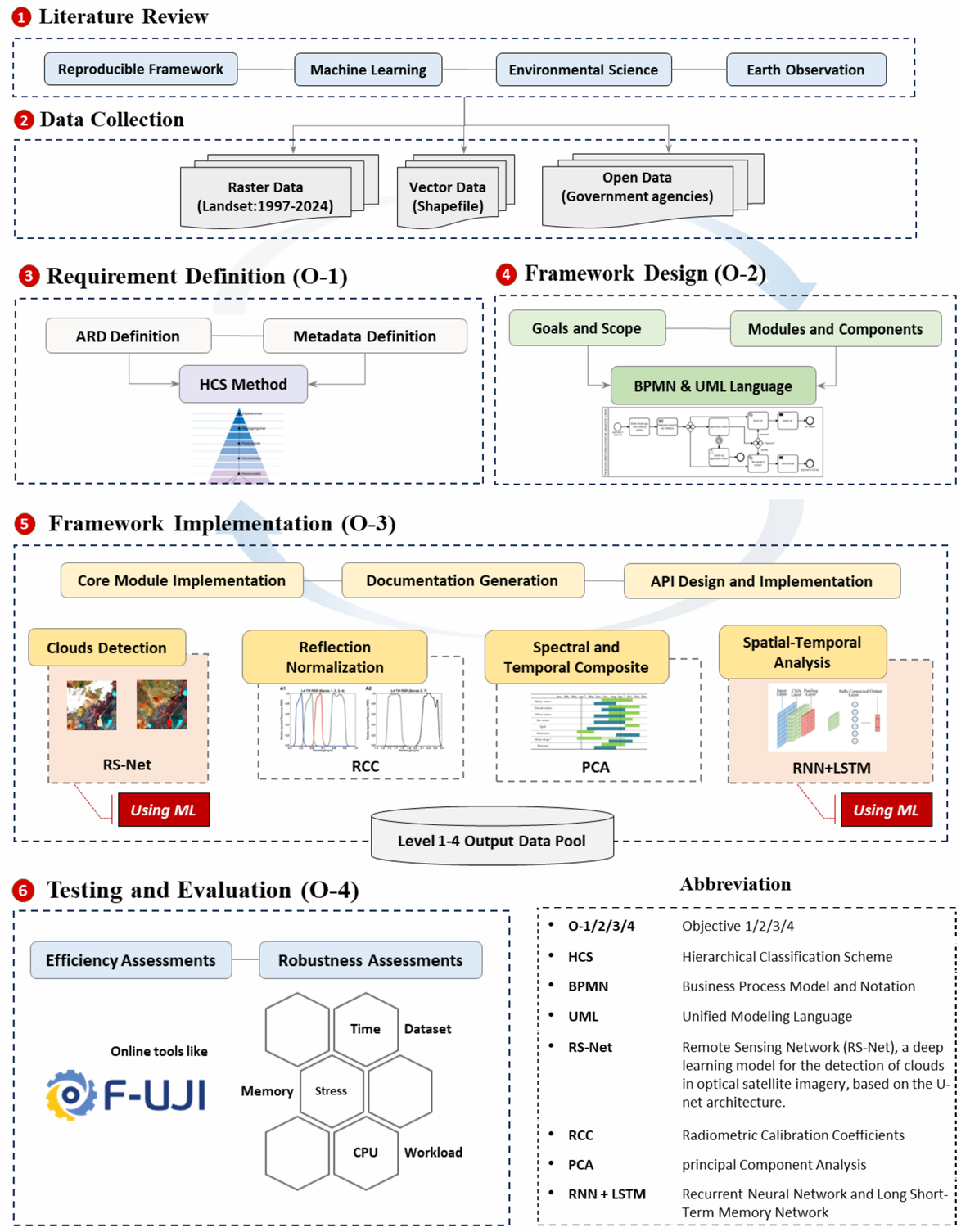

Fig. 2. The flowchart of the whole research

References

[1] Musen, M.A., O'Connor, M.J., Schultes, E. et al. (2022). *Modeling community standards for metadata as templates makes data FAIR*. Scientific Data, 9(1), 696.

[2] Devaraju, A., & Huber, R. (2021). An automated solution for measuring the progress toward FAIR research data. Patterns, 2(11), 1-15.

[3] Blanchi, C., Gebre, B., & Wittenburg, P. (2022). *Canonical workflow for machine learning tasks*. Data Intelligence, 4(2), 173-185.

[1] School of Earth and Planetary Science, Curtin University; [2] FrontierSI; [3] School of Design and the Built Environment, Curtin University;
[4] Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University

Australian Government Department of Industry, Science and Resources | Cooperative Research Centres Program