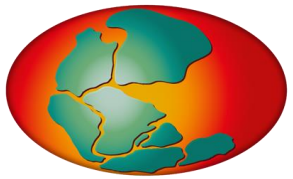
Semantic Annotation of Physical Quantities in the Context of PANGAEA Data Infrastructure

Anusuriya Devaraju
adevaraju@marum.de



PANGAEA.
Data Publisher for Earth & Environmental Science

ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG

marum
Zentrum für Marine Umweltwissenschaften

Universität Bremen

Outline

- Background
  - PANGAEA Data Publisher for Earth & Environmental Science
  - Physical Quantity (aka. Parameter in PANGAEA)
- Motivation
- Automatic Semantic Annotation
  - Text analytics of parameters
  - Inference of quantity kinds through units of measurement
- Conclusions

Outline

- Background
  - PANGAEA Data Publisher for Earth & Environmental Science
  - Physical Quantity (aka. Parameter in PANGAEA)
- Motivation
- Automatic Semantic Annotation
  - Text analytics of parameters
  - Inference of quantity kinds through units of measurement
- Conclusions

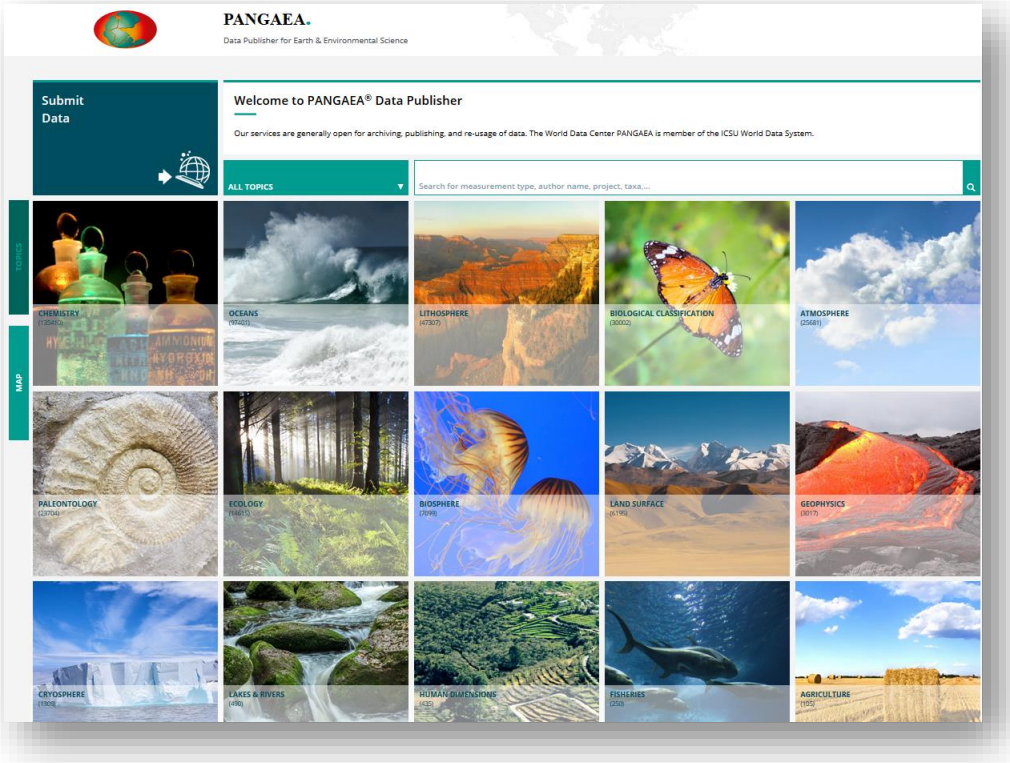
About PANGAEA

- An information system for long-term archiving and publication of Earth & Environmental Science datasets.
- Founded in 1993 and hosted by AWI and MARUM.
- Accreditation and Recognition
  - 2001 - Accredited by the International Council for Science (ICSU) as Publisher for Earth & Environmental Science (ICSU WDS World Data Center).
  - 2007 - Accredited by the World Meteorological Organisation (WMO) as World Radiation Monitoring Center (WRMC).
  - 2011 - Accredited by the WMO as the Data Collection and Processing Center (DCPC).
  - 2013 – Become a data repository for the German Federation for Biological Data (GFBio).
  - 2015: Selected for the German Network for Bioinformatics Infrastructure (de.NBI) Service Center “Biodata” as data resources.

PANGAEA Data

- Data sources, e.g., individual researchers, research projects, data centres and infrastructures.
- Data types, e.g., time series, spatial, images, audio, video.
- Datasets are published with Digital Object Identifiers (DOIs), and are accessible via the PANGAEA portal.
- Web services and APIs are available for advanced interaction, e.g., OAI-PMH metadata harvesting.

377,484 datasets > 13 billions measurements

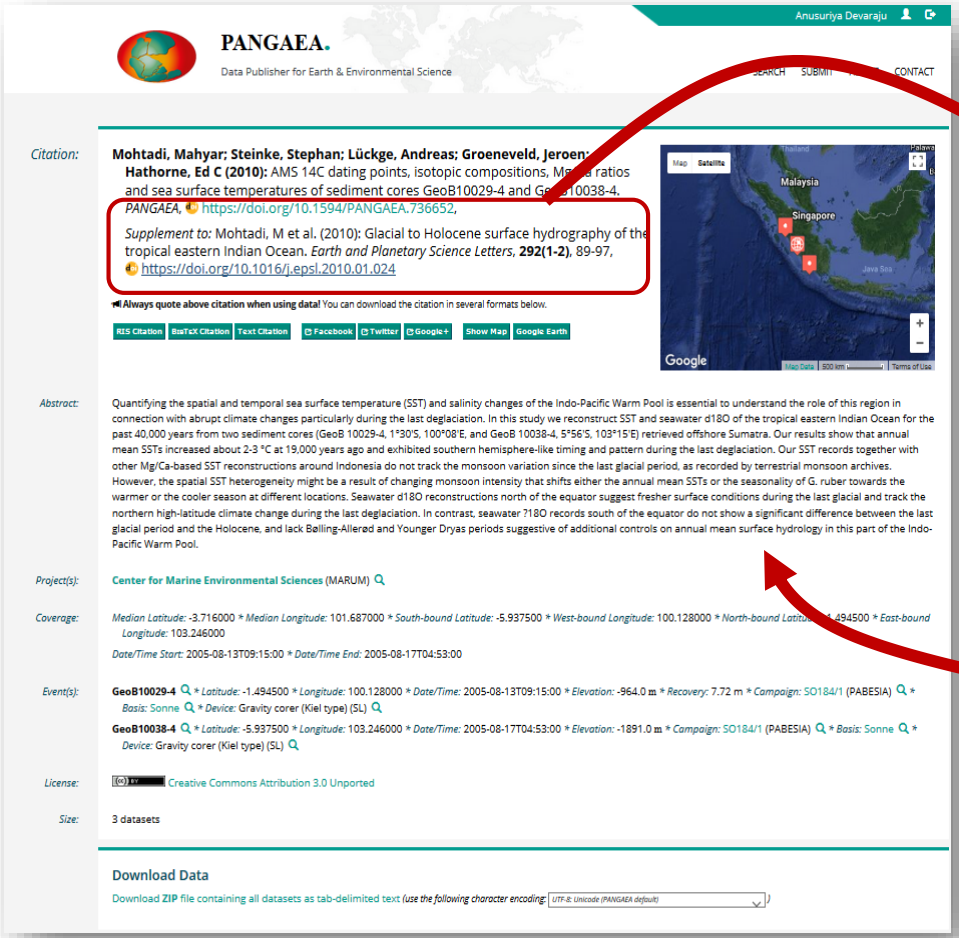


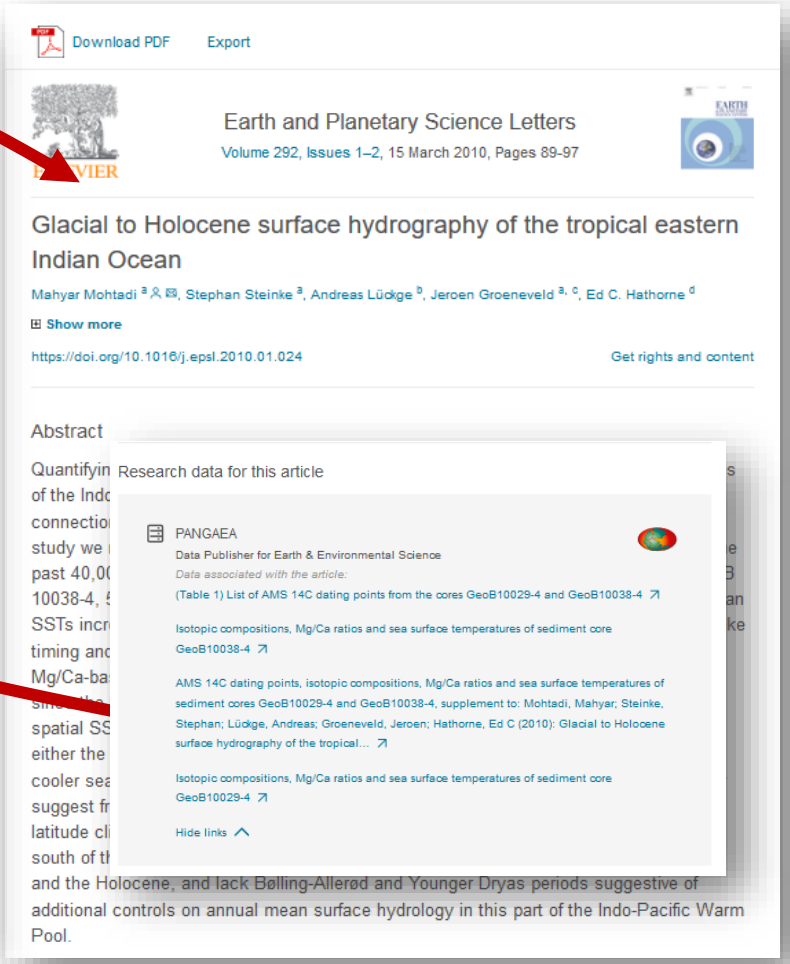
The PANGAEA Data Portal (https://pangaea.de/) enables data search through the ElasticSearch full-text engine.

Cross Referencing

An example of published dataset

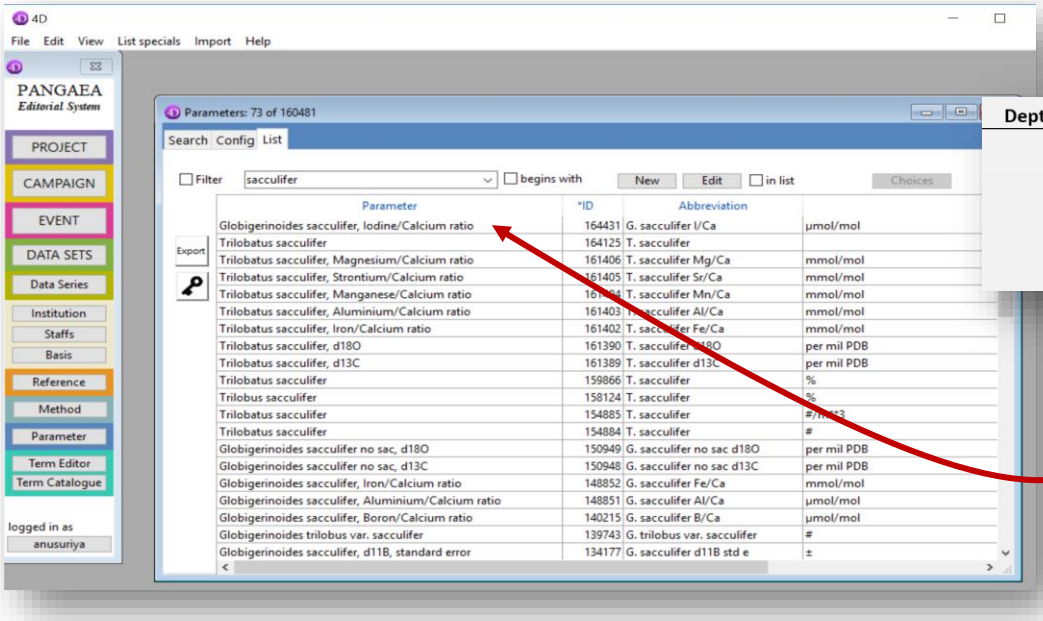
https://doi.org/10.1016/j.epsl.2010.01.024





PANGAEA Data Editorial (4D)

- Data curators import and review datasets submitted by users through the 4D client.
- The review includes transforming user-specified parameters into the relevant PANGAEA parameters.
- A parameter has name, abbreviation and units.



PANGAEA 4D Editorial Interface

An example of user supplied data file

Depth (cm)	*G. sacculifer* (%)	*G. glutinata* (%)	*G. bulloides* (%)
1	4,63	10,65	2,78
5	3,74	1,28	0,77
9	8,68	0,64	0
14	5,59	10,34	7,54

PANGAEA Parameters

- In PANGAEA system, parameters are observed properties/physical quantities.
  - Complex, heterogeneous and are colloquially expressed.
  - A parameter name may consist of one or more 'contexts' such as quantity, features, method/device, aggregates, prepositions, etc.
  - Examples of parameters:
    - Temperature, air, calculated
    - Calanus finmarchicus, egg production rate
    - Accumulation rate, sediment, mean
    - Bacterial biomass production of carbon, standard deviation
    - Oithona hebes, female, length
    - Fugacity of carbon dioxide in seawater

Outline

- Background
  - PANGAEA Data Publisher for Earth & Environmental Science
  - Physical Quantity (aka. Parameter in PANGAEA)
- Motivation
- Automatic Semantic Annotation
  - Text analytics of parameters
  - Inference of quantity kinds through units of measurement
- Conclusions

Motivation

- Semantic annotation of parameters
  - Curators annotate parameters with one or more standard terms (ontological concepts) to improve data indexing and discovery.
  - Added value - integration of data from different providers is facilitated.
- There are 20 ontologies in the PANGAEA system:
  - External ontologies, e.g., WoRMS, ITIS, GCMD keywords, QUDT, CHEBI.
  - In-house terminologies, e.g., Feature ontology, PANGAEA.
- The manual annotation is time-consuming and labour intensive.
- Not feasible with a growing number of parameters (> 160000 parameters at present) and complex ontologies in the PANGAEA data system!

Motivation

- The goal of the study is to largely automate the semantic annotation of parameters.
- The solution comprises
  1. Automatic discovery of the standard terms of the parameters through text analytics.
  2. Inference of quantity kinds through units of measurement of the parameters.

Outline

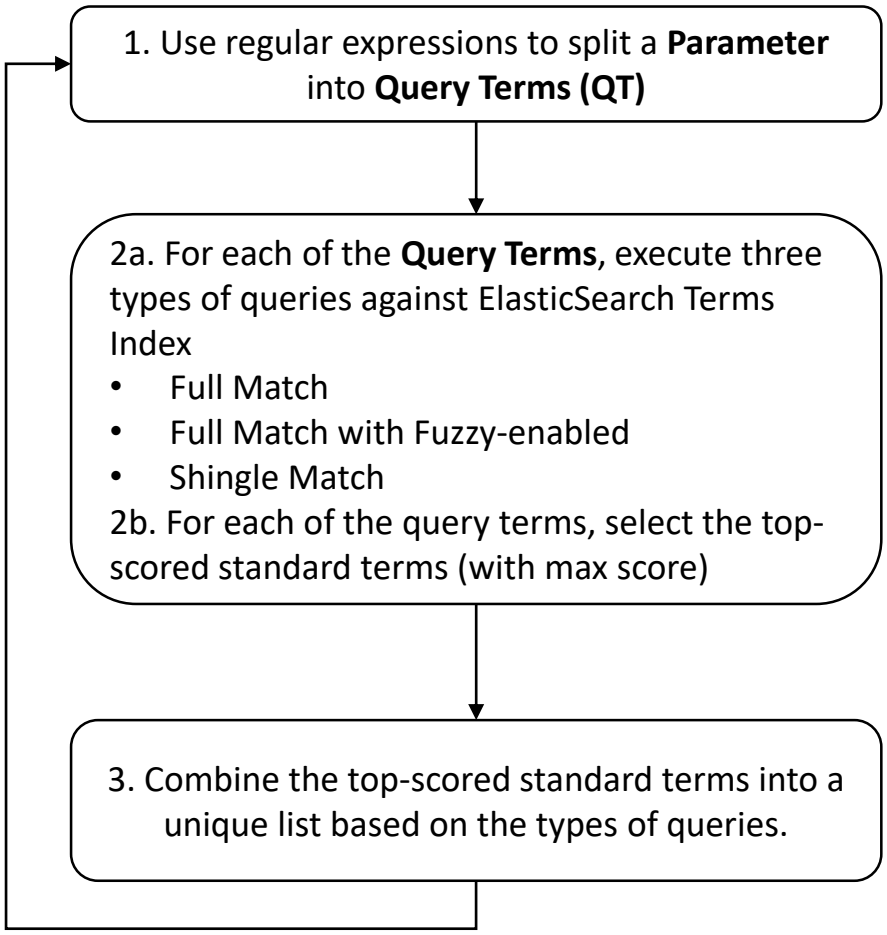
- Background
  - PANGAEA Data Publisher for Earth & Environmental Science
  - Physical Quantity (aka. Parameter in PANGAEA)
- Motivation
- Automatic Semantic Annotation
  - Text analytics of parameters
  - Inference of quantity kinds through units of measurement
- Conclusions

Text Analytics of Parameters

- We indexed all standard terms (ontological concepts) in ElasticSearch.
- We defined and implemented three kinds of ElasticSearch queries.
  - Full Match :
    - Treats a query term as a single unit
    - Supports filters (lowercase, trim, whitespace, ascii)
  - Full Match with Fuzzy-enabled
    - Full Match combined with auto fuzziness
  - Shingle Match
    - Converts a query term into a list of individual terms (based on the Unicode Text Segmentation algorithm) and constructs shingles (n-grams)
    - Supports filters (lowercase, trim, whitespace, ascii) and auto fuzziness

Standard Terms refer to ontological concepts.
Query Terms refer to search string included in ElasticSearch requests.

Text Analytics of Parameters



Repeat the steps (1, 2a-b and 3) for all parameters

EXAMPLE

Parameter:
Chironomidae indeterminata per unit sediment volume
Query Terms (QT):
['Chironomidae indeterminata', 'sediment volume']

	QT1	QT2
Full	[ ]	[ ]
Full Fuzzy	[ ]	[ ]
Shingle	[1034279, 1056978]	[1081536]

[ ] – empty results
1034279, 1056978, 1081536 are standard term ids with max scores, resulted from ElasticSearch query requests.

Parameter:
Chironomidae indeterminata per unit sediment volume
Scored Results:
Full – [ ]
Full Fuzzy – [ ]
Shingle – [1034279, 1056978, 1081536]

Evaluation (Preliminary)

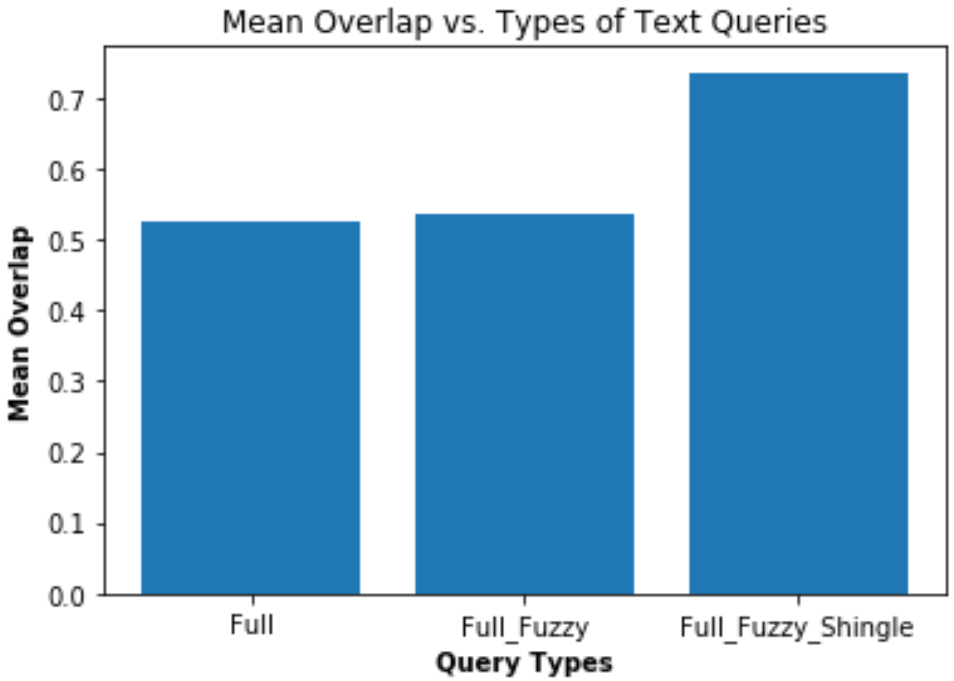
- Compare the retrieved standard terms with the relevant terms (standard terms annotated by curators).
- Test data: 97251 parameters

Parameter	Truth	FullMatch	FullFuzzyMatch	ShingleMatch	FullFuzzyShingleMatch
10:2 fluorotelomer alcohol	[40041]	[40041]	[40041]	[40041]	[40041]
10-Methyldodecanoic acid, particulate	[37962, 38338]	[38338, 37962]	[38338, 37962]	[38338, 37962]	[38338, 37962]
10-methyl-Hexadecanoic acid	[37985]	[]	[37985]	[37983]	[37985, 37983]
10-methyl-Hexadecanoic acid, d13C	[37985, 1073091]	[1073091]	[37985, 1073091]	[1073091, 37983]	[37985, 1073091, 37983]
10-methyl-Hexadecanoic acid of total fatty aci...	[37904, 37985, 38253]	[]	[37985]	[37985, 37983]	[37985, 37983]

Evaluation (Preliminary)

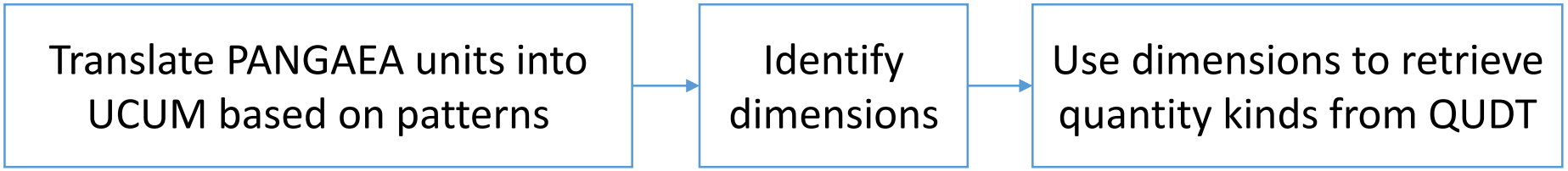
- Overlap refers to the fraction of correct predictions over the total number of a parameter's test data.

$$\text{overlap} = \frac{|\{\text{relevant terms}\} \cap \{\text{retrieved terms}\}|}{|\{\text{relevant terms}\}|}$$



Inference of Quantity Kinds

- A Quantity Kind represents the physical nature or type of a measured quantity, e.g., length, density
- Unified Code for Units of Measure (UCUM) is a code system of units of measures.
- QUDT – a formal specification of Units of Measure, Quantity Kinds, Dimensions and Data Types.
- Leverage units of measurement of the parameters to identify quantity kinds associated with parameters.



Inference of Quantity Kinds

- Total parameters with units: 124618, Total unique units: 1347
- Units translated into UCUM: 1112/1347
- UCUM units with quantity kinds: 941/1112 (84%)

Request: http://seprojects.marum.de:3838/pucum/v1/api/quantity/pg/m**3

Response

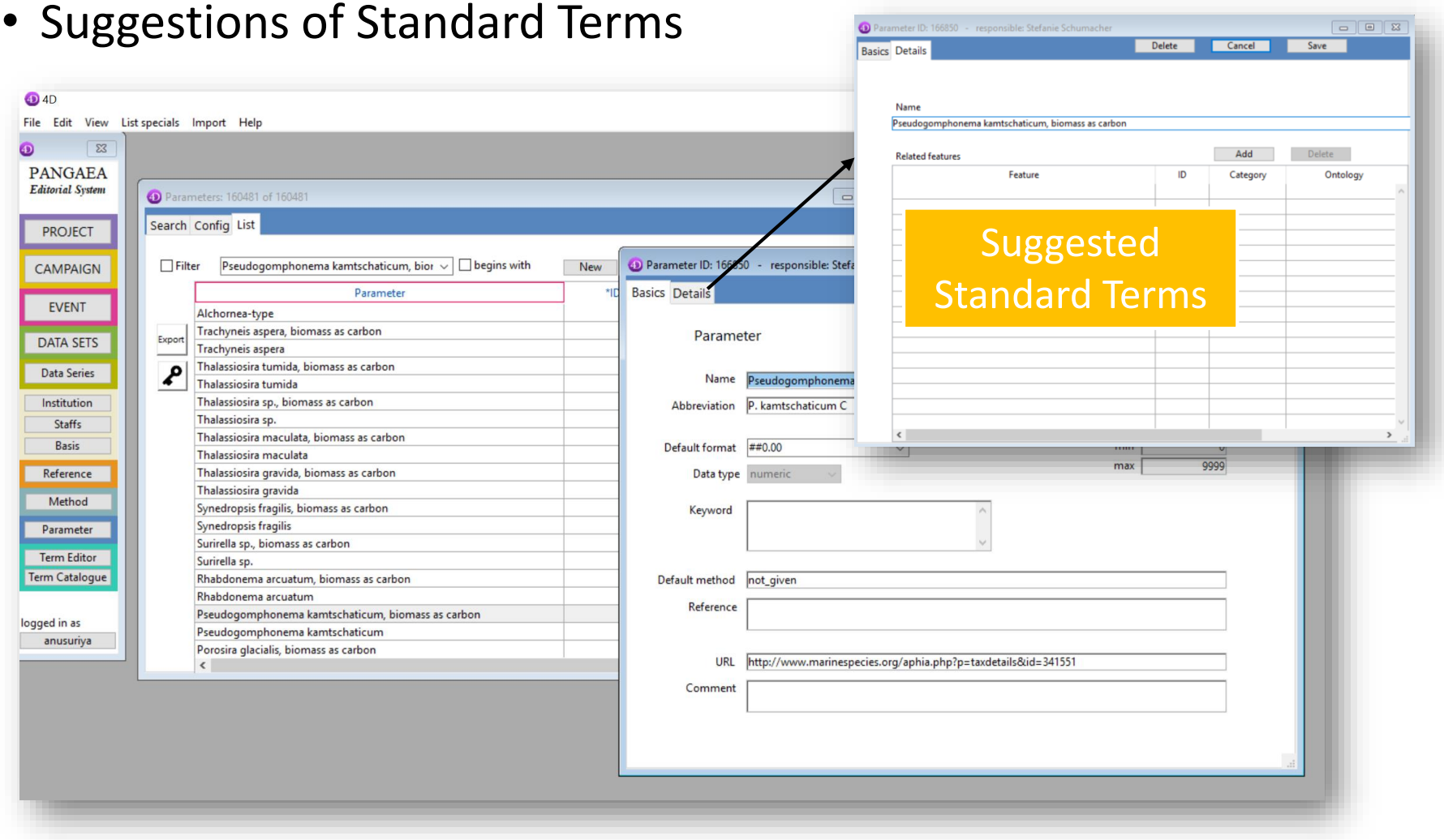

```
input:                  "pg/m**3"
status:                 "201_QUANTITY_FOUND"
status_msg:             "The quantities (UCUM and/or QUDT) are available."
ucum:                   "pg/m3"
fullname:               "(picogram) / (meter ^ 3)"
canonicalunit:          "g.m-3"
dimension:              "M.L-3"
verbosecanonicalunit:   "Mass &middot; Length<sup>-3</sup>"
qudtQuantities:
  0:                    "Density"
```

Outline

- Background
  - PANGAEA Data Publisher for Earth & Environmental Science
  - Physical Quantity (aka. Parameter in PANGAEA)
- Motivation
- Automatic Semantic Annotation
  - Text analytics of parameters
  - Inference of quantity kinds through units of measurement
- Conclusions

Expected Output

- Suggestions of Standard Terms

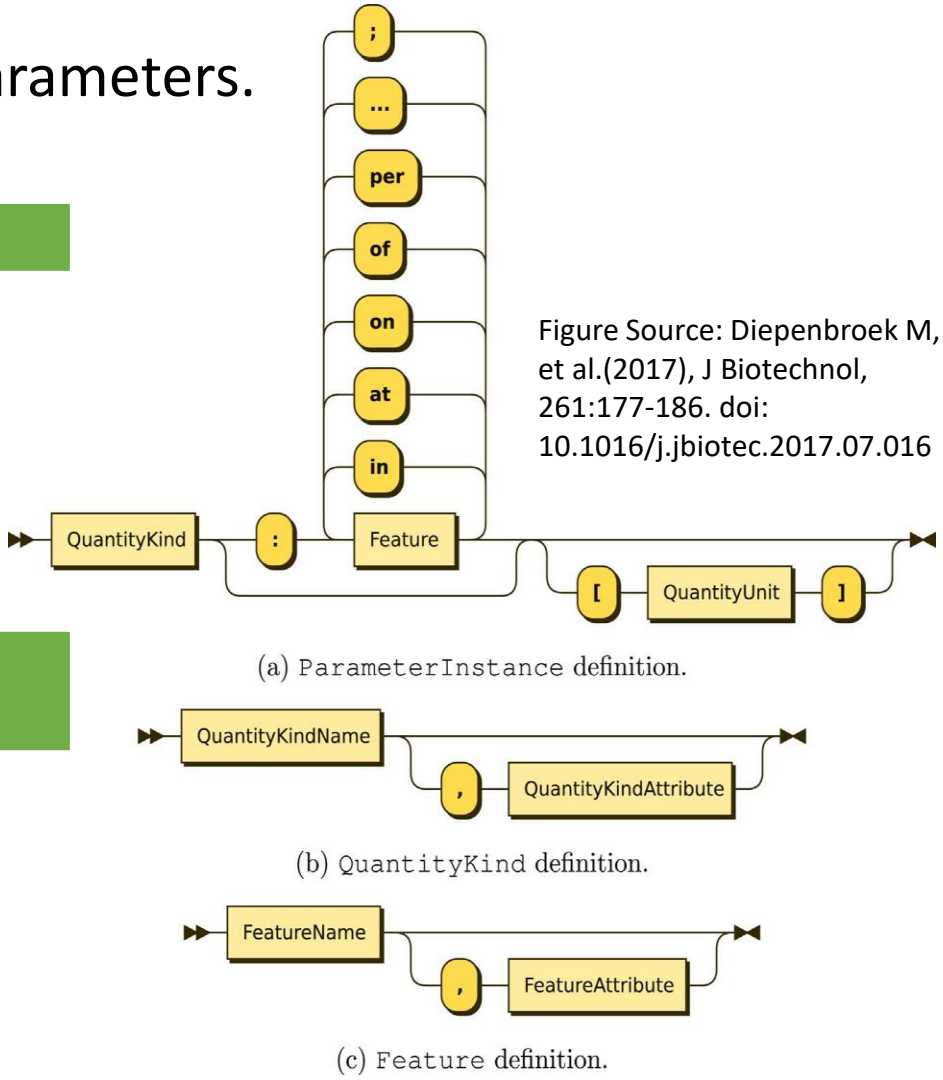


Expected Output

- Model for the definition of parameters.

UNSTRUCTURED PARAMETERS
Calanus finmarchicus, fecal pellet volume
Calanus finmarchicus, egg production rate
Calanoides carinatus, d13C
Calanoides carinatus, d15N

STANDARDIZED & MACHINE-UNDERSTANDABLE PARAMETERS
Volume: fecal pellet of Calanus finmarchus
Production rate: egg of Calanus finmarchus
Ratio: d13C in Calanoides carinatus
Ratio: d13N in Calanoides carinatus



(a) `ParameterInstance` definition.

(b) `QuantityKind` definition.

(c) `Feature` definition.

Figure Source: Diepenbroek M, et al.(2017), J Biotechnol, 261:177-186. doi: 10.1016/j.jbiotec.2017.07.016

Conclusions

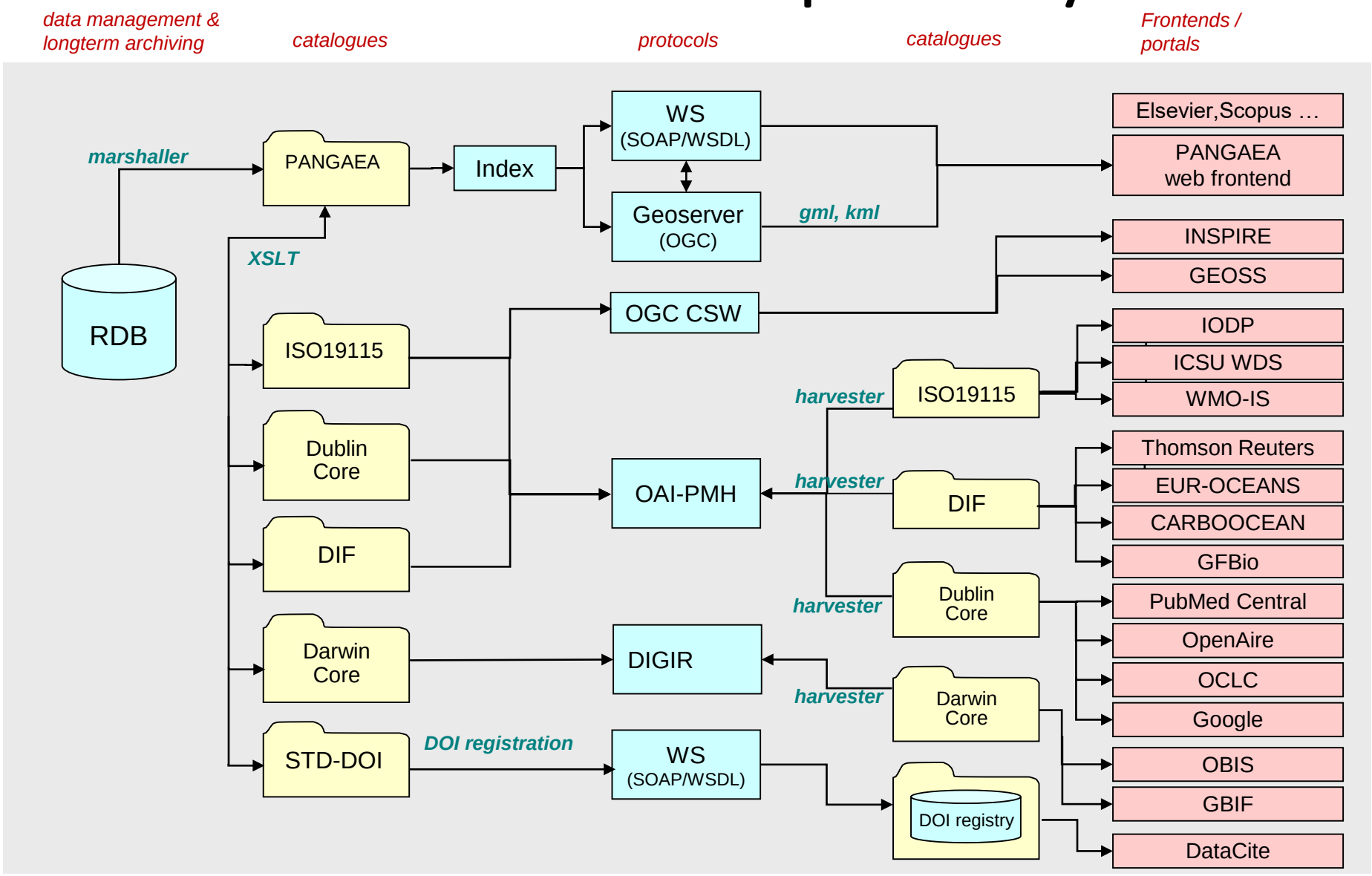
- Developed a solution to support semantic annotation of Earth and Environmental parameters.
  - Text analytics of parameter names
  - Inference of quantity kinds of parameters through units of measurement
- Practical Significance
  - Built upon existing tools in the PANGAEA system, e.g., ElasticSearch, ontologies.
- Future work
  - Improve patterns and term queries
  - Conduct an offline evaluation of unranked results (precision, recall, f-measure)
  - Translate the solution developed into practice.

Acknowledgement

Semantic Enablement of PANGAEA Data
(Team Members)
- Michael Diepenbroek (mdiepenbroek@pangaea.de)
- Robert Huber (rhuber@uni-bremen.de)
- Uwe Schindler (uschindler@pangaea.de)
- Melanie Buß (mbuss@marum.de)
- Janine Felden (jfelden@marum.de)
- Andree Behnken (abehnken@marum.de)

Dissemination of Data & Metadata

PANGAEA Interoperability



Data Publication

- The PANGAEA data editorial ensures the integrity and authenticity as well as a high usability of your data.