Optimizing Laboratory Data Processing at the ZMT and Guidelines for the General Community

Marion Rieken, Birte Hemmelskamp-Pfeiffer, Donata Monien

This work was supported by the German Research Foundation DFG under the grant agreement number 442032008 (NFDI4Biodiversity). The project is part of NFDI, the National Research Data Infrastructure Programme in Germany.

Abstract

Over time, the historical development of work processes can lead to inefficiencies in data collection and data processing. This was also the case for the laboratory data from the MAREE at the Leibniz Centre for Tropical Marine Research (ZMT). To address this issue, a pilot project was initiated as part of the Use Case 27 of the NFDI4Biodiversity. The project aimed to systematically record and improve the processing of research data from the ZMT's experimental laboratories, with a particular focus on establishing a structured collection of nutrient data from the MAREE's water tanks. This paper discusses the challenges encountered in this data processing, outlines the procedure applied to overcome them, and provides recommendations that are applicable to other data processing challenges in research institutes.

1 Background

The efficient processing and management of scientific data is critical to ensure the reliability and accuracy of research results. However, historical practices and the evolution of work processes can lead to inefficiencies, especially when dealing with large and complex datasets. At the Leibniz Centre for Tropical Marine Research (ZMT), the laboratory data collection and processing methods used to monitor water tank nutrients in the MAREE were found to be fragmented and prone to error. This situation highlighted the need for a more structured and streamlined approach to data management. As part of the Use Case 27 of the NFDI4Biodiversity, this pilot project aimed at identifying the challenges, providing recommendations and consolidating existing nutrient data in a database.

The initial challenge involved the fragmented storage of weekly nutrient measurements across multiple Excel files. Although each file contained the relevant data, the decentralised structure significantly hindered the efficient and reliable generation of time-series visualisations. Manual data aggregation via copy-and-paste methods – already used in some instances – proved to be not only labour-intensive, but also prone to transcription errors and inconsistencies, resulting in incomplete datasets. A key requirement for subsequent analysis was the reconstruction of a comprehensive chronological dataset encompassing MAREE water tanks, nutrient concentrations, and associated standard substances.

2 Procedure

In order to address the aforementioned challenges, nutrient data from the years 2023 and 2024 were consolidated into a unified PostgreSQL database. This integration facilitates comprehensive data analysis and enhances the identification of temporal correlations. The consolidation process involved the systematic import of numerous individual Excel files into the database. For this purpose, an ETL (Extract, Transform, Load) tool – specially Pentaho – was employed to extract and integrate the datasets. This approach enabled the automated and reusable merging of heterogenous data sources. Throughout the implementation, several technical and procedural challenges emerged, which are examined in detail in the following sections. Based on these insights, a set of recommendations is provided to support the prevention or more efficient resolution of such issues in future projects.

3 Common Challenges and Recommendations

The following section summarizes the challenges associated with the preparation of the dataset.

3.1 Filenames and Metadata Extraction

The filenames of Excel files containing laboratory data frequently encode essential metadata such as measurement dates, responsible personnel, and versioning information. To ensure the preservation and accessibility of this metadata, it is systematically extracted and stored alongside the laboratory data within the central database. This extraction process is implemented using the ETL tool Pentaho which interprets filenames as structured text strings according to a set of predefined parsing rules and naming conventions:

- An underline (_), used exactly once between each component, serves as the delimiter separating distinct metadata elements within the filename.
- Relevant files begin with a standardized prefix in the form “xx_NUT_MAREE_”, where “xx” denotes a two-digit numerical code indicating the measurement run or rerun sequence. This prefix is immediately followed by an underscore and the fixed identifier “NUT_MAREE_”.
- The subsequent string encodes the date if laboratory analysis in the format “YYYYMMDD”.
- If the underscore following the date is not immediately succeeded by the string “ver”, the intermediated substring (from this underscore to the one preceding “ver”) is interpreted as the project identifier.
- Finally, the filename concludes versioning information. The substring “ver” is followed by a numerical version code in the format “X.XX” (a single-digit integer, a decimal point, and two decimal places), indicating the specific version of the Excel file and the processing state of its content.

Despite the clear structure of the naming conventions, inconsistencies and minor deviations in filenames – although typically easy for humans to recognize and interpret – can significantly hinder automated metadata extraction. Table 1 illustrates this issue using three illustrative examples from the MAREE dataset. The left column lists filenames that conform to the established conventions and can be

processed correctly by the import tool. The middle column presents filenames with minor errors or deviations, which are difficult for rule-based parsing algorithms to interpret accurately. The right third column identifies and explains the specific issues in each non-conforming filename.

Table 1: Ilustrative Examples of Non-Conforming Filenames

Compliant File Names	Incorrect File Names	Identified Error
01_NUT_MAREE_20230116_Ana_Grillo_ver1.06.xlsb	01_NUT_MAREE_20230116_Ana_Grillo_ver1.6.xlsb	"..ver1.6…" : zero missing
01_NUT_MAREE_20230123_Ana_Grillo_ver1.06.xlsb	01_NUT_MAREE_20230123__Ana_Grillo_ver1.06.xlsb	".. 23__Ana …": double underline
01_NUT_MAREE_20231009_ver.1.09.xls	01_NUT_MAREE_202310009_ver.1.09.xls	"…EE_202310009_v…" not a date, because of redundant zero

Recommendation 1: Rule-based software systems, such as the ETL processes employed here, rely on strict adherence to predefined filename patterns; even slight deviations can result in failed parsing or misclassification of data.

3.2 Variable Naming

A persistent challenge in managing the dataset stems from the involvement of multiple technicians over time, each bringing unique backgrounds and recording habits. Consequently, the same variable is frequently labelled using different nomenclature or spelling conventions across entries. This lack of standardization introduces a significant barrier to data harmonization and analysis, as it requires comprehensive knowledge of all synonymous or semantically equivalent terms.
For instance, the following pairs represent cases where identical data points have been inconsistently labelled:

- *shrimp tank* vs. *shrimptank*
- *AQ03* vs. *Glasrosen*
- *EA (Experimental Aquarium)* vs. *ET (Experimental Tank)*

These discrepancies can result in erroneous data classification, where semantically identical records are mistakenly treated as separate categories.

Recommendation 2: It is essential to develop a controlled vocabulary or synonym mapping to ensure terminological consistency. Additionally, defining the acceptable values for each descriptive column and storing them as predefined lists is recommended to standardize data entry and prevent future inconsistencies.

3.3 Data Formats

In Microsoft Excel, columns are typically defined by headers that implicitly specify the expected data type for the values contained within each column. For instance, a column labelled "insert date" is expected to contain temporal data entries (i.e., dates), rather than free-text or alphanumeric strings. Despite this implicit expectation, Excel does not enforce strict data type validation, thereby allowing

users to input values that deviate from the intended format. From the perspective of the Excel file, this may be acceptable as long as data entries are mutually exclusive, i.e., they don't overlap in meaning or data type and do not cause errors in the intended processing workflows.

These structural and semantic inconsistencies present significant challenges in the context of processing the MAREE data. In the specific case examined here, quantitative laboratory measurements – typically reported in micromoles per liter (µmol/L) – are expected to populate dedicated columns within a measurement series. However, in the provided Excel files, certain columns intended to contain measured values of nutrient concentrations (specifically NOx), were instead filled with yellow-highlighted textual interpretations, rather than actual data. This blending of numerical and textual content within a single column introduces inconsistencies that obstruct automated parsing and transformation routines required for database integration. An illustrative example of this issue is provided in Table 2.

Table 2: Blending of Numerical and Textual Content Within a Single Column

NO_x[µmol/L]
24,29
779,99
10,63
0,00
0,00
too high
too high
174,83
116,55
80,13
119,86
123,52

Recommendation 3: Text interpretation is not qualified for evidence-based data storage. Every information of importance should be a discrete notation with a dedicated position within the Excel worksheet.

Recommendation 4: Each row within an Excel worksheet should represent a single, consistently defined data record, with all values conforming to predefined data types and formats across columns.

3.4 Naming and Identifiers

A recurring issue within the MAREE Excel sheets for nutrient data is the inconsistent use of names, abbreviations, and project identifiers. These inconsistencies complicate data integration and hinder traceability. The following examples illustrate the variation in how this information is currently recorded:

- Marilyn Meier
- Project 006/2023
- EA850

The absence of a unified naming convention across MAREE datasets represents a substantial barrier to effective data consolidation and integration with other data sources. In many cases, project references rely on the name of the responsible contact person, which, while practical in an operational context, is insufficient for systematic data linkage. Instead, the use of a unique, institutionally assigned project number – such as the ZMT project ID – would provide a solution. Such identifiers can be linked to associated metadata, including contact persons, project descriptions and other relevant information associated with the project.

Recommendation 5: To ensure reliable identification of datasets and consistent referencing of shared information, a standardized schema for project names, project numbers, and abbreviations should be implemented.

Recommendation 6: Several value ranges, such as those in the MAREE tanks, are not limited to individual datasets, but recur across different data sources. These should be recognized as cross-cutting contextual variables and treated accordingly. A ZMT-wide agreement should be reached on naming conventions for tanks and water types, as well as similar classification attributes.

3.5 Structure of the Excel Sheets

The Excel files intended for import into the database were originally designed by technicians as tools to support the nutrient analysis process. The initial goal was not focused on periodic database imports. As simple table structures proved insufficient for a working process, the Excel files became increasingly complex, containing multiple sheets and numerous formulas. However, this complexity has made them less suitable for straightforward database transfers.

Over time, the requirements for the Excel file as a working tool have evolved, leading to changes in their content. In particular, the data structure has become unstable and the position of the data within the sheets has shifted over time. This problem has been intensified by the frequent insertion of comment lines, which makes the consistency of the data structure more difficult. This means that the import program has to be adapted regularly, even if there are no fundamental changes to the data structure.

Recommendation 7: For reliable and periodical database imports, a simple and stable data structure in the files should be established – one that contains mainly fields and data, with minimal formulas or references. The Excel sheets for static data should be separated from those supporting the working process.

Recommendation 8: Each column should have a clearly defined meaning, and once column name and meaning are established, they should never change. However, columns may remain empty if they are no longer needed.

Recommendation 9: To maintain a stable data structure, new information or columns should only be

appended at the end of the sheet. Ad hoc information does not belong in a standardized Excel file (e.g., avoid writing notes between the data to be imported). If necessary, a defined area for notes should be created.

3.6 Number, Date and Time Conventions

Differences in Excel settings between the English and German versions are also relevant to the data import process. Variations in data formats and numerical notations can lead to misinterpretations if not recognized. The issue is made even more complex, because of multiple photometers and connected PCs which are configured differently. Depending on which PC is connected to a particular photometer, a numerical value such as '1.000' may be interpreted either as “one thousand” or “one”, leading to potential inconsistencies. Similarly, discrepancies in date formats between language or regional settings of various PCs and associated devices add another layer of complexity.

Recommendation 10: To avoid mistakes and misinterpretation of data, it is recommended to standardize the regional settings of all computers, PCs, and laptops at ZMT to English, as the most widely used language.

4 Data Model

4.1 Motivation of the Model

Nutrient measurement data represent only a small fraction of the scientific and other datasets collected at ZMT. At the same time, they are directly or indirectly connected to a wide range of additional information:

- Projects: Some measurements were commissioned within the scope of specific projects and can therefore be linked to them. This allows connections between different laboratory processes to be established via project assignment – provided that such links have been methodically prepared.

- People: Projects involve researchers for whom the nutrient data are relevant. Consequently, information such as project duration, research objectives, or funding can also be associated with the measurement data.

- Tanks: Samples for nutrient analyses originate from tanks, which may be categorized as permanent or experimental. These properties are included in the data model and can be extended with further details if needed – for example, the plants or animals inhabiting a tank and their native regions.

Since all information at ZMT is interrelated and interdependent, a data model has been documented and implemented within this pilot project. It describes the relevant entities and their relationships, ensuring that nutrient data can later be correlated with other ZMT datasets. This, in turn, enables integrated queries and reports across different domains of knowledge.

4.2 Design of the Model

To ensure a reliable, consistent, and efficient data import process, the data model has been designed as an Entity-Relationship Model (ERM) (see Figure 1). This modeling approach is widely used for semantic data modeling, representing objects, their attributes, and the relationships between them.

The objectives of the ERM are:

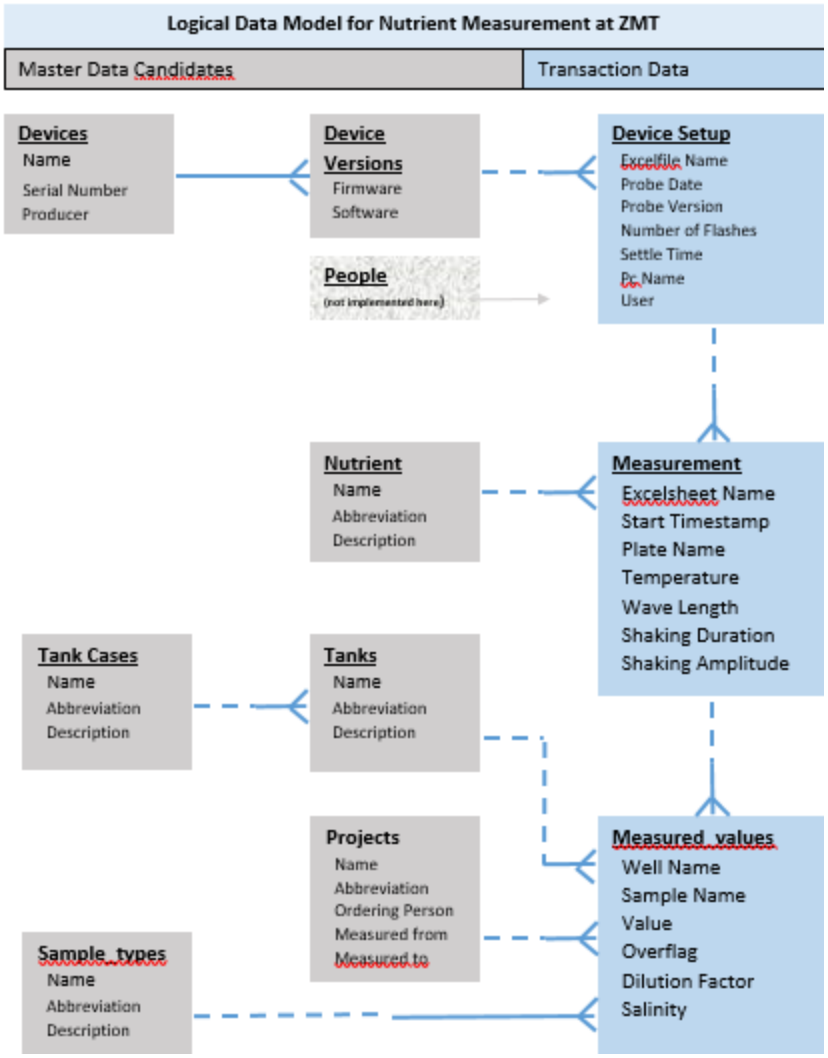
- to avoid redundancy and ensure consistency in the typing of objects and attributes,
- to provide a clear graphical representation of entities and their relationships,
- and to allow for flexible extensions that capture additional logic and business rules specific to ZMT.

A central aspect of the ERM is the distinction between transaction data and master data, which together form the structural backbone of the model:

- Transaction Data: Measurement results are classified as transaction data. Each measurement represents an independent and self-contained event. While older data lose direct relevance once new measurements are taken, they remain important for historical reference and time-series analysis.
- Master Data: Core objects such as tanks, nutrients, projects, and operational devices are modeled as master data. These are stable over time, uniquely identifiable, and consistently referenced across multiple measurement events. Standardized master data ensure semantic consistency, contextual accuracy, and uniform terminology throughout the database.

This separation between transaction and master data provides a solid foundation for consistent imports, reliable queries, and the long-term integration of datasets. To achieve full interoperability, the identified master data candidates should be incorporated into ZMT's central master data system, which is currently under development. This integration will allow nutrient data to be harmonized with other institutional datasets, enabling cross-domain analyses and comprehensive reporting.

Figure 1: Data Model for Nutrient Measurement at ZMT



Recommendation 11: A data model should clearly distinguish between master data that is relatively static, and transaction data that are dynamic.

5 Establishment of a Stable Import Process

In light of the identified challenges and building upon the outlined recommendations, the structure of the Excel spreadsheets used for laboratory data imports has been significantly simplified. Raw measurement data, calculations and staff-generated reports are now stored in separate files, marking a major step toward automating the import process. The only remaining manual input is the mapping between the position in the measuring sequence and the corresponding samples. This mapping must be entered once into the machine-exported file. Furthermore, the worksheet names must be manually adjusted to match the expected format. Despite these minimal manual steps, the import process is now much more robust.

During the ETL process, the following implicit, but contextually valuable information (metadata) is derived from the original data (also see Figure 1):

Sample Type: The Sample Type indicates the origin or nature of the analysed sample – for example, whether it represents a reference material, water from a MAREE tank, or water from an experimental tank. This information can be systematically inferred from the sample name and from the attributes of the tank the sample was taken from.

1) A sample name that matches the name of a tank serves as a reference to that specific tank. Based on the tank's attributes, the sample type can be categorized as either a MAREE tank or an experimental tank.

2) Some identifiers correspond to "calibration samples" or "sample matrix reference materials" that are used for the external calibration of analytical methods or quality assurance.

Project Assignment: The association between experimental tanks and specific research projects can be derived through defined programmatic procedures, for example: (1) Project Assigment based on sample name; (2) Project Assignment based on time interval.

6 Temporal Data Visualisation

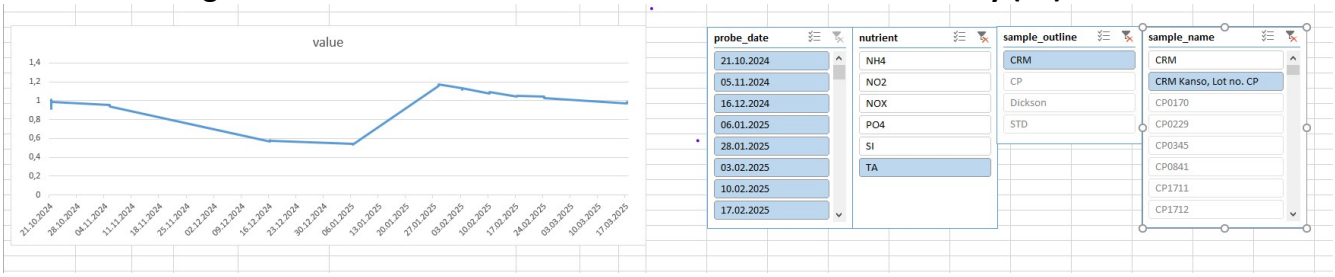
Once the laboratory data are stored in a structured database, they are ready for a wide range of analytical purposes. This includes the generation of graphical representations, in which the previously inferred metadata (e.g. sample type, project affiliation) can be visualised within temporal sequences.

Examples

1) Standard Substance

Figure 2 is a visualisation of measured Total Alkalinity (TA) values in the **C**ertified **R**eference **M**aterial (CRM Kanso, Lot no. CP) recorded over a selected period of five months. In the model, by using the check boxes the nutrient, the standard substance and the period can be changed. Such visualization makes it easier to monitor the behavior of reference materials, e.g., the degradation process and the need to use a new bottle.

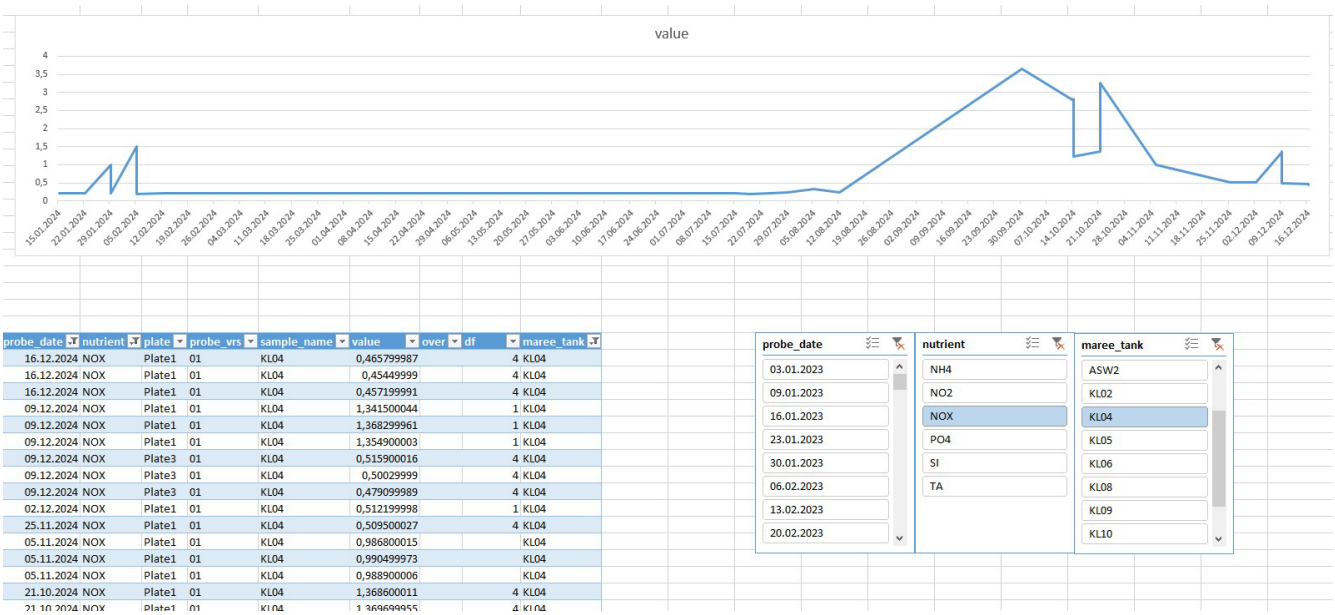
Figure 2: Time series visualisation of measured Total Alkalinity (TA) values



2) MAREE water tanks

Figure 3 illustrates the measured concentrations for nitrogen oxide "NO_X" in a MAREE water tank designated *KL4* over the year 2024. Additionally, a table with the exact values and the dilution factors is shown. By using the check boxes the nutrient, the MAREE water tank and the period can be changed in the generate other visualisations over time. These easily modifiable time periods, tanks and parameters offer a quick overview of the tanks, enabling measures to be taken to improve water quality or monitor long-term developments.

Figure 3: Time series visualisation of NO_X Concentrations in then MAREE Tank KL4 (2024)

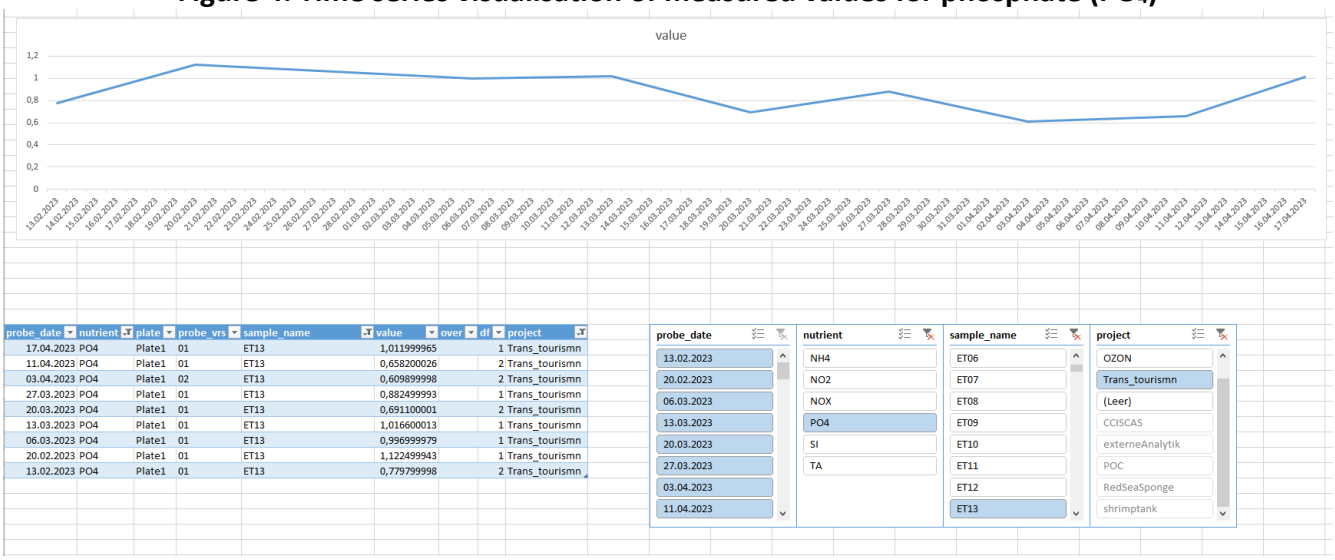


probe_date	nutrient	plate	probe_vrs	sample_name	value	over	df	maree_tank
16.12.2024	NOX	Plate1	01	KL04	0,465799987		4	KL04
16.12.2024	NOX	Plate1	01	KL04	0,45449999		4	KL04
16.12.2024	NOX	Plate1	01	KL04	0,457199991		4	KL04
09.12.2024	NOX	Plate1	01	KL04	1,341500044		1	KL04
09.12.2024	NOX	Plate1	01	KL04	1,368299961		1	KL04
09.12.2024	NOX	Plate1	01	KL04	1,354900003		1	KL04
09.12.2024	NOX	Plate3	01	KL04	0,515900016		4	KL04
09.12.2024	NOX	Plate3	01	KL04	0,50029999		4	KL04
09.12.2024	NOX	Plate3	01	KL04	0,479099989		4	KL04
02.12.2024	NOX	Plate1	01	KL04	0,512199998		1	KL04
25.11.2024	NOX	Plate1	01	KL04	0,509500027		4	KL04
05.11.2024	NOX	Plate1	01	KL04	0,986800015			KL04
05.11.2024	NOX	Plate1	01	KL04	0,990499973			KL04
05.11.2024	NOX	Plate1	01	KL04	0,988900006			KL04
21.10.2024	NOX	Plate1	01	KL04	1,368600011		4	KL04
21.10.2024	NOX	Plate1	01	KL04	1,369699955		4	KL04

3) Project

Figure 4 is a visualization of measured phosphate (PO_4) concentrations in the experimental tank *ET13* over the duration of the ZMT project *Trans_Tourismn*. In addition, a corresponding table presents the exact measured values alongside the applied dilution factors. By using the check boxes the nutrient, the experimental tank, the period and the project can be changed. MAREE tanks can easily be filtered by project, as project and storage tank data are measured together. There is no need to copy and paste or create multiple files.

Figure 4: Time series visualisation of measured values for phosphate (PO_4)



probe_date	nutrient	plate	probe_vrs	sample_name	value	over	df	project
17.04.2023	PO4	Plate1	01	ET13	1,011999965		1	Trans_tourismn
11.04.2023	PO4	Plate1	01	ET13	0,658200026		2	Trans_tourismn
03.04.2023	PO4	Plate1	02	ET13	0,609899998		2	Trans_tourismn
27.03.2023	PO4	Plate1	01	ET13	0,882499993		1	Trans_tourismn
20.03.2023	PO4	Plate1	01	ET13	0,691100001		2	Trans_tourismn
13.03.2023	PO4	Plate1	01	ET13	1,016600013		1	Trans_tourismn
06.03.2023	PO4	Plate1	01	ET13	0,996999979		1	Trans_tourismn
20.02.2023	PO4	Plate1	01	ET13	1,122499943		1	Trans_tourismn
13.02.2023	PO4	Plate1	01	ET13	0,779799998		2	Trans_tourismn

7 Conclusion

The outcomes of this pilot project highlight the critical importance of implementing structured and standardized data management practices in research settings, particularly when dealing with large and complex datasets. The challenges encountered during the processing of nutrient data from the MAREE at ZMT reflect common issues faced by many research institutions, where legacy workflows and evolving procedures often result in inefficiencies and data inconsistencies.

The integration of distributed data segments into a centralized Master Data System represents a key step toward the creation of a unified and interoperable information infrastructure. Such a system offers substantial benefits not only for the management of nutrient data but also for broader research data contexts. Cross-functional master data will serve as the link between databases and operational systems at ZMT.

By consolidating and optimising data collection and processing workflows, this project not only addressed specific technical or workflow issues, but also established a foundation for long-term improvements in lab data management. Integrating the lessons learned from this project into the NFDI4Biodiversity framework will further strengthen the infrastructure for data management, promoting more effective collaboration and facilitating data sharing within the scientific community.