Journal of Biotechnology 261 (2017) 177–186

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Review

Terminology supported archiving and publication of environmental science data in PANGAEA

Michael Diepenbroek*, Uwe Schindler, Robert Huber, Stéphane Pesant, Markus Stocker, Janine Felden, Melanie Buss, Matthias Weinrebe

PANGAEA Data Publisher for Earth & Environmental Science, MARUM Center for Marine Environmental Sciences, University of Bremen, P.O. Box 33 04 40, 28334 Bremen, Germany

ARTICLE INFO

Keywords:
Data publishing
Data interoperability
Data findability
Terminologies
Semantics

ABSTRACT

Exemplified on the information system PANGAEA, we describe the application of terminologies for archiving and publishing environmental science data. A terminology catalogue (TC) was embedded into the system, with interfaces allowing to replicate and to manually work on terminologies. For data ingest and archiving, we show how the TC can improve structuring and harmonizing lineage and content descriptions of data sets. Key is the conceptualization of measurement and observation types (parameters) and methods, for which we have implemented a basic syntax and rule set. For data access and dissemination, we have improved findability of data through enrichment of metadata with TC terms. Semantic annotations, e.g. adding term concepts (including synonyms and hierarchies) or mapped terms of different terminologies, facilitate comprehensive data retrievals. The PANGAEA thesaurus of classifying terms, which is part of the TC is used as an umbrella vocabulary that links the various domains and allows drill downs and side drills with various facets. Furthermore, we describe how TC terms can be linked to nominal data values. This improves data harmonization and facilitates structural transformation of heterogeneous data sets to a common schema. Technical developments are complemented by work on the metadata content. Over the last 20 years, more than 100 new parameters have been defined on average per week. Recently, PANGAEA has increasingly been submitting new terms to various terminology services. Matching terms from terminology services with our parameter or method strings is supported programmatically. However, the process ultimately needs manual input by domain experts. The quality of terminology services is an additional limiting factor, and varies with respect to content, editorial, interoperability, and sustainability. Good quality terminology services are the building blocks for the conceptualization of parameters and methods. In our view, they are essential for data interoperability and arguably the most difficult hurdle for data integration. In summary, the application of terminologies has a mutual positive effect for terminology services and information systems such as PANGAEA. On both sides, the application of terminologies improves content, reliability and interoperability.

1. Introduction

Initiated in the 1990s, PANGAEA[1] (Diepenbroek et al., 2002) has evolved from a paleoclimate data archive to a multidisciplinary data publisher for environmental sciences. PANGAEA is a World Data Center accredited by the International Council for Science World Data System (ICSU WDS)[2] and a World Radiation Monitoring Center (WRMC)[3] within the World Meteorological Organisation Information System (WIS).[4] From its earliest stages, data were archived consistently and annotated according to how they were produced, including information about principal investigators, measurement and observation types, sampling and analysis methods, and devices as well as references to literature. Later, the data structure was adapted to store descriptions of data sets with metadata elements comparable to bibliographic descriptions. Lacking standards, the concept of data entity granularity changed several times. This required reorganisation of the metadata but

* Corresponding author.
E-mail addresses: mdiepenbroek@pangaea.de (M. Diepenbroek), uschindler@pangaea.de (U. Schindler), rhuber@uni-bremen.de (R. Huber), spesant@marum.de (S. Pesant), mstocker@marum.de (M. Stocker), jfelden@marum.de (J. Felden), mbuss@marum.de (M. Buss).

[1] https://www.pangaea.de/.

[2] https://www.icsu-wds.org/.

[3] http://bsrn.awi.de/.

[4] http://www.wmo.int/pages/prog/www/WIS/.

http://dx.doi.org/10.1016/j.jbiotec.2017.07.016
Received 16 February 2017; Received in revised form 17 July 2017; Accepted 19 July 2017
Available online 23 July 2017

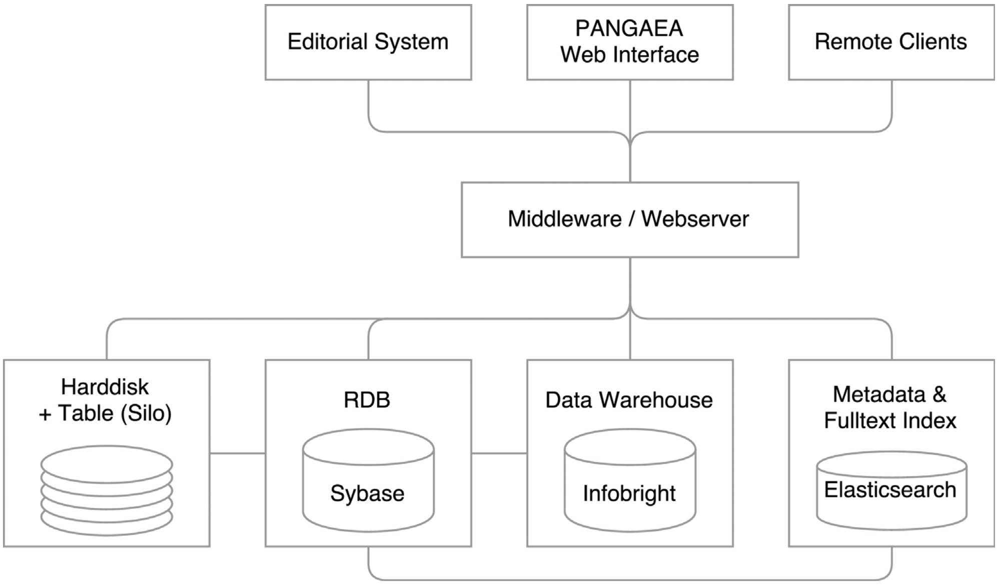


Fig. 1. Basic technical architecture of PANGAEA.

never led to information losses or deficiencies in the metadata. In January 2005, the first data sets were registered with a DOI Name. Today, PANGAEA holds around 360 thousand citable data sets comprising more than 11 billion data items—numerical and textual data as well as binaries such as images, videos, or files with community specific mime types. Each data item is a georeferenced record including the parameter value, parameter type, and the spatial and temporal coverage. Unless they are values of parameters, spatio-temporal values themselves are not data items. Currently, over 20% of published data sets include at least one author linked to ORCID and almost 8% of published data sets have all authors linked to ORCID.

Structured data and metadata with consistent semantics are prerequisites for data usability, in particular for interoperability of data and efficient data integration (Cruz and Xiao, 2005; Wilkinson et al., 2016; Ćwiek-Kupczyńska et al., 2016). Large-scale questions in science, such as global warming, invasive species spread, and resource depletion, increasingly require the collection of disparate data sets from various data sources. To a certain extent, relational models provide a basis for consistent storage of data and metadata, and also allow marshalling and mapping to different structures and data models. Nevertheless, with data inventories growing rapidly over the last decades, careful harmonization of data and metadata becomes increasingly important.

In PANGAEA, tens of thousand different measurement and observation types have been defined, in particular for biodiversity related data. Most of these data types—known as parameters—are complex and consist of terms that often can be matched with terms in existing, externally curated ontologies, thesauri, and vocabularies. Further metadata fields that can be matched with external resources include methods and devices. They are part of the lineage information. Complementing PANGAEA metadata with a component that supports using and managing terminologies not only improves the consistency of archived data but also makes data more reliable and interoperable, significantly improves the findability of data sets, in particular enables comprehensive and consistent search facets, and finally facilitates long term curation of data and metadata.

Ontologies, thesauri, and vocabularies for environmental sciences have been evolving tremendously during the last decade. They have emerged as a key tool to formally represent and semantically organize aspects of the real world. Such terminologies facilitate communication between experts and enable the application of computational techniques that extract useful information from available data and metadata (del Mar Roldán García et al., 2016; Mate et al., 2015). Taxonomies, chemical compounds, traits, or geographical locations: there is hardly a domain not being covered. For some domains several terminologies are available.

Features and usability of terminologies and terminology services vary greatly. To integrate terminologies into applications, interoperability is indispensable. Ideally, services allow to look up, download, and submit new terms. Some services have APIs with flexible parameterization of requests, others are restricted to bulk downloads of the whole terminology, and some terminologies can only be searched and browsed on a proprietary website. Terminology editorial workflows range from peer reviewed by expert communities to restricted to specialized closed groups. Also critical is sustainability, in particular the persistence of term descriptions and identifiers. Sustainability is not only important for technical operations, but also for maintenance and actuality of contents. Less problematic are heterogeneities due to differences in syntax and schemata. Here, mapping, matching and alignment techniques can be applied and tools exist to support such processes (Bergman, 2014). In cases where different terminologies cover the same domain, semantic inconsistencies might raise additional practical issues when integrating terminologies into specific applications (Saripalle, 2008; Kotis and Lanzenberger, 2008).

Using PANGAEA as an example, this paper describes the application of terminologies in a data archiving and publication environment. We show how terminologies are integrated into the system and discuss the requirements and benefits for data ingest, curation and dissemination. Finally, we describe some of the current limitations in applying terminologies.

2. Technical background of PANGAEA

The basic technical structure of the information system PANGAEA corresponds to a three-tiered architecture with a number of backends, clients and middleware components controlling the information flow and quality (Fig. 1). On the backend, a relational database management system (RDBMS) and file systems are used for information storage. Tape silos are used as tertiary storage medium. To ensure fast access to subsets of the PANGAEA data inventory, numerical and textual data are mirrored into a data warehouse. Metadata and textual data are also marshalled and indexed for the PANGAEA search engine (Elasticsearch[5]). On the frontend, PANGAEA is furnished with an editorial

[5] https://www.elastic.co/products/elasticsearch.

system for data ingest and curation. As part of the middleware, an API[6] is used by the PANGAEA web interface as well as by numerous clients worldwide (e.g. the GEOSS data portal[7]). All interfaces to the information system are based on web services, including map support (Google Earth, Google Maps).

The challenge of managing the heterogeneous and dynamic bio- and geosciences data was met in PANGAEA through a flexible data model. It reflects the data processing steps in the environmental science domains and can handle any relevant analytical data (Fig. 2). The system uses a normalized relational structure for data and metadata. Ingest and recompilation of complex data sets for display or download is achieved through several middleware components—using the original data matrix configuration, stored during ingest as part of each data set.

For the system to conform, interoperate, and make use of external (third-party) terminology services, the conception and implementation of new system components, as well as restructuring of existing components, was required. The conversion affected technical components on all tiers but also included extensive work on the data and metadata holdings. The main areas of work comprise (1) a new component allowing to manage terminologies, called terminology catalogue (TC); (2) interfaces that allow to synchronize the TC with external terminologies; (3) embedding the TC into the current editorial workflows for ingesting and archiving data into PANGAEA; and finally (4) embedding the TC into the data access and dissemination workflows.

3. Terminology catalogue

The implementation requirements for the terminology catalogue (TC) were:

- Minimal effort to embed the new component into the PANGAEA relational data model.
- Editorial interface (GUI) seamlessly embedded into the editorial system, allowing to define or update terms.
- Interfaces for the ingest of new or updated terms of external terminologies.
- API that handles requests for TC content, whereby outputs should comply to RDF (Schreiber and Raimond, 2014) and OWL (Hitzler et al., 2012) recommendations.
- Interfaces for using TC terms in the conceptualization of parameter, methods and devices.

3.1. Data model and structure

For the data model there were two options: (1) relational model or (2) formal ontology model. Interfacing the latter approach with the current PANGAEA editorial system would have been laborious and complex. Moreover, a simple marshalling routine is sufficient to produce RDF output from our relational model. We therefore chose a relational model consisting of tables for the terms (`Term`); for relations between terms (`TermRelation`); for information about relation types (`RelationType`: properties); for general information about the terminologies in the catalogue (`Terminology`); to categorize terms (`TermCategory`: class, attribute, etc.); and to assign a term status (`TermStatus`: pending, accepted, etc.) (Fig. 4).

Columns in the `Term` table allow to specify a term label and an additional identifier. These are term attributes. Further attributes can be assigned to individual terms using relations between terms, e.g. for the species *Canis lupus* the taxonomic level can be described as *Canis lupus* has rank species. The conceived structure is simple and flexible. Because many terms are primarily for PANGAEA internal use we settled on internal IDs as primary keys. Terms that have been replicated from external terminologies have URIs as additional keys. They are used to match TC contents with external terminologies, e.g. to update TC terms (see Section 3.3).

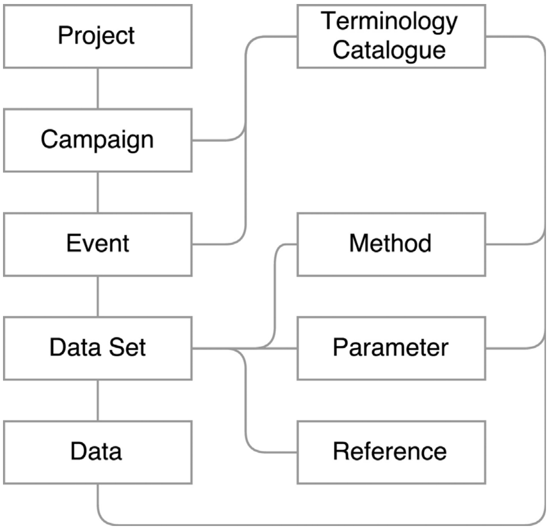


Fig. 2. Simplified data model of PANGAEA. Lines show relationships between the different entities. They are partly 1:n or m:n relationships.

3.2. Editorial interface for terminology curation

The TC is not only used to replicate external terminologies but also to curate PANGAEA terminologies. Curation relies on an own editorial interface. The tool we created for this purpose is fully integrated into the PANGAEA editorial system. Compared to a full-featured ontology management system (e.g. Protégé (Musen, 2015)) the TC is more simple. Still, the editorial interface enables efficient work with terminologies. Specifically, it allows to search, browse, create, update, or delete terms, and to set relations between terms, e.g. to build and display hierarchies (Fig. 3).

3.3. Synchronization of the TC with external terminologies

The TC was initially populated with terms of a number of terminologies, in particular WoRMS,[8] ChEBI,[9] EnvO,[10] QUDT,[11] (Hodgson et al., 2014) and PATO.[12] Replication was done either using supplied APIs or bulk downloads of RDF files. With WoRMS, PANGAEA has for some time maintained a bidirectional workflow that includes submission of new species names and regular downloads. More recently, in particular within the German e-infrastructure project de.NBI,[13] similar workflows have been created with the information systems BacDive,[14] Brenda,[15] and Silva.[16]

For long term maintenance, bilateral workflows are, however, inefficient. Already now, PANGAEA requires regular updates from various terminologies. Moreover, more relevant and good quality terminology services do emerge. Considering that interfaces change and PANGAEA is only one among potentially many service clients, one-for-all terminology services are in principle the better choice. Aggregating or mediating services such as Ontobee[17] or BioPortal[18] offer lookups and downloads. For automatic replication, however, standard APIs that

[6] https://ws.pangaea.de.

[7] http://www.geoportal.org/.

[8] http://www.marinespecies.org/.

[9] https://www.ebi.ac.uk/chebi/.

[10] http://environmentontology.org/.

[11] http://www.qudt.org/.

[12] http://www.obofoundry.org/ontology/pato.html.

[13] https://www.denbi.de/.

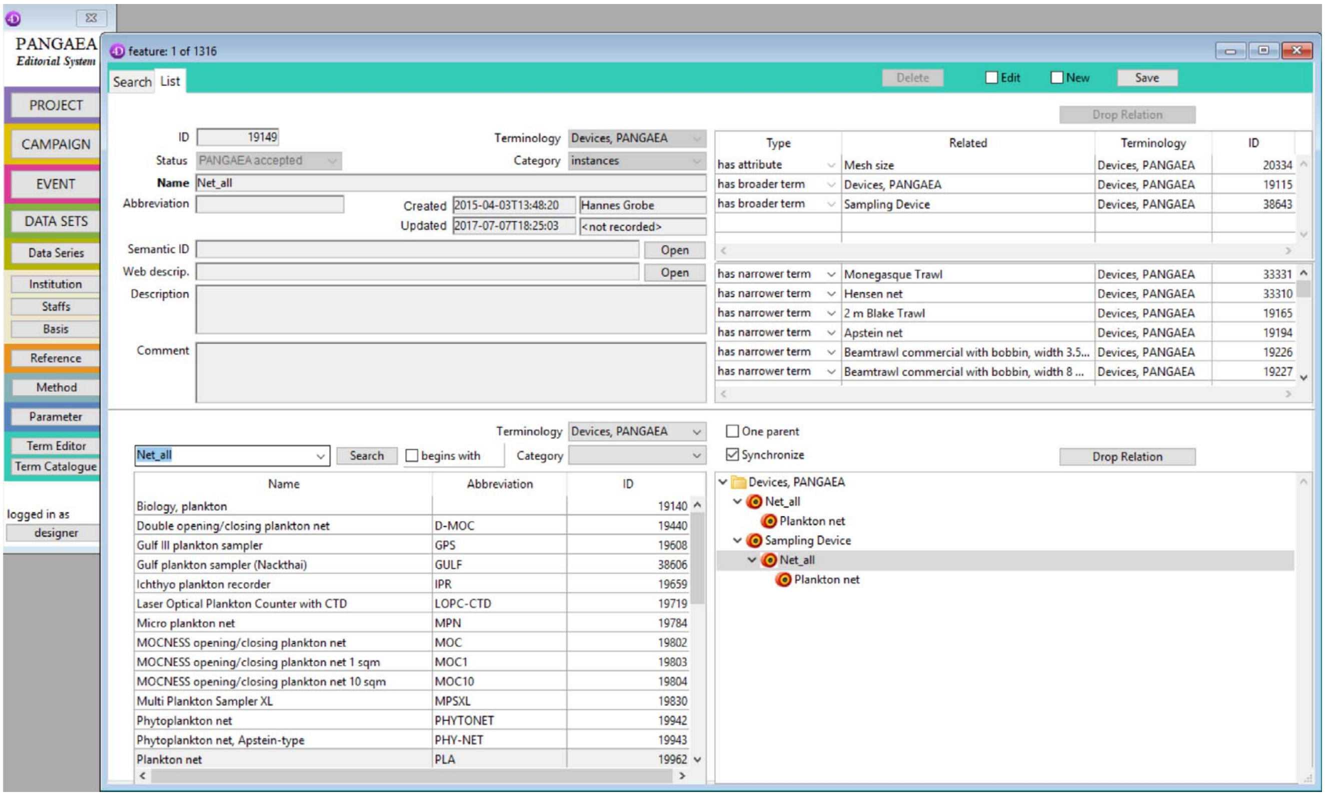
[14] https://www.dsmz.de/bacterial-diversity/bacdive.html.

[15] http://www.brenda-enzymes.org/.

[16] https://www.arb-silva.de/.

[17] http://www.ontobee.org/.

[18] https://bioportal.bioontology.org/.



feature: 1 of 1316

□ Edit □ New

ID: 19149
Terminology: Devices, PANGAEA
Status: PANGAEA accepted
Category: instances
Name: Net_all
Abbreviation:
Created: 2015-04-03T13:48:20 Hannes Grobe
Updated: 2017-07-07T18:25:03 <not recorded>
Semantic ID:
Web descrip.:
Description:
Comment:

logged in as: designer

Type	Related	Terminology	ID
has attribute	Mesh size	Devices, PANGAEA	20334
has broader term	Devices, PANGAEA	Devices, PANGAEA	19115
has broader term	Sampling Device	Devices, PANGAEA	38643

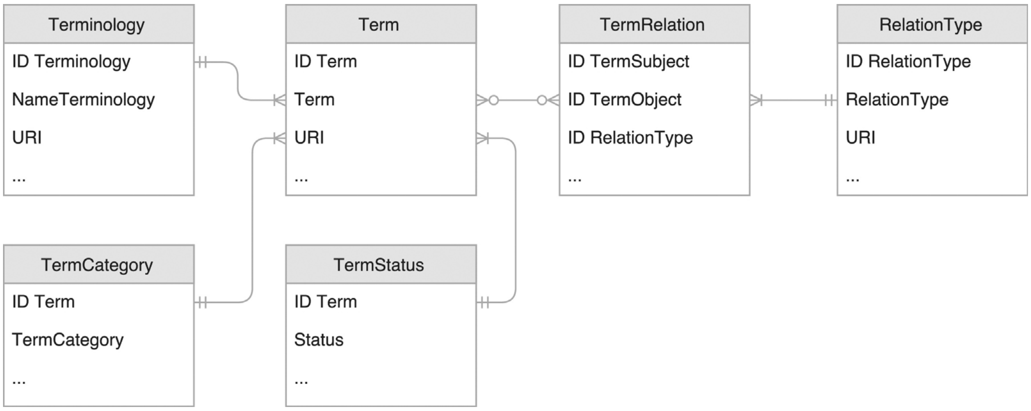
has narrower term	Monegasque Trawl	Devices, PANGAEA	33331
has narrower term	Hensen net	Devices, PANGAEA	33310
has narrower term	2 m Blake Trawl	Devices, PANGAEA	19165
has narrower term	Apstein net	Devices, PANGAEA	19194
has narrower term	Beamtrawl commercial with bobbin, width 3.5...	Devices, PANGAEA	19226
has narrower term	Beamtrawl commercial with bobbin, width 8 ...	Devices, PANGAEA	19227

Terminology: Devices, PANGAEA
Net_all — □ begins with
Category:
□ One parent
☑ Synchronize

Name	Abbreviation	ID
Biology, plankton		19140
Double opening/closing plankton net	D-MOC	19440
Gulf III plankton sampler	GPS	19608
Gulf plankton sampler (Nackthai)	GULF	38606
Ichthyo plankton recorder	IPR	19659
Laser Optical Plankton Counter with CTD	LOPC-CTD	19719
Micro plankton net	MPN	19784
MOCNESS opening/closing plankton net	MOC	19802
MOCNESS opening/closing plankton net 1 sqm	MOC1	19803
MOCNESS opening/closing plankton net 10 sqm	MOC10	19804
Multi Plankton Sampler XL	MPSXL	19830
Phytoplankton net	PHYTONET	19942
Phytoplankton net, Apstein-type	PHY-NET	19943
Plankton net	PLA	19962

- Devices, PANGAEA
  - Net_all
    - Plankton net
  - Sampling Device
    - Net_all
      - Plankton net

Fig. 3. Screenshot of the TC interface as part of the PANGAEA editorial system.


Terminology
ID Terminology
NameTerminology
URI
…

Term
ID Term
Term
URI
…

TermRelation
ID TermSubject
ID TermObject
ID RelationType
…

RelationType
ID RelationType
RelationType
URI
…

TermCategory
ID Term
TermCategory
…

TermStatus
ID Term
Status
…

Fig. 4. Relational structure of the terminology catalogue (TC). For simplicity, only the essential table columns are shown in the ER diagram (Crow's foot notation).

support harvesting new or updated terms, or even subscription services, would be better solutions. Vice versa, submission of terms to corresponding services, for review and ingest, is also a desirable functionality. With the GFBio Terminology Service (TS)[19] (Karam et al., 2016), the German GFBio project[20] is currently heading for such an environment. The GFBio TS serves a number of terminologies. For the PANGAEA TC, we are currently setting up a prototype for automated replication of terms with the GFBio TS (Fig. 5). Regular incremental replication of terms into PANGAEA will occur using the GFBio TS API.[21] For this purpose, the next version of the API will support requesting terms that are new or have been updated within a specific time interval. For the submission of new TC terms to the various terminology services, APIs are presently missing and are not likely to be supplied in the near future because of the heterogeneity of terminologies. In addition, the submission of terms often requires communication with experts.

3.4. Editorial work

Although most of the needed terms could be replicated from external terminologies, manual work on terminologies cannot be avoided. As part of the lineage information of a data set, methods and devices are examples requiring editorial work (cf. below). Additional examples include terms that fall into the scope of an existing terminology service. We try to first retrieve such terms from these terminologies using supplied web interfaces for searching and browsing. If unsuccessful, terms are defined in the TC and subsequently submitted to the corresponding terminology for review and ingest. Currently, we are mostly submitting new terms to WoRMS and ChEBI. These terms get a preliminary status in the TC. After acceptance, terms can be replicated back to the TC. A particular problem occurs if TC terms with preliminary status do not have persistent identifiers, because they were not

[19] http://terminologies.gfbio.org/.
[20] http://www.gfbio.org/.
[21] http://terminologies.gfbio.org/developer_section/api.html.

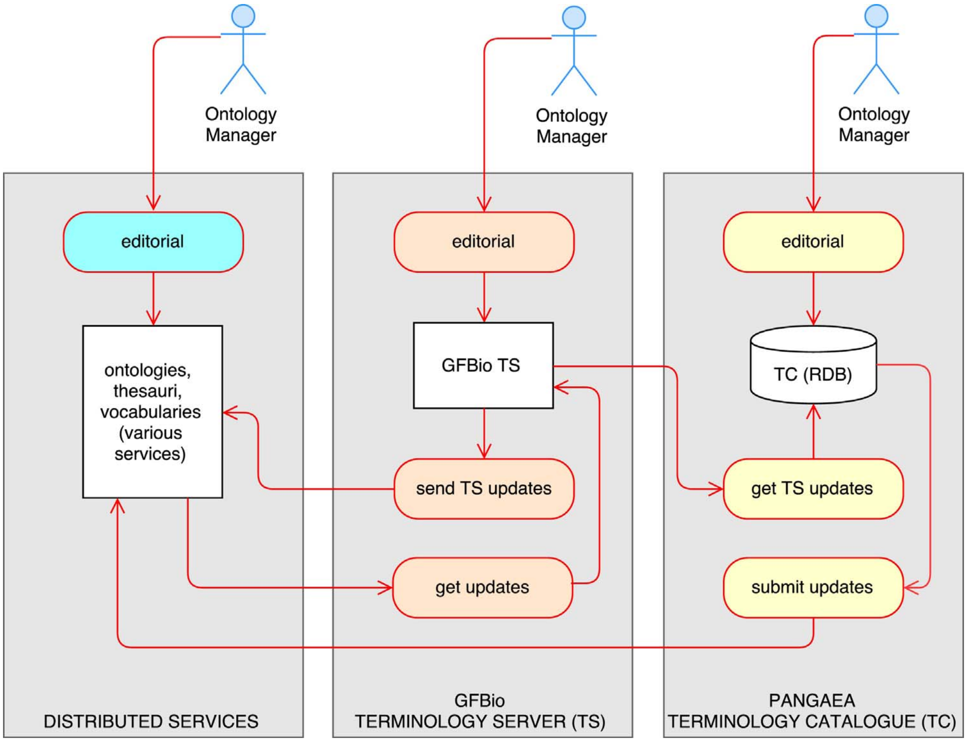


Fig. 5. UML activity diagram showing the replication and submission of terms between the PANGAEA TC, the GFBio TS, and distributed terminology services.

available at the time of submission. This complicates identification of terms when they are replicated back into the TC. The time needed for review and acceptance of terms varies. Together with the time needed to release new terms it might take several months before they get the final URI.

4. Use case: data ingest and archiving

4.1. Workflow description

The PANGAEA workflow for data ingest and archiving (Fig. 6) complies with the OAIS standard (CCSDS, 2012). Data are submitted using a ticket system (Jira[22]) and assigned to an editor who is a specialist in the corresponding data domain. Preparation of the data for import is done with the editorial system. Data editors check the completeness and validity of data and metadata, and reformat data according to the PANGAEA ingest format. The editorial review is complemented by inviting authors and external reviewers (e.g. reviewers of articles supplemented by the data). After being accepted, the data are archived, provided with a DOI, and metadata are registered at DataCite.[23] Besides manual data submission and ingest, the system also supports submission and staging of pre-processed data by external, authenticated processes (e.g. recurrent submission of monitoring data).

Preparation of data for ingest is supported by the PANGAEA TC, namely in the conceptualization of parameters and methods as well as the definition of campaigns (cruises or field campaigns) and events (sampling events or experiments). It is furthermore used to define locations (gazetteers) and keywords. The latter can be assigned to nearly all types of records in the data model (Fig. 2). Finally, it is used to relate TC terms directly to data values using the term ID, thus ensuring consistency of, e.g., lists of species. For this purpose, the nominal values of a data matrix are matched with TC terms.

4.2. Definition of parameters

Terminologies support the consistency and integrity of metadata. Toward this aim, controlled vocabularies (CV) are often used as an integral part of standards (e.g. feature catalogues (ISO/TC 211, 2016) as part of ISO 19115 (ISO/TC 211, 2014)). However, a practice and generally accepted model for thorough and comprehensive conceptualization of parameters and methods are so far missing. Parameter selections such as the Essential Climate Variables (Bojinski et al., 2014) or the more recent Essential Biodiversity Variables (Pereira et al., 2013) only reflect a small section of possible measurement types.

PANGAEA maintains more than 140 thousand different parameters as string values, with an estimated 30 thousand unique terms that can be roughly categorized as quantity kinds (cf. definition given in QUDT); terms (features) belonging to nomenclatures such as taxonomies, chemical compounds or traits; and unit specifications. Frequently, quantity kinds are for SI and derived units but also include other types for which data can be measured at all levels of scale.

Most of the parameters are complex, meaning that terms of several categories are combined into a compound concept. For the definition of parameters, we conceived a model that allows to link terms according to their category following a syntax schema (Fig. 7). Formal notations comply with the Extended Backus-Naur Format (EBNF) in W3C notation[24] and are shown in Listing 1.

The schema implies a set of rules:

- A `ParameterInstance` must begin with a `QuantityKind` which is any string, and may be followed by (1) a colon (':') in case the `ParameterInstance` is further defined by associated features or (2) a `QuantityUnit`, or (3) both.
- A `QuantityKind` is specified by a `QuantityKindName` and can be complemented by a `QuantityKindAttribute`, e.g. a statistical constraint such as 'mean' or 'standard deviation', separated by a comma from the `QuantityKindName`.
- A `Feature` is specified by a `FeatureName` and can appear multiple

[22] https://en.wikipedia.org/wiki/Jira_(software).

[23] https://datacite.org/.

[24] http://www.w3.org/TR/xml/#sec-notation.

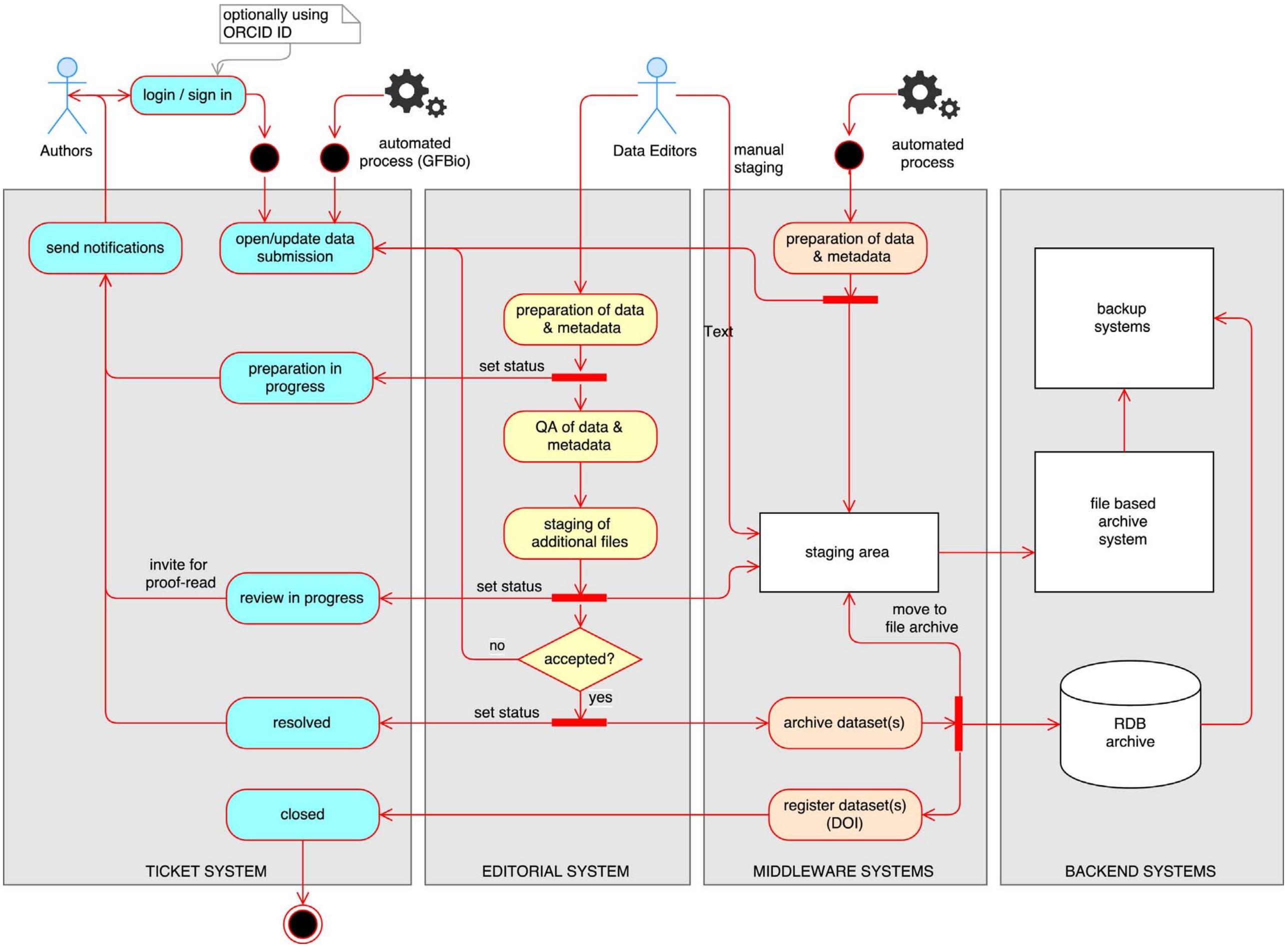


Fig. 6. UML activity diagram for the PANGAEA ingest and archiving workflow.

times. Features are optional.

- A simple controlled vocabulary of prepositions (‘in’, ‘at’, ‘per’, etc.) is used to specify relations between features. Alternatively, a semicolon (‘;’) can be used if a relation is unclear.
- A `Feature` can be further specified by adding an optional `FeatureAttribute` following the `FeatureName`, separated by a comma.
- A `ParameterInstance` may include a `QuantityUnit`, which must be noted in square brackets.

Each parameter instance is represented by an internal ID and a string for the concatenated linked terms. The latter is necessary to generate human-readable output and for compliance of newly generated parameter instances with existing parameter names that have not yet been normalized. Parameter examples include:

- Assimilation rate: 14C of carbon dioxide [mg{C}/kg/day]
- Production rate: egg of Calanus finmarchicus per female [#/#{Ind}/day]
- Ratio: d_{13}C in Calanoides carinatus [per mil PDB]
- Volume: fecal pellet of Calanus finmarchicus [µm^3]

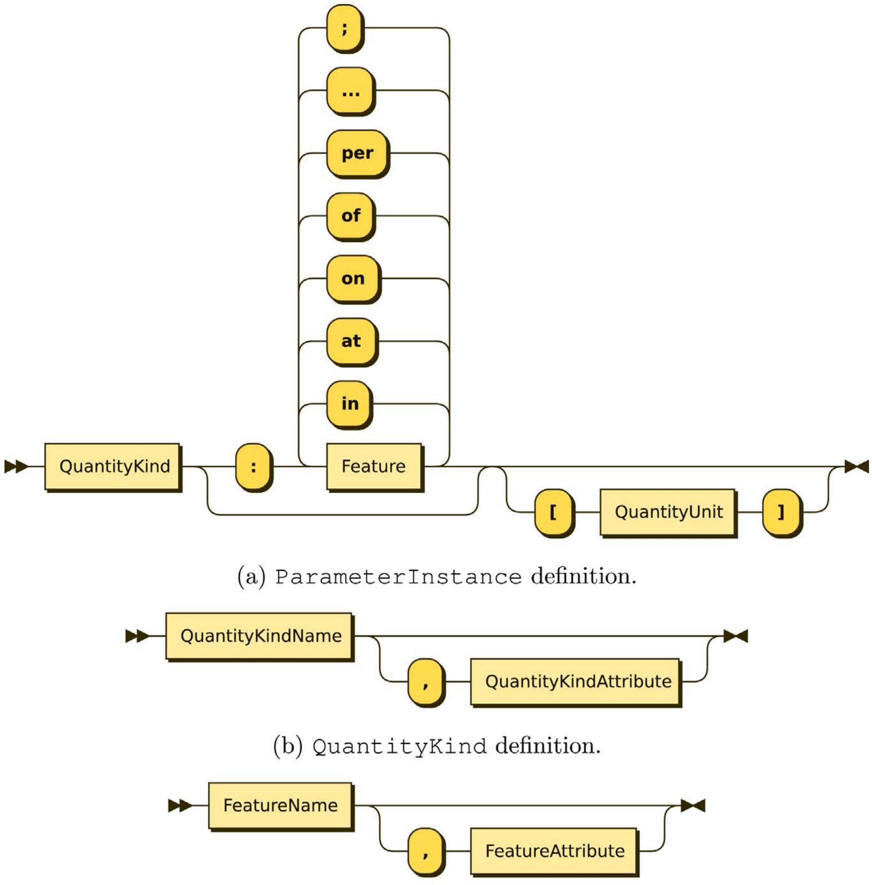
In the latter example ‘Volume’ is the quantity kind, ‘fecal pellet’ and ‘Calanus finmarchicus’ are the features, and ‘µm^3’ is the quantity unit.

The model and rule set are basic. It is arguably the minimum required for a parameter definition. However they, can be extended and refined with the ongoing curatorial work in PANGAEA. We have analysed the unique terms contained in PANGAEA parameter strings and identified a number of subcategories. It is foreseen that terms such as ‘ratio’ or ‘flux’ imply further semantics that influence the proposed syntax and corresponding rules. We plan to evolve the list of parameters into a full-featured ontology with parameters corresponding to concepts. However, it is important to remember that developments also have to be aligned with the practicalities of the curatorial work. Ultimately, they should lead to more efficiency.

As a first step, we applied a parser routine to match TC terms with parameter strings. Relations between a parameter and a TC term were stored in an intermediate table with status pending to indicate that evaluation by a data editor is needed. New parameters are defined as required by the TC (Fig. 8). For existing parameters a considerable effort will be needed. Due to the high number of parameters and potential damages to the integrity of archived data it is unlikely that all parameters can be transformed. Eventually the process needs to involve data providers and communities.

4.3. Methods

Specification of methods used to make measurements or observations are an indispensable part of the lineage information. Method specifications comprise procedures, devices and conditions. Over the last decades, more than 5000 methods have been assembled in PANGAEA. One of the difficulties arises due to the inconsistent level of detail supplied by data providers. For instance, data sets might be produced using a method ‘mass spectrometry’ or ‘mass spectrometry on acidified combusted sample (centrifuged)’. Data editors need to evaluate the validity of methods applied. Similar to parameters, methods need to be structured following a single model. At PANGAEA, this work is in an early stage. At present, the basic and most frequent terms are



(a) `ParameterInstance` definition.

(b) `QuantityKind` definition.

(c) `Feature` definition.

Fig. 7. Model for the definition of parameters.

assembled in the TC.

Listing 1
Parameter definitions.

```
ParameterInstance::= QuantityKind
  (’:’ Feature (
(’in’ | ‘at’ | ‘on’ | ‘of’ | ‘per’ | ‘...’ | ‘;’)
Feature)*
  )?
  (’[’ QuantityUnit ‘]’)?

QuantityKind::= QuantityKindName
  (’,’ QuantityKindAttribute)?

Feature::= FeatureName
  (’,’ FeatureAttribute)?
```

4.4. Events

Measurement and observation events, such as sampling or experiments, are part of the lineage information and are characterized by a set of attributes including a name, coverage, and devices used. As for methods, the level of supplied details varies as device specification is sometimes only at the level of device type and sometimes also at the level of device instance. Formerly a simple list, devices were now structured and categorized, and incorporated into the TC. The new structure allows to define for each device a default set of attributes (keys) that can be used to specify further details about the device. For instance, the sampling device ‘plankton net’ has various attributes, e.g. ‘mesh size’, also abstracted from narrower terms such as ‘bongo net’ (Fig. 3). An intermediate table between event and the TC holds the values for the attributes, which formerly had to be stored in the comment field of the event, provided this information was stored at all. In summary, the developments led to a flexible, structured, and more detailed description of events. Because of the considerable overlap between devices and methods we plan to merge them.

4.5. Campaigns

Field campaigns, cruises, and other types of research series and endeavours are part of the lineage information and also need a flexible description using key-value pairs linked to the TC. A structure identical to the one for devices was implemented. Campaign types have been defined in the TC and furnished with a set of attributes, e.g. a ‘cruise’ typically has a departure and arrival harbour.

4.6. Gazetteers

Areas and locations are assembled as gazetteers in the TC. Data set coverages, given as geographical positions in the data or related events, are automatically matched with areas using polygons representing them. Selected is the area with the best fit. Areas are organized hierarchically. For the marine part, the IHO sea areas[25] and the GEBCO Undersea Feature Names[26] are used. On top, GeoNames (Wick, 2006) is used as a repository for tracking synonyms of all locations and area names. This allows to search for foreign names of places, countries and oceans in the PANGAEA Elasticsearch index. Besides, on the event level,

[25] International Hydrographic Organization (IHO) and the Intergovernmental Oceanographic Commission (IOC) of UNESCO.

[26] GEBCO Gazetteer of Undersea Feature Names, http://www.gebco.net/data_and_products/undersea_feature_names/.

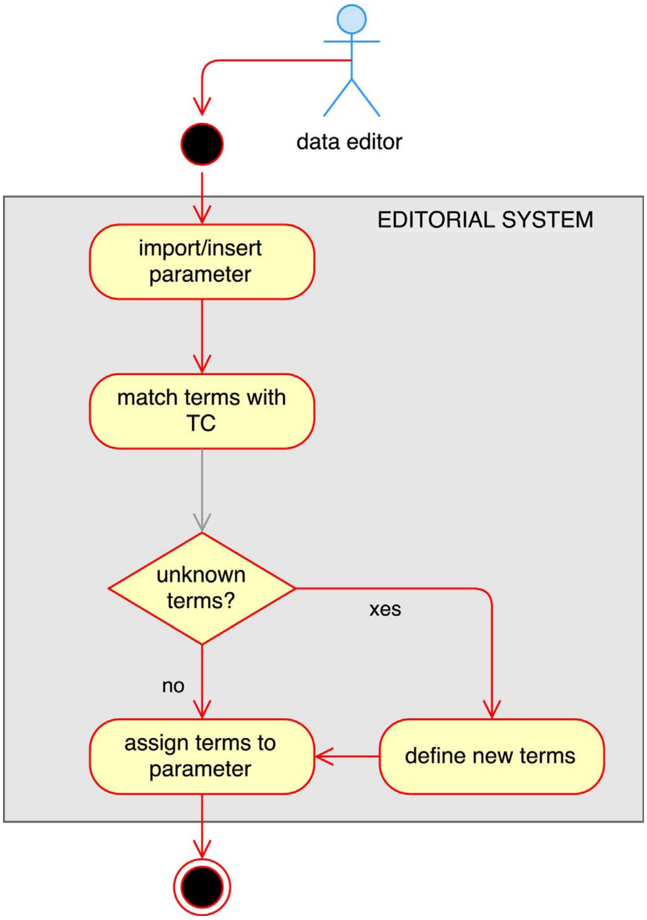


Fig. 8. UML activity diagram for the definition of new parameters.

editors may specify additional locations (typically local names). Encountered problems include inconsistencies in language use and variations in transcriptions. Harmonization of terms in the TC will include such variations as synonyms, ideally curated in external gazetteers.

4.7. TC terms as data values

Data are usually submitted as spreadsheets having complex parameter names heading the columns. In principle, it is possible to reformat the data in order to split this complex information into separate columns. For instance, the matrix with columns as in the data set by Findlay et al. (2010) and shown here in Table 2 could also be written as shown in Table 3.

In order to ensure the compliance of species names (as they are given in the second matrix) with the standard terms in the TC, names are matched during import. For this purpose, we use a specialized TC full text index that includes synonyms, stemming, folding, and n-grams. This allows to deliver ranked results of possible term matches. Incorrect or unrecognized names need to be adjusted. Alternatively, term IDs or URIs could be used in the ingest matrix, instead of the names. Eventually, such data are stored as term IDs. It is thus possible to match any nominal values against TC terms. The approach is not only suitable for taxon names but also for, e.g., lithologies or any kind of nominal scale. The new functionality for linking data values with TC terms fosters harmonization of data. However, dependencies between columns in a matrix are not made explicit in the metadata. In the reformatted example matrix (Table 3), ‘growth rate’ might be related to values of other columns. In the data set by Ow et al. (2016), ‘Growth rate’ in column 20 may refer to ‘Species’ or, say, ‘Specific leaf area’. Not to have such dependencies explicit in the metadata can lead to misinterpretations of data.

Table 2
Data matrix as typically submitted, with complex parameter names heading the columns.

Species A, growth rate	Species B, growth rate
Value x1	Value y1
Value x2	Value y2

Table 3
Reformatted data matrix, with complex information split into separate columns.

Species	Growth rate
Species A	Value x1
Species A	Value x2
Species B	Value y1
Species B	Value y2

5. Use case: access and dissemination of data and metadata

5.1. Workflow description

Data and metadata are marshalled from the relational database into XML files according to the PANGAEA internal XML schema. They are subsequently indexed using Elasticsearch, whereby the metadata are semantically enriched using the marshalled and indexed content of the TC. To support fast and 24/7 access, the marshalled data sets are cached on the website. Based on the PANGAEA internal schema, metadata for various other content standards are produced on demand via XSLT and also enriched using the TC (Fig. 9).

5.2. Metadata enrichment

Initially, metadata enrichment is done by retrieving TC terms matching terms occurring within a metadata description. In a second step, related, broader or synonym terms are retrieved and added to the index. In parallel, for each retrieved term, a lookup is made for a corresponding term in the thesaurus used to classify data according to subject, environment, and various other categories. This thesaurus is a special terminology of the TC. If the lookup is successful, the term is added to a corresponding tag in the XML scheme used to enable comprehensive and well structured faceted searches on metadata. Currently, there are only three facets, namely ‘topic’, ‘device’, and ‘location’ that are implemented in this way. An additional facet for taxonomic trees is in preparation. Other facets, such as ‘author’ and ‘project’, are derived directly from the original metadata. The PANGAEA thesaurus of classifying terms is currently integrated with a corresponding terminology of the GFBio TS—created, for the same purpose, by the larger GFBio consortium.

The lookup between the various terminologies and the PANGAEA thesaurus is done using string comparison. This is facilitated by duplicating facet-relevant terms to this thesaurus. Advantages are that no explicit relations between terms are needed, and maintenance of the thesaurus requires less effort. The few cases that might cause incorrect matches due to possible homonyms are neglected. So far this problem has not materialized. Should the problem occur, it can be resolved by adding an explicit relation to the correct term.

5.3. Access to data and metadata

Access to data is provided via the PANGAEA frontend by first retrieving metadata results (Fig. 9, step 1) and subsequently selecting a specific data set for display or download (Fig. 9, step 2). Moreover, the

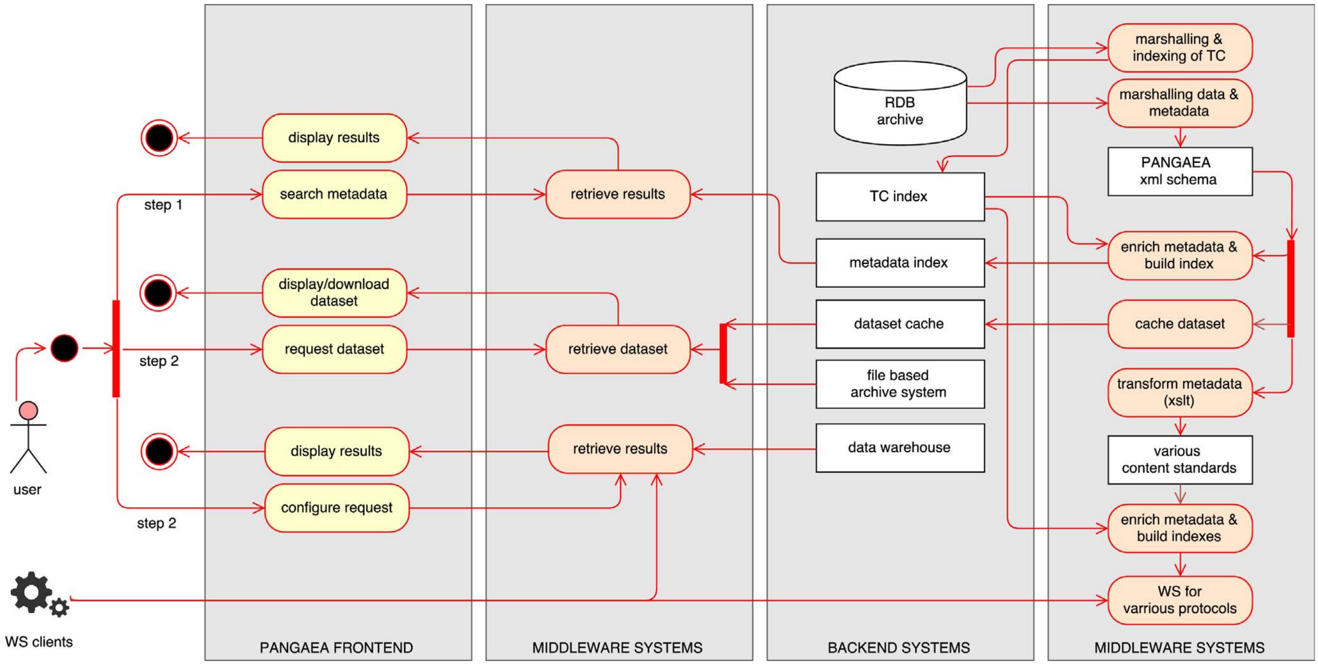


Fig. 9. UML activity diagram showing access and dissemination of data and metadata.

result set can be used for the parameterization of a PANGAEA data warehouse query. This service supports the integrated retrieval of data from all data sets in the result.[27] The list of parameters and geocodes available for the configuration of the data matrix to be downloaded is generated dynamically from the result set using the metadata index.

The data warehouse is a very efficient tool to compile data, even from tens of thousands of different data sets. Nevertheless, users must be aware that such compilations cannot be taken as final data products. Firstly, users need to ensure that data sets in the result set are fit for the specific scientific application. For this, users may utilize the list of citations included in the download package (zip). Secondly, downloaded data do not take into account the original dependencies between parameters (see Section 4.7). Indeed, misinterpretations could be avoided, or at least reduced, if the conceptualization of parameters is extended to the description of possible dependencies between parameters, respectively relationships between different concepts. Such dependencies could be inserted into the TC as relationships between terms and be used to set additional constraints for the configuration of data warehouse queries.

5.4. Further dissemination of data and metadata

PANGAEA supplies a number of web services for access to data and metadata. Consumers can harvest the PANGAEA metadata using the OAI-PMH protocol.[28] The service supports various international content standards for metadata, such as Dublin Core and ISO 19139 (implementation of ISO 19115), but also formats such as GCMD DIF.[29] PANGAEA thus contributes to various portals, including GEOSS,[30] ICSU-WDS, OA-ICC, and GFBio. Proprietary controlled vocabularies can pose significant difficulties when mapping metadata to other formats. For instance, GCMD DIF requires that parameter names, platform names, or geographical locations are chosen from the GCMD thesaurus.[31] To address this limitation, the TC allows to map PANGAEA's own terms to those from external controlled vocabularies. PANGAEA has mirrored the GCMD terms into the TC and manually linked PANGAEA's own terms to the imported terms using an "equivalent to" relation. The metadata schema mapping code (XSL) uses these relations to enrich DIF metadata with controlled GCMD terms.

All PANGAEA APIs are available through a single entry point,[32] using several techniques: REST-based endpoints (OAI-PMH, Elasticsearch) and SOAP APIs for well structured, proprietary access. The API to access PANGAEA's data warehouse is an example. It allows consumers to quickly extract large amounts of data for processing. For example, the German project C3Grid used this API in its portal (Kindermann, 2013) to drive climate models and the cabled seafloor monitoring infrastructure Neptune Canada[33] used PANGAEA in this way as complementary data source to cover physical oceanography.

6. Conclusion

We have embedded the structure and functions of a terminology catalogue (TC) into the PANGAEA system. A considerable effort and several iterations were needed to meet the requirements of a highly efficient environment for the editorial and publication of biodiversity related data. The effort is well balanced and is adding value to PANGAEA.

For data ingest and archiving, we described how the TC can be applied to various types of metadata, namely to the definition of parameters, methods, and devices. For data access and dissemination, the added value is clearly on data findability, largely a result of enriching metadata with TC terms. With science getting increasingly complex, simple keywording is insufficient (Fernández et al., 2011). Semantic annotations, such as adding whole term concepts (including synonyms and hierarchies) as well as mapping terms between different terminologies, facilitate comprehensive data retrievals, even if users do

[27] Example result set: https://www.pangaea.de/?t=Biosphere&q=Ocean+Acidification. Select 'data warehouse' or use https://www.pangaea.de/advanced/datawarehouse.php?t=Biosphere&q=Ocean+Acidification (supplied query example links might be subject to changes).

[28] https://ws.pangaea.de/oai/.

[29] Directory Interchange Format, http://gcmd.gsfc.nasa.gov/add/difguide/.

[30] http://www.geoportal.org/.

[31] Global Change Master Directory (GCMD) Keywords, https://earthdata.nasa.gov/about/gcmd/global-change-master-directory-gcmd-keywords.

[32] https://ws.pangaea.de/.

[33] http://www.oceannetworks.ca/.

not have domain expertise. The PANGAEA thesaurus of classifying terms, which is part of the TC, in particular, is used as an umbrella terminology linking the various domains and allowing drill downs and side drills with various facets. The PANGAEA thesaurus is currently integrated in a corresponding thesaurus elaborated as part of the GFBio project. The integrated product, also to be used by PANGAEA, will contain contributions from several project members.

We have shown how TC terms can be linked to nominal data values. As for metadata, this linking not only leads to improved harmonization of data and increased reliability in the usage of data and metadata but also offers the opportunity to overcome structural differences of archived data sets. Parameter components that are identical to nominal data values can be transformed into a data matrix that keeps these components as nominal values in separate columns. Such transformations facilitate the integration of structurally different data sets and, in general, provide a better starting point for statistical analysis.

Key for further development is the conceptualization of parameters and methods. So far, we have implemented a basic syntax and rule set that allows defining parameters and methods using the TC. We have thoroughly analysed terms used for parameters and methods. Results indicate that further refinements of syntax and rule set will be needed.

Technical developments must be complemented by work on the metadata content, which is in fact the most time consuming and difficult task because of the huge legacy but also due to the dynamics within science. Over the last 20 years, more than 100 new parameters had to be defined on average per week. During the last years, new terms are increasingly submitted to various terminology services. On one hand, persistent meaning of terms and their identification improves the preservation of the integrity of published data sets. On the other hand, there is a risk that adaptation of old parameter names to the new syntax model might lead to changes in the semantics. Important in this respect will be to improve our current routines to match terms of external terminology services with our parameter or method names. Ultimately, however, expert knowledge will be needed to manually confirm matches.

An additional limiting factor is the quality of terminology services, which varies in content, editorial, interoperability, and sustainability. Good quality terminology services are the building blocks for the conceptualization of parameters and methods. They are, in our view, essential for data interoperability and arguably the most difficult hurdle for the integration of data from different providers. Our future work will therefore have a strong focus on the further development of our models for the definition of parameters and methods.

In summary, the application of terminologies has a mutual positive effect for terminology services and services such as PANGAEA. On both sides, the application of terminologies improves content, reliability and interoperability.

References

Bergman, M., 2014. 50 Ontology Mapping and Alignment Tools. http://www.mkbergman.com/1769/50-ontology-mapping-and-alignment-tools/.

Bojinski, S., Verstraete, M., Peterson, T.C., Richter, C., Simmons, A., Zemp, M., 2014. The concept of essential climate variables in support of climate research, applications, and policy. Bull. Am. Meteorol. Soc. 95 (9), 1431–1443. http://dx.doi.org/10.1175/BAMS-D-13-00047.1.

CCSDS, 2012. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2. The Consultative Committee for Space Data Systems, Washington, DC, USA. https://public.ccsds.org/Pubs/650x0m2.pdf.

Cruz, I., Xiao, H., 2005. The role of ontologies in data integration. Eng. Intell. Syst. 13 (4), 245–252. http://www.cs.uic.edu/~advis/publications/dataint/eis05j.pdf.

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., Fiorani, F., Frohmberg, W., Junker, A., Klukas, C., Lange, M., Mazurek, C., Nafissi, A., Neveu, P., van Oeveren, J., Pommier, C., Poorter, H., Rocca-Serra, P., Sansone, S.-A., Scholz, U., van Schriek, M., Seren, Ü., Usadel, B., Weise, S., Kersey, P., Krajewski, P., 2016. Measures for interoperability of phenotypic data: minimum information requirements and formatting. Plant Methods 12 (1), 44. http://dx.doi.org/10.1186/s13007-016-0144-4.

del Mar Roldán García, M., García-Nieto, J., Aldana-Montes, J.F., 2016. An ontology-based data integration approach for web analytics in e-commerce. Expert Syst. Appl. 63, 20–34. http://dx.doi.org/10.1016/j.eswa.2016.06.034.

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., Wefer, G., 2002. PANGAEA – an information system for environmental sciences. Comput. Geosci. 28 (10), 1201–1210. http://dx.doi.org/10.1016/S0098-3004(02)00039-0.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E., 2011. Semantically enhanced information retrieval: an ontology-based approach. Web Semant. Sci. Serv. Agents World Wide Web 9 (4), 434–452. http://dx.doi.org/10.1016/j.websem.2010.11.003. jWS special issue on Semantic Search.

Findlay, H.S., Kendall, M.A., Spicer, J.I., Widdicombe, S., 2010. Seawater carbonate chemistry and biological processes during experiments with postlarvae of barnacle of *Semibalanus balanoides* and *Elminius modestus*. Mar. Biol. 157 (4), 725–735. http://dx.doi.org/10.1007/s00227-009-1356-1. Supplement to: Findlay, H.S. et al., 2010: Postlarval development of two intertidal barnacles at elevated CO_2 and temperature. doi:10.1594/PANGAEA.758699.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S., 2012. OWL 2 Web Ontology Language Primer, 2nd ed. Recommendation, W3C doi:10.1594/PANGAEA.758699.

Hodgson, R., Keller, P.J., Hodges, J., Spivak, J., 2014. QUDT – Quantities, Units, Dimensions and Data Types Ontologies. Tech. Rep. TopQuadrant, Inc.. http://qudt.org/.

ISO/TC 211, 2014. Geographic Information – Metadata – Part 1: Fundamentals. Tech. Rep. 19115-1. International Organization for Standardization. https://www.iso.org/standard/53798.html.

ISO/TC 211, 2016. Geographic Information – Methodology for Feature Cataloguing. Tech. Rep. 19110. International Organization for Standardization. https://www.iso.org/standard/57303.html.

Karam, N., Müller-Birn, C., Gleisberg, M., Fichtmüller, D., Tolksdorf, R., Güntsch, A., 2016. A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data. Datenbank-Spektrum 16 (3), 195–205. http://dx.doi.org/10.1007/s13222-016-0231-8.

Kindermann, S., 2013. Data Discovery: Identifying, Searching and Finding Data. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 21–32. http://dx.doi.org/10.1007/978-3-642-37244-5_4.

Kotis, K., Lanzenberger, M., 2008. Ontology matching: current status, dilemmas and future challenges. 2008 International Conference on Complex Intelligent and Software Intensive Systems 924–927 10.1109/CISIS.2008.28.

Mate, S., Köpcke, F., Toddenroth, D., Martin, M., Prokosch, H.-U., Bürkle, T., Ganslandt, T., 2015. Ontology-based data integration between clinical and research systems. PLOS ONE 10 (1), 1–20. http://dx.doi.org/10.1371/journal.pone.0116656.

Musen, M.A., 2015. The ProtÉGÉ Project: a look back and a look forward. AI Matters 1 (4), 4–12. http://dx.doi.org/10.1145/2757001.2757003.

Ow, Y.X., Collier, C.J., Uthicke, S., 2016. Responses of three tropical seagrass species to CO_2 enrichment. Mar. Biol. 162 (5), 1005–1017. http://dx.doi.org/10.1007/s00227-015-2644-6. Supplement to: Ow, YX et al. (2015): Responses of three tropical seagrass species to CO_2 enrichment. doi:10.1594/PANGAEA.859062.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science 339 (6117), 277–278. http://dx.doi.org/10.1126/science.1229931.

Saripalle, R.K., 2008. Current Status of Ontologies in Biomedical and Clinical Informatics. Student Project in Biomedical Informatics (CSE5095). School of Engineering, University of Connecticut. http://www.engr.uconn.edu/~steve/Cse300/saripalle.pdf.

Schreiber, G., Raimond, Y., 2014. RDF 1.1 Primer. Working Group Note, W3C. .. https://www.w3.org/TR/rdf11-primer/.

Wick, M., 2006. GeoNames. http://www.geonames.org.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci. Data 3. http://dx.doi.org/10.1038/sdata.2016.18.