Challenges of Tracking Provenance in Marine Data

Tanja Auge[1] (tanja.auge@ur.de), Fajar J. Ekaputra[2,3] (fajar.ekaputra@wu.ac.at), Susanne Feistel[4] (susanne.feistel@io-warnemuende.de), Susanne Jürgensmann[4] (susanne.juergensmann@io-warnemuende.de), Meike Klettke[1] (meike.klettke@ur.de), Laura Waltersdorfer[2,3] (laura.waltersdorfer@tuwien.ac.at)

[1]University of Regensburg (Germany)
[2]TU Vienna (Austria)
[3]WU Vienna (Austria)
[4]Leibniz Institute for Baltic Sea Research Warnemünde (Germany)

Long-term interdisciplinary data studies in the Baltic Sea area

The *Leibniz Institute for Baltic Sea Research, Warnemünde (IOW)* is a non-university research institution dedicated to interdisciplinary marine research in coastal and marginal seas divided into four scientific sections. The IOW focuses on the Baltic Sea ecosystem and holds data from more than 130 years of research, both collected by the institute itself and provided by other research institutes.

Due to their *provenance* in terms of where they were collected, who collected them and when they were collected, the data are in different formats and have different measuring devices, methods and standards for data collection and management. Provenance generally describes the information supporting the reproducibility of research data processes. The Physical Oceanography Section, for example, deals with long-term observations of the marine environment. For decades now, the IOW has deployed a *CTD-Probe* (*Conductivity*, *Temperature* and *Depth*) to explore the water column. The device contains multiple sensors to continuously measure parameters and a varying number of water bottles to take samples from the water column at specific depths.

From Sensor to Publication

Any measurement by sensor is initially a voltage value which gets converted by the operating software *Seasoft* to a numeric value of the measured parameter. To assure quality and accuracy of the measurements several measures are taken. One is the deployment of double sensors for the same parameter. Another measure is the comparison of CTD-measurements with measurements of the same parameter by different methods. A third way to ensure the quality of CTD-measurements is the precise calibration of the sensors before and after a deployment or time period. Calibration factors are determined and have to be applied in the post-processing. The results of a CTD measurement, as well as their analyses, are initially stored locally on the scientists' computers in files of various formats and structures such as *txt*, *csv* or *xlsx*, and then in an evolving relational database (see data science pipeline in Figure 1, highlighted by green arrows). To make the data available as quickly as possible, the raw data will be regularly updated/published in real-time databases such as *Copernicus*[44]. In addition, the data processed in the *IOWDB*[45] will be made available to IOW researchers and the most important part of the data will be made available to the public. In summary, (raw) data is published at different times, in different formats and at a different level of analysis (see orange arrows in Figure 1). To make sure that IOW data can be used by others in the long term, the provision of data at

[44] Coperincus: https://www.d-copernicus.de
[45] IOWDB: https://odin2.io-warnemuende.de

IOW is subject to a *data policy*[46] for handling research data. Therefore, *metadata* (geolocation, time, sensors, calibration, operating persons, processing software, etc.) have to be collected and stored during the entire process to enable comparability in the aftermath. To ensure the collection, storage, and reuse of all data and metadata in accordance with *FAIR (Findable, Accessible, Interoperable, Reusable) principles* a box of easy-to-use, standardized tools that accompany the research process is required.

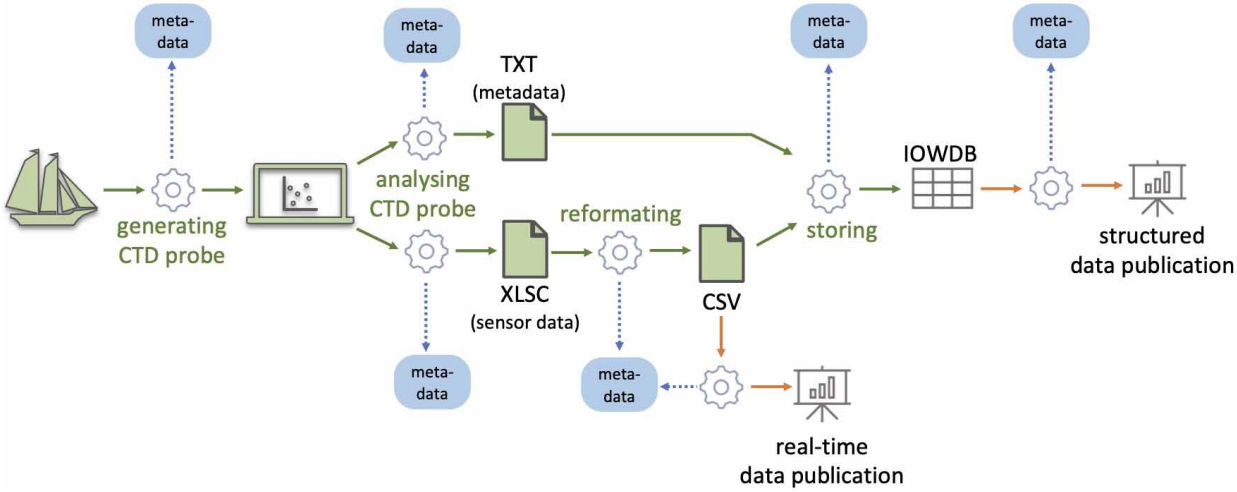

Figure 1 Example for a possible data science pipeline as it might be anchored at the IOW.

Challenges and Use Cases for a Marine Data Provenance Tool

When designing and implementing such a toolbox, several challenges must be addressed:

- As the standards for data collection differ between institutions, countries, and application scenarios, a high degree of heterogeneity within the data is to be expected.
- In long-term studies, schema changes cannot be avoided.
- A centralized collection and management of data is often expensive and time-consuming.
- The storage structures within a project sometimes do not match the desired formats of the funders or the institution-wide storage.
- Data collection is largely automated, depending on the collection method, and manual review of the data collected is simply not possible.
- While metadata is collected throughout the data science process, it is usually collected independently and not collected, managed, and stored in a structured, centered way.

We believe that these challenges are typical for the entire scientific landscape, since in many contexts data are collected in time-limited projects and made available in other formats, in the long term. Our goal is to better support researchers and data managers of institutes like the IOW in their (computational) scientific work to answer specific (provenance) questions such as:

- Q1: Which datasets are affected by an error or bug, and how?
- Q2: Who was involved in generating the data?
- Q3: Which data or scripts are needed to repeat a workflow or (re-)produce a result?

[46] IOW Data Policy:
https://www.io-warnemuende.de/files/forschung/mediathek/iowdb/IOWDataPolicy_20180411.pdf

We therefore aim to capture metadata throughout the whole data science process and preserve it using (workflow- and data-) provenance techniques. For this, we are developing concepts and tools based on existing initiatives (SeaDataNet, EMODnet, EMBRC, ...), standards (W3C PROV, ...) and ontologies in cooperation with the IOW, the University of Regensburg, TU Vienna and WU Vienna.