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

### Anusuriya Devaraju

adevaraju@marum.de





Data Publisher for Earth & Environmental Science









2

- 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



- 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

PANGAEA.



The PANGAEA Data Portal (<u>https://pangaea.de/</u>) 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.	Download PDF Export
ation:	Data Publisher for Earth & Environmental Science CONTACT Mohtadi, Mahyar; Steinke, Stephan; Lückge, Andreas; Groeneveld, Jeroen:	Earth and Planetary Science Letters Volume 292, Issues 1–2, 15 March 2010, Pages 89-97
	Hathorne, Ed C (2010): AMS 14C dating points, isotopic compositions, Marger ratios and sea surface temperatures of sediment cores GeoB10029-4 and Geom038-4. <i>PANGAEA</i> . Chttps://doi.org/10.1594/PANGAEA.736652, Supplement to: Mohtadi, M et al. (2010): Glacial to Holocene surface hydrography of the tropical eastern Indian Ocean. Earth and Planetary Science Letters, <b>292(1-2)</b> , 89-97, Chttps://doi.org/10.1016/j.epsl.2010.01.024 Mawya quote above citation when using datal You can download the citation in several formati below. Ris Catalon EntroCatation TextCatation (Excelose) Entwitter (Geoglet Show Map Geogle Earth Google (Determined Content of Catalon Content of Cat	Glacial to Holocene surface hydrography of the tropical eastern Indian Ocean Mahyar Mohtadi <sup>a</sup> A <sup>SB</sup> , Stephan Steinke <sup>a</sup> , Andreas Lückge <sup>b</sup> , Jeroen Groeneveld <sup>a, c</sup> , Ed C. Hathorne <sup>d</sup> El Show more https://doi.org/10.1016/j.epsl.2010.01.024 Get rights and conte
bstract:	Quantifying the spatial and temporal sea surface temperature (SST) and salinity changes of the Indo-Pacific Warm Pool is essential to understand the role of this region in connection with abrupt climate changes particularly during the last deglaciation. In this study we reconstruct SST and seawater d180 of the tropical eastern Indian Ocean for the past 40,000 years from two sediment cores (GeoB 10029-4, 1305, 100'08; and GeoB 10038-4, 5565, 103'15) (prevised offinore Sumara. Our results show that annual mean SSTs increased about 2-37 cs 11000, years ago and exhibited southern hemisphere-like iming and pattern during the last deglacial contor ST records together with other Mg/Ca-based SST reconstructions around Indonesia do not track the monston variation since the last glacial period. as recorded by terrestrial monstoon archives. However, the spatial SST heterogeneity might be a result of changing monsson intensity that shifts either the annual mean SSTs or the seasonality of G, ruber towards the warmer or the color season at different locations. Seawater d180 reconstructions north of the equator do not show a significant difference between the last glacial period and the Holocene, and lack Belling-Allered and Younger Dryas periods suggestive of additional controls on annual mean strace hydrology in this part of the Indo- Pacific Warm Pool.	Abstract Quantifyin Research data for this article of the Indc connectio study we   Data Publisher for Earth & Environmental Science Data Publisher for Earth & Environmental Science e
oject(s):	Center for Marine Environmental Sciences (MARUM) Q	10038-4, { (Table 1) List of AMS 14C dating points from the cores GeoB10029-4 and GeoB10038-4 7 an
verage:	Median Latitude: -3.716000 * Median Langitude: 101.687000 * South-bound Latitude: -5.937500 * West-bound Langitude: 100.128000 * North-bound Latitus: 1.494500 * East-bound Langitude: 103.246000	SSTs incrine Isotopic compositions, Mg/Ca ratios and sea surface temperatures of sediment core KE timing and GeoB10038-4 71
rent(s):	Date/Time Stort: 2005-08-13T09:15:00 * Date/Time End: 2005-08-17T04:53:00  GeoB10029-4 Q * Latitude: -1.4,94500 * Longitude: 100.128000 * Date/Time: 2005-08-13T09:15:00 * Elevation: -964.0 m * Recovery: 7.72 m * Compaign: SO184/1 (PABESIA) Q * Basis: Sonne Q * Device: Gravity corer (Niel type) (SL) Q  GeoB10038-4 Q * Latitude: -5.937500 * Longitude: 103.246000 * Date/Time: 2005-08-13T04:53:00 * Elevation: -1891.0 m * Compaign: SO184/1 (PABESIA) Q * Basis: Sonne Q * Device: Gravity corer (Niel type) (SL) Q  device: Gravity corer (Niel type) (SL) Q	Mg/Ca-ba         AMS 14C dating points, isotopic compositions, Mg/Ca ratios and sea surface temperatures of sediment cores GeoB10029-4 and GeoB10038-4, supplement to: Mohtadi, Mahyar, Steinke, spatial SS           Stephan; Lüdige, Andreas; Groeneveld, Jeroen; Hathorne, Ed C (2010): Glacial to Holocene surface hydrography of the tropical 7
cense:	(e)	COOLER SE2 Isotopic compositions, Mg/Ca ratios and sea surface temperatures of sediment core Surgest fr GeoB10029-4 7
Size:	3 datasets	latitude cli Hide links A
	Download Data Download ZIP file containing all datasets as tab-delimited text (use the following character encoding [r.#. Unioder (MARAM default)])	and the Holocene, and lack Bølling-Allerød and Younger Dryas periods suggestive of additional controls on annual mean surface hydrology in this part of the Indo-Pacific Warm Pool



# 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.

GAEA							
l System	Parameters: 73 of 160481			Depth (cm)	G. <u>sacculifer</u> (%)	G. glutinata (%)	G. bulloides (
ECT	Search Config List						
				1	4,63	10,65	2,78
PAIGN	Filter sacculifer v be	egins with New Edit in	list Choices	-		1.20	0.77
	Parameter	*ID Abbreviation		5	5,74	1,28	0,77
NT	Globigerinoides sacculifer, lodine/Calcium ratio	164431 G. sacculifer I/Ca	µmol/mol				
	Trilobatus sacculifer	164125 T. sacculifer		9	8,68	0,64	0
SETS	Trilobatus sacculifer, Magnesium/Calcium ratio	161406 T. sacculifer Mg/Ca	mmol/mol				
ariar	Trilobatus sacculifer, Strontium/Calcium ratio	161405 T. sacculifer Sr/Ca	mmol/mol	14	5,59	10,34	7,54
Series	<ul> <li>Trilobatus sacculifer, Manganese/Calcium ratio</li> </ul>	161-04 T. sacculifer Mn/Ca	mmol/mol				
tion	Trilobatus sacculifer, Aluminium/Calcium ratio	161403 1. acculifer Al/Ca	mmol/mol				
affs	Trilobatus sacculifer, Iron/Calcium ratio	161402 T. sacc. Efer Fe/Ca	mmol/mol				
	Trilobatus sacculifer, d180	161390 T. sacculifer 1180	per mil PDB				
215	Trilobatus sacculifer, d13C	161389 T. sacculifer d13C	per mil PDB				
rence	Trilobatus sacculifer	159866 T. sacculifer	%				
	Trilobus sacculifer	158124 T. sacculifer	96				
hod	Trilobatus sacculifer	154885 T. sacculifer	#/11/**3				
eter	Trilobatus sacculifer	154884 T. sacculifer					
	Globigerinoides sacculifer no sac, d180	150949 G. sacculifer no sac d18O	per mil PDB				
Editor	Globigerinoides sacculifer no sac, d13C	150948 G. sacculifer no sac d13C	per mil PDB				
talogue	Globigerinoides sacculifer, Iron/Calcium ratio	148852 G. sacculifer Fe/Ca	mmol/mol				
	Globigerinoides sacculifer, Aluminium/Calcium ratio	148851 G. sacculifer Al/Ca	µmol/mol				
a as	Globigerinoides sacculifer, Boron/Calcium ratio	140215 G. sacculifer B/Ca	µmol/mol				
suriva	Globigerinoides trilobus var. sacculifer	139743 G. trilobus var. sacculifer	#				
Juliya	Globigerinoides sacculifer, d11B, standard error	134177 G. sacculifer d11B std e	±	~			
	<			in <			



## **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



- 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.



- 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



13

PANGAEA.

# 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)**:

PANGAEA.

['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]



PANGAEA.

# **Evaluation (Preliminary)**

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

		•	•		
FullFuzzyShingleMatch	ShingleMatch	FullFuzzyMatch	FullMatch	Truth	Parameter
[40041]	[40041]	[40041]	[40041]	[40041]	10:2 fluorotelomer alcohol
[38338, 37962]	[38338, 37962]	[38338, 37962]	[38338, 37962]	[37962, 38338]	10-Methyldodecanoic acid, particulate
[37985, 37983]	[37983]	[37985]	Π	[37985]	10-methyl-Hexadecanoic acid
[37985, 1073091, 37983]	[1073091, 37983]	[37985, 1073091]	[1073091]	[37985, 1073091]	10-methyl-Hexadecanoic acid, d13C
[37985, 37983]	[37985, 37983]	[37985]	0	[37904, 37985, 38253]	10-methyl-Hexadecanoic acid of total fatty aci



# **Evaluation (Preliminary)**

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





# 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"
<pre>status_msg:</pre>	"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 · Length <sup>-3</sup> "
▼qudtQuantities:	
0:	"Density"



- 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

PANGAEA. Data Publisher for Earth & Environmental Science

### **Expected Output**

D						Name			
Edit View Li	st specials In	nport Help				Pseudogomphonema kamtschaticum, bioma	ass as carbon		
23						Delated for the second		Add	Delete
NGAFA					7	Related features	n	Category	Ontology
torial System	( Parama	eters: 160481 of 160481				. cotore		concycry	childigy
	Parame								
PROJECT	Search C	Config List				Sugar	actod		
				O Damaster ID 155		Sugge	esteu		
AMPAIGN	Filter	Pseudogomphonema kamtschaticum, bior V Degins with	New	Parameter ID: 1005	50 - responsible: Stera		1.77		
EVENT		Parameter	*10	Basics Details		Standar	d lerms		
EVENI		Alchornea-type Trachyneir yspera, hiomarc ar carbon				-			
DATA SETS	Export	Trachyneis aspera, biomass as carbon Trachyneis aspera		Parame	eter				
Data Series	٥	Thalassiosira tumida, biomass as carbon							
ou series	<b>•</b>	Thalassiosira tumida		Name	Pseudogomphonema				
Institution		Thalassiosira sp., biomass as carbon		Abbreviation	P. kamtschaticum C				
Staffs		i naiassiosira sp. Thalassiosira maculata, biomass as carbon				<			1
Basis	i i	Thalassiosira maculata		Default format	##0.00	~		0	
Reference	1	Thalassiosira gravida, biomass as carbon		Data type	numeric 🗸		max 9	1999	-
Method		Thalassiosira gravida							
	9	Synedropsis fragilis, biomass as carbon		Keyword		^			
Parameter		Surirella sp., biomass as carbon				~			
Ferm Editor	9	Surirella sp.							
rm Catalogue	F	Rhabdonema arcuatum, biomass as carbon		Default method	not_given				
	F	Rhabdonema arcuatum		Reference					
ed in as	E F	Pseudogomphonema kamischaticum, biomass as carbon Pseudogomphonema kamischaticum							
anusuriya	F	Porosira glacialis, biomass as carbon							
		<		URL	http://www.marinespec	ies.org/aphia.php?p=taxdetails&id=341551			
				Comment					





21

Universität Bremen

PANGAEA.



#### Universität Bremen

### 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, fmeasure)
  - Translate the solution developed into practice.



PANGAEA.

# 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)





### **PANGAEA** Interoperability



### Data Publication

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



PANGAEA.