Technical Report

Quality assurance in the MARLIN system

Authors: Paul Kloss, Hendrik Pehlke, Dr. Jan Beermann, Dr. Alexa Wrede, Dr. Jennifer Dannheim

(Technical report of the project "ANsätze zur Kostenreduzierung bei der ERhebung von Monitoringdaten für Offshore Vorhaben" (ANKER), Arbeitspaket BENTHOS; FKZ 0325921)

Last update: 03. June 2020

Citation: Kloss P, Pehlke H, Beermann J, Wrede A, Dannheim J (2020). Quality assurance in the MARLIN system. Technical Report of the project ANKER FKZ 0325921, pp. 19.

Alfred-Wegener-Institut, Helmholtz-Zentrum
für Polar- und Meeresforschung (AWI)
Am Handelshafen 12
27570 Bremerhaven

1. Aim

The goal is to establish biological and technical criteria for plausibility checking of benthos and fish data in MARLIN, always focusing on the automation of these actions. The necessary mechanisms and algorithms were developed and implemented transparently in MARLIN in close cooperation with SCOPELAND Technology. Furthermore, the data quality in MARLIN and the implementation of the criteria and algorithms were evaluated. Attention has always been paid to the compatibility of the systems MARLIN and CRITTERBASE, particularly with regard to future high-quality data exchange. In particular, the following report concerns benthic data (ger. „Schutzgut Benthos“).

2. Introduction

The intense use of our coastal systems has led to an increase of human pressures, such as transport, sand extraction and dumping, laying of cable, plumbing and pipelines, eutrophication, pollution, fisheries and, more recently, the use for marine renewable energy by the area-intense offshore windfarms. Thereby, every kind of human utilisation is an interference with the respective ecosystem. Nowadays, marine ecosystem management and environmental protection aim at a sustainable use of our coastal areas and such management approaches are therefore implemented in national guidelines such as the marine strategy framework directive (MSFD), marine spatial planning (MSP) and regulations related to offshore wind farm construction and operation.

Sustainable management plans can be related to local impacts, such as for offshore wind farm environmental assessments, or on larger scales such as MSFD, spatial planning or conservation issues of habitats. Consequently, data information systems must be able to provide sound scientific databases that enables the provision of advisory information that supports regulation and assessments for decision makers at different scales. Thus, the tasks for applied science have changed fundamentally and call for a rethinking in the scientific application of data in the way of large database infrastructures in order to give scientific advice. Large data information systems supports support our understanding on how the benthos responds to pressures, how we determine the level of change and ultimately, whether there are any effects (Halpern et al. 2008, Ban et al. 2010).

However, individual studies from the past are often restricted in the amount of data they can generate, but by combining the results from many studies, massive databases can be created that make analysis on a much-enhanced scale possible (Grassle 2000, Vanden Berghe et al. 2009). Combining datasets in data information system requires detailed thinking on the

structure of the database, the way data should be organised and which data are mandatorily required.

One of the main difficulties in integrating and comparing different datasets from various data providers is the harmonisation of the data itself, but also the metadata. Metadata are of high importance in order to describe data in a standardised way and to create a searchable metadata inventory which facilitate data exploration and the provision of information on the benthos via user-defined products. On a basic level, metadata must be able to provide information on where and when the data were collected, how data were collected and in which format data are available (Vanden Berghe et al. 2009). The harmonisation of data thus applies at different levels which are mainly the (a) taxonomic, (b) geographical and temporal, as well as the level of (c) sampling methodology (see also Vanden Berghe et al. 2009). Thereby, Harmonisation and quality assurance of data includes different levels such as the plausibility of data (e.g. spatial and temporal coverage), the format of the data (e.g. geographical position as latitude and longitude in decimal degrees, units of attributes such as biomass) and the same attribute categories (e.g. the names of the sampling gear). Central is also the taxonomy in biological data which requires plausibility checks and quality-assurance of species occurrences and unique names (Kloss et al. 2020).

Over the last decade, AWI and BSH have been working together in several projects (StUKplus[1]: see Dannheim et al. 2013, BSH-AWI project[2]: see Dannheim et al. 2016) to harmonise a plausibility and quality checked information system. Existing data from environmental impact studies on current wind farm projects and other monitoring and research projects have been harmonised over the last years. During the ANKER project, the collected data were imported into a comprehensive uniform and quality-controlled data information system. The data model was adjusted during the project and data were fitted to data information system needs in an iterative manner during the whole project. Finally, this enables us now to manage metadata, environmental data and biological data in a searchable and sustainable way in the MARLIN system in order to provide information for the benthos and demersal fish for different stakeholders.

[1] StUKplus project "Joint evaluation of data on benthos and fish for ecological effect monitoring at the offshore test field "alpha ventus"

[2] BSH-AWI project "Evaluation approaches for regional planning and approval procedures with regard to the benthic system and habitat structures"

3. Implementation in MARLIN

During previous projects, data were already harmonized, as well as plausibility and quality checked. However, implementation of the MARLIN information system and data transfer into the new MARLIN system required further in-depth harmonization, quality control and tests on plausibility of data during the ANKER project. Based on the previous experience of data handling, AWI and BSH developed further tools and implemented strategies where and how to store and handle data and metadata. The technical implementation of the data model was carried out by SCOPELAND based on advice from AWI and BSH biologists and computer scientists.

3.1. The data model

Since the quality of the data is very much dependent on the structure of the data model, it plays a central role. In the following section the structure of the data model is described and evaluated. During the project ANKER, data format and data plausibility checks were carried out continuously by AWI, in close collaboration with BSH, which consequently led to a harmonized and high-quality data stock in the MARLIN system.

Clearly identifiable entries/records

In order to clearly assign measured values to the corresponding metadata (e.g. measurement location and measurement method), these must have a distinct identifier. This guarantees that the individual data records are uniquely identifiable.

Data identification is achieved by a clean definition of the metadata framework (main tables):

- SURVEY (unique identifier of a cruise)
- STATION (unique identifier for the station approached by the ship)
- HAUL (unique name for the catch)

The actually measured data is stored here in this data package:

- SAMPLE (unique name for samples taken)

During the project, unique identifiers were established and data sets were checked and assigned to the correct identifiers. This enables that data are searchable in the future by the metadata.

Information on sampling

Information on sampling methods on board and the sampling procedures in the lab is available in the MARLIN system. These metadata thus provide information on the way

samples were taken and the quality and comparability of the samples. The following information on the entire sampling procedure is stored:

- Information on the sampled taxa: captures information on sample treatment of the evaluating laboratory (METHODS_SAMPLE).
- This concerns the method of sampling, which stores data from sampling on the ship (METHODS_SAMPLING).

Gear catalogue

MARLIN comprises a general catalog in which various information is stored. Subcategories are separated by the attribute “catalog key”. Each sub-catalogue is uniquely defined by the triple Catalogue Key, Excel Entry and Database Entry. The gear catalog is one of these sub-catalogs with the key GER. In addition to the bilingual defined gears name, the gear catalog contains the following three fundamentally important functionalities:

1. A plausibility check assignment
2. A translation rule from name in Excel spreadsheet, Catalog name to name in database
3. A control functionality for visibility in the individual objects of protection (“Schutzgut”)

In this case, the plausibility check function indicates whether the corresponding device is visible or usable in the plausibility check module. The definition of visibility in the individual objects of protection means that certain entries in certain objects of protection (“Schutzgut”) cannot be seen or used.

Figure 1 shows some gear catalog entries of available sampling methods.

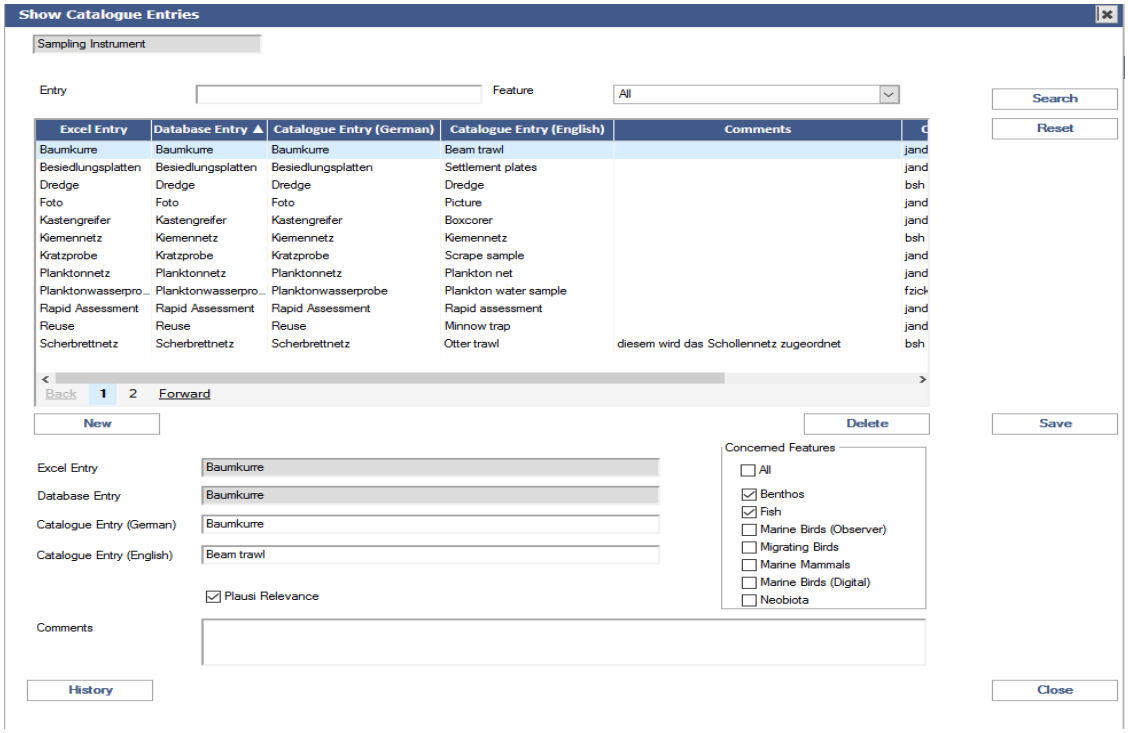

Figure 1: Gear data in general catalog of the MARLIN system

Responsible laboratory

The laboratories responsible for processing the samples are also recorded in a separate table (QUALITY_ASSESSMENT). Both, the evaluating persons and their level of qualification can be noted. This enables the quantification of the samples quality.

Commentaries

Furthermore, comments, e.g. taxonomic literature used for species identification, is stored together with the benthic data (COMMENTS). In addition to this, each of the main tables mentioned above has a dedicated comment entry option.

The entire structure of the data model provides a clean and consistent data storage, beginning with the recording of the metadata, the actual measurement data, the storage of the methods (on board and in the lab) and the information of the processing laboratory. This framework ensures that the data is stored at the highest quality level possible.

3.1.1. Iterative construction process of the data model

The actual structure of the MARLIN data model and the development of the plausibility tests were the result of years of close cooperation between BSH, AWI and the development company SCOPELAND Technology. The data model was created and refined over a number of iterations, starting with the preparation of research data. Research data was cleaned over years and led to an increasing understanding of the MARLIN data model that was developed

within the project. In the following, a description of the development work AWI has done in recent years is given. This work has directly and indirectly influenced the development of the current data model of MARLIN.

Data cleaning of the research data

All underlying data sets were cleaned, harmonised and brought into a consistent state. For example, the corresponding hauls were clearly assigned to their cruises. All data changes were logged into a text file, so that backwards tracking to the original data of the data deliverer is always guaranteed. During this data cleansing process the basics for the MARLIN data model were developed.

3.1.1. Data cleaning and harmonisation

The following is an extract of the information provided by AWI, which led to the development and improvement of the data model in MARLIN. Much of this information was implemented in the plausibility check module of MARLIN later on. Special attention was paid to the consistency of the data fields, which are important for a later selection of data in the frontend.

3.1.1.1. Improving the structure of Excel import workbooks

The original import files that BSH provides for the data provider of environmental impact assessments was kept. However, for the import of these data in the MARLIN system, several adoptions to the data model structure have been made for improvement.

Data sheet: Zoobenthos measured values (Zoobenthos-Messwerte)

- Originally, zoobenthos data were stored by linking them to the station. However, storing data linked to each haul data, i.e. the counted and weighed taxa per grab, keeps the information of each haul (e.g. geo-data as latitude and longitude). Thus, for the MARLIN system import data were adjusted and linked to haul entries.
- The actual sampled area for each haul has to be stored. In the original data of the import sheet, usually the total sampled area is given for each station. As the structure was changed in the MARLIN system to haul-linked zoobenthos data, the sampled area for each haul is now given in the MARLIN system together with the biological data.
- The specification of taxa (e.g. cf., sp., indet. etc.) has to be stored with each taxon, but in a separate column to guarantee that scientific taxon names are searchable by taxonomic databases (e.g. WoRMS).

Data sheet: Sediment measured values (Sediment-Messwerte)

- Sediment data were originally linked to the station metadata. However, sediment samples are taken from hauls (e.g. single grabs). Thus sediment measured values were linked to hauls in the MARLIN system.

Data sheet: Species reference list (Spezies-Liste)

- A complete species taxonomic reference was developed over the last years for macrozoobenthic and fish taxa. Within the project, specific attributes were added to the reference list (e.g. information on endangered species, charakterising species and neobiota). Further, the species reference list was linked to a web-based taxonomic database, here WoRMS (www.marinespecies.org) which will be used in the future.

3.1.1.2. General cleaning procedures

Migration of the historical data into the MARLIN system required some general cleaning and harmonization of data. In the following, general different classifications of data cleaning work is given.

Purely textual harmonisation (non-lookup data)

Some data was harmonized purely textually (e.g. the standardization of different spellings) without the effort of creating a dedicated lookup table. These were:

- haul: condition
- haul: weather
- haul: rain
- haul: clouds
- haul: beaufort
- haul: wind direction
- sample: specification
- sample: development stage

Lookup table for various data types (free text instead of code)

Some data values were given inconsistently. For these cases, lookup tables were created to guarantee consistent data values and to enable a queryable database. In particular, the following data fields were standardised:

- survey: purpose of a project

- survey: station area type
- haul: sampling instruments

Time and date format

All dates and times were converted to ISO standard 8601 (DD.MM.YYYY and hh:mm). This concerned in particular the following data:

- station: start/end date
- haul: start/end time

Harmonisation of data types

Some attributes of the raw data were stored as different data types (e.g. integer and text). These were unified into a well defined data type such as the information on wind force (haul: beaufort).

List definitions

Data fields designed to hold a single data unit (e.g. used instrument technique) have been misused as a list definition. Individual data was entered into such a field separated by commas. These data have been cleaned and harmonised.

Check for valid value ranges

In particular, the valid value ranges were checked. Values ranges were consecutively defined in the data model and plausibility module of MARLIN. In particular, the following types of data are checked on their values range and implemented in the plausibility procedures of MARLIN:

- survey: transect sums
- station: number of parallel hauls
- haul: sample weight with more than one decimal digit are not allowed
- haul: water temperature
- haul: beaufort
- haul: wind velocity

Ranges in scalar change

In the original raw data, ranges of values were entered where data should have been converted into a scalar value which was harmonised for migration of historical data into the MARLIN system:

- haul: wind velocity
- haul: sea state (PETERSEN)
- haul: wave height
- haul: wind velocity

Testing for sufficient value range of a scalar

For floating point values, lengths of the digits before and after the decimal point are given. Sometimes these are not sufficient to cover the specified value ranges and had to be adjusted:

- sample: drywprobe1
- sample: wetwprobe1

Test on consistent units

Some data had differing units to the MARLIN system and had to be recalculated, e.g. [g] conversion to [mg]. This was the case for:

- sample: wetprobe1

Check for missing data

All data was checked whether and where obligatory data were missing or were leading to inconsistent data states. The following data fields were affected by this procedure. If the absence would lead to an inconsistent state, the data was added, if possible, before migration of historical data into the MARLIN system:

- survey: positioning system
- survey: precision of geo references
- survey: reference system
- station: area type
- station: type
- station: distance to coast
- station: end date/time
- haul: method index of used instrument
- haul: taxa sum
- haul: water depth mean

- haul: turbidity
- sample: specimenhaul1
- sample: drywprobe1

Application of the institutionalized nothing (NULL)

In many cases, the use of the dedicated value NULL is allowed and, if necessary, mandatory to indicate missing data. Example for the variable type integer: the value 0 would not define the absence of a value but the value 0 itself. In the original data, many of these cases existed in various attributes and had to be resolved:

- haul: water temperature
- haul: water oxygenation bottom
- haul: water salinity bottom
- haul: wind direction
- haul: wind velocity

Splitting up attributes containing multiple data types

There are attributes that encode several data types and contents. For example, data on numbers of taxa (count data) requires number format. However, if taxa were only registered by presence/absence or only parts of individuals were found (e.g. tail of a polychaete), there were indicated by codes (x = colony (presence/absence), t = part of animal). These data types were split into individual attributes:

- sample: specimenhaul (T= part, X = colony)

Geo references

The following tests on geo reference data were conducted:

- station: for grab samples, the nominal coordinates should be in table station and the actual coordinates in table haul.
- haul: geo references had more than 6 decimal digits (pseudo precision) and were adjusted

General conservation of data in commentary area

In order to stay on the safe side, some data were preserved in comment fields. This ensures that the information is not lost and can be reintegrated into the data model, if needed, at a later stage of MARLIN again.

- sample: specification

3.1.1.3. Specific cleaning procedures

For the integration of the original data into the MARLIN data model, some more specific checks to special data types or attributes were conducted.

Sampled area

The sampled area is registered in the table sample. Adjustment was not necessary, because the calculation is linked to information on sampling instrument (gear specification such as for grabs, width of trawled gears) and distances from geo references in haul.

Sample depth of instruments

Important for quality check of the sample is the penetration depth of the grab instrument. Penetration depth cannot exceed the size of the used instrument (e.g. grab). If the grab is too full or if the sample is too small, samples have to be discarded.

Wind force in Beaufort

The wind force is relevant for the sampling of benthos as a quality characteristic. For example, bad weather conditions may lead to jumping of trawled fishery gears and thus the size of the real sampling area might be affected. However, wind force influence on sampling depends also on the size of the ship. The wind force data are usually not used for selection, but provide important metadata information (in combination with ship size) for quality checks of sampling.

Laboratories

In exceptional cases, an expedition can be carried out by two laboratories. The MARLIN data model only allows one entry, i.e. the main laboratory. The data were checked and cleaned to ensure persistent identification of laboratories.

Bathymetry data used for water depth mean in hauls

Water depth is an important environmental parameter for some benthic analysis and thus products. However, at one single station (same geographical location) we registered differences up to six meter due to tides and swell. Hence, data on water depth from the original data were of minor use as they have not been tide or swell corrected. Thus bathymetry data from BSH were used to fill data gaps and to analyse biological data for products.

Omission of pre-calculated data

In the historical data of the benthos database, some attributes contained calculated values based on raw data of other attributes. In older versions of the benthic MS Access tables these data kept to be stored, as they contain the history of dataset revisions. In the current MARLIN system, these columns are still kept for documentary reasons. However, for analysis these values are always recalculated based on the actual data.

(and thus the calculated values) were omitted, as data were often calculated incorrectly and calculation can be easily carried out based on the attributes in the current MARLIN system. Omitted columns concerned:

- sample: density
- sample: sum
- sample: wetweight
- sample: wetsum
- sample: dryweight
- sample: drysum
- sample: AFDW
- sample: AFDWSUM
- sample: HolPraesenz

4. Technical descriptions

4.1. The data model

Meta data tables

The uniqueness of the metadata in the MARLIN system is achieved by storing the respective identifiers unique in the following tables.

- T_SURVEY.SURVEY_ID
- T_STATION.STATION_ID
- T_HAUL.HAUL_ID

As the respective identifiers are already globally unique, the combinations of the three identifiers to which a sample is attached to, are also unique.

Measurement data tables

The sample name is also unique and is attached to a haul. Thus, the combination of survey, station, haul and sample is globally unique and clearly identifies a measured value.

- T_SAMPLE.SAMPLE_ID

Further metadata information

The catalog tables contain master data that can be linked to other records of tables. These tables contain important information on the quality of data such as sample procedures in the lab and sampling procedures on board of a ship, as well as on the expertise of the lab and the taxonomic benthologist. Further, metadata enables searchability of data. Metadata information is stored in the following tables.

- T_METHODS_SAMPLE (description of the treatment of samples)
- T_METHODS_SAMPLING (description of the sampling, e.g. instruments)
- T_QUALITY_ASESSMENT (information on the processing laboratory)
- T_COMMENTS (further comments on a sample)
- CATALOGS (identifiers of the instruments available for use)

Comparison of the data models of MARLIN and CRITTERBASE

In the following, a brief comparison of the corresponding data locations in MARLIN and CRITTERBASE is given. This is particularly important for the interoperability of the two different information systems.

Table 1: Comparison of data tables and corresponding stored information in the MARLIN and CRITTERBASE information system.

MARLIN	**CRITTERBASE** *corresponds directly to ...*	*data is distributed in ...*
T_SURVEY	*cruise*	
T_STATION	*station*	
T_HAUL	*sample*	
	subset	
T_SAMPLE	*biota*	

T_METHODS_SAMPLE		*sample, gear*
T_METHODS_SAMPLING		*sample, dataset, gear*
T_QUALITY_ASESSMENT		*dataset*
T_COMMENTS		*cruise, station, sample, biota*
CATALOG (GER)		*gear*

Figure 2 shows the data model of CRITTERBASE. Including the corresponding tables of table 1.

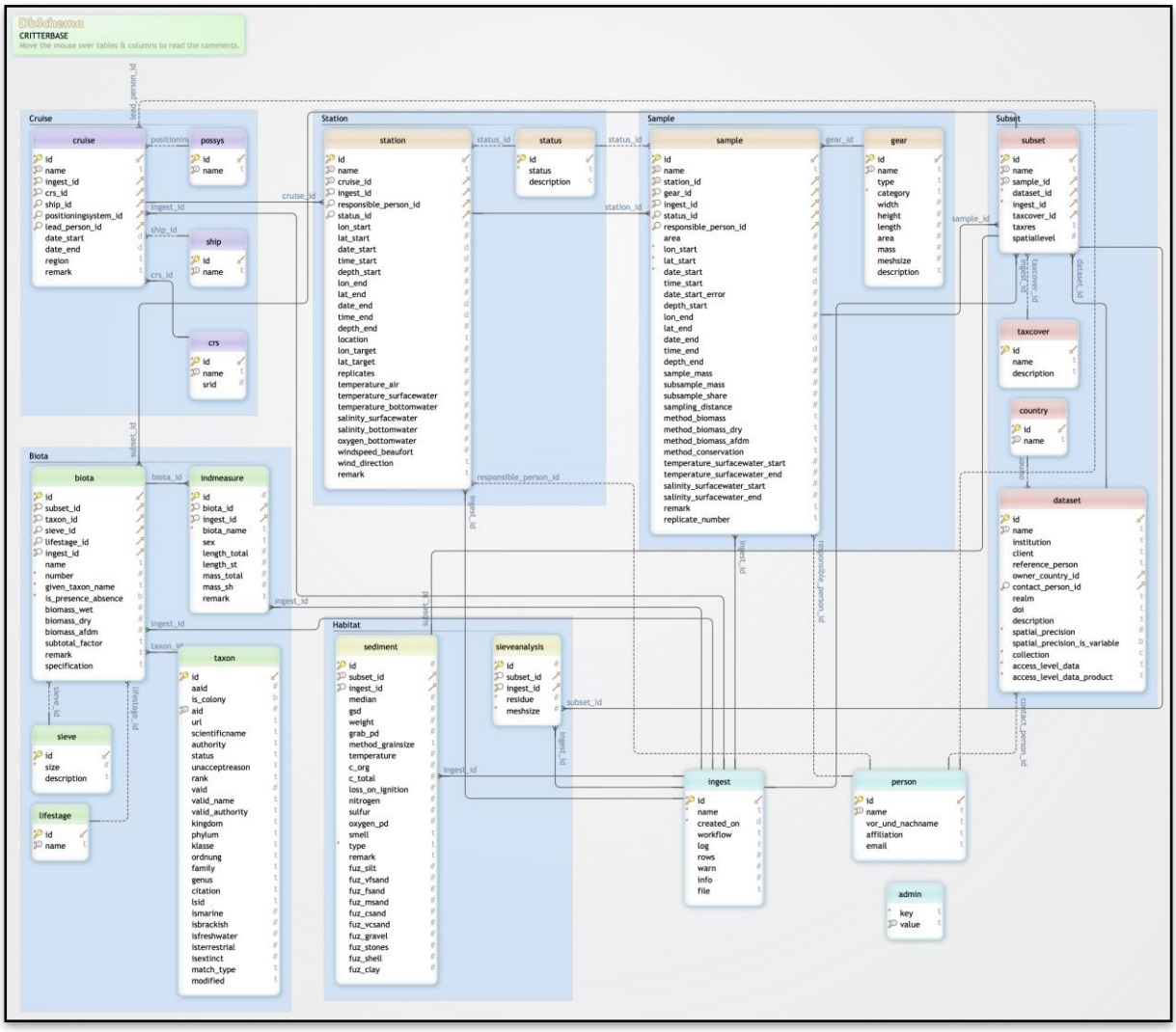

Figure 3: Data model of CRITTERBASE

4.2. Plausibility checks in MARLIN

Based on the results presented in chapter 3.1.1. and on the experience of data harmonisation and controlling, various plausibility tests have been implemented in MARLIN. The tests were generically defined, as far as possible, in the administration area of MARLIN. Deeper functionalities can be extended via SQL or using a programming language. However, this requires an increased amount of development work. Figure 4 shows an input mask to define a plausibility test.

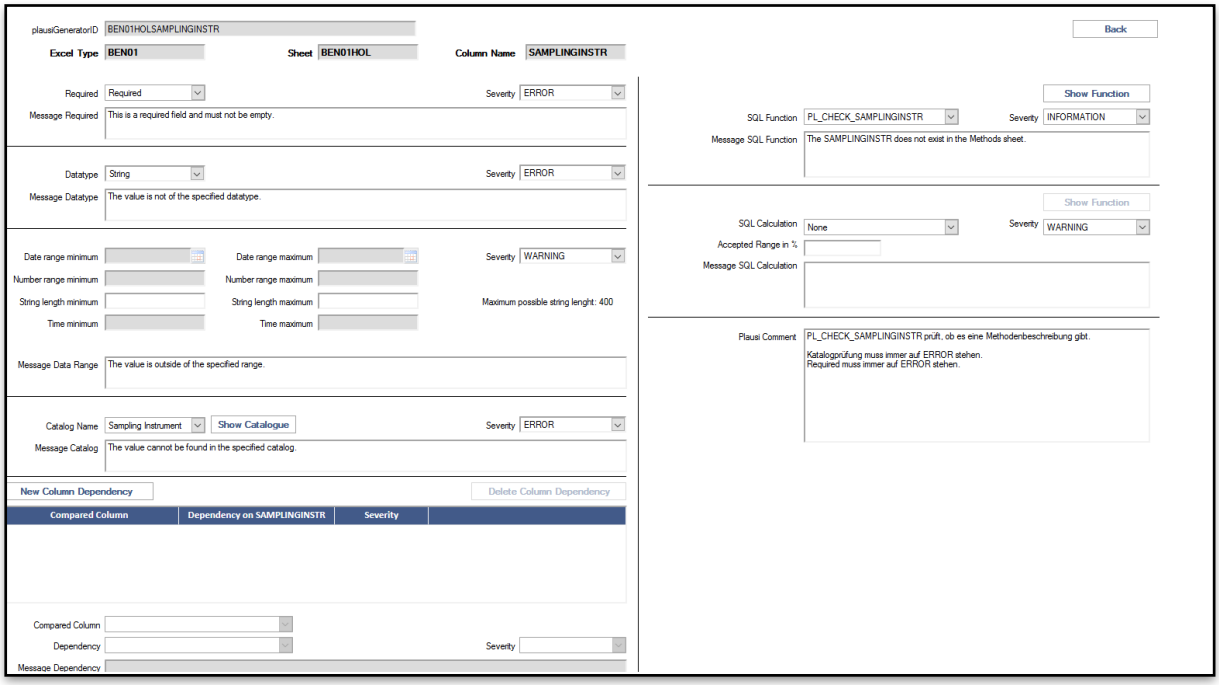

Figure 4: Input mask of the definition of a plausibility test

5. Conclusion

Consequent harmonization, plausibility and quality checks of data enabled us to integrate historical data at a high-quality level into the MARLIN information system. Further, cleaning of the data in iterative steps and the gained experience enabled us to establish routines on plausibility and quality of data for the MARLIN system which led to automated tools checking the data. Further, during the ANKER project the data model of MARLIN, as well as of CRITTERBASE, were consistently adjusted in an iterative manner according to the needs of data handling and to the needs of the data information system.

However, this is still work in progress with a steep learning curve and future potential tools we see for the MARLIN system. For example, a quality indicator of the imported data could be calculated which could reflect the quality level of the benthic analyses based on the used

or imported data. Up to now, the project work enables us now to manage metadata, environmental data and biological data in a searchable and sustainable way in the MARLIN system in order to provide information for the benthos and demersal fish for different stakeholders.

6. Supplementary material

An interactive version of the data model of CRITTERBASE (critterbase_data_model.zip) can be found in the digital annex.

7. Literature

Kloss P, Beermann J, Pehlke H, Wrede A, Dannheim J (2020). Taxonomic quality control in MARLIN system. Technical Report of the project ANKER FKZ 0325921, pp. 22.

Ban NC, Alidina HM, Ardron J A (2010). Cumulative impact mapping: advances, relevance and limitations to marine management and conservation, using Canada's Pacific waters as a case study. Marine Policy, 34: 876–886.

Dannheim J, Schröder A, Wätjen K, Gusky M (2013). Gemeinsame Auswertung von Daten zu Benthos und Fischen für das ökologische Effektmonitoring am Offshore-Testfeld „alpha ventus", Schlussbericht zum Projekt Ökologische Begleitforschung am Offshore-Testfeldvorhaben „alpha ventus" zur Evaluierung des Standarduntersuchungskonzeptes des BSH (StUKplus), FKZ: 0327689A, pp. 66.

Dannheim J, Gusky M, Holstein J (2016). Bewertungsansätze für Raumordnung und Genehmigungsverfahren im Hinblick auf das benthische System und Habitatstrukturen. FKZ: 10016990, pp. 38.

Grassle JF (2000). The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modelling and mapping marine biological data in a multidimensional geographic context. Oceanography (Wash DC) 13: 5–9

Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, D'Agrosa C, Bruno JF, et al. (2008). A global map of human impact on marine ecosystems. Science, 319: 948–952.

Vanden Berghe E, Claus S, Appeltans W, Faulwetter S, Arvanitidis C, Somerfield PJ, Aleffi IF, et al. (2009). MacroBen integrated database on benthic invertebrates of European continental shelves: a tool for large-scale analysis across Europe. Marine Ecology Progress Series, 382: 225-238.