Geospatial Sensing Conference 2019
From Sensing to Understanding our World
September 2 – 4, 2019 Münster, Germany

Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining

Anusuriya Devaraju, Uwe Schindler, Robert Huber, Michael Diepenbroek
{adevaraju, uschindler, rhuber, mdiepenbroek}@marum.de
PANGAEA Data Publisher for Earth & Environmental Science,
MARUM - Center for Marine Environmental Sciences, University of Bremen, Germany.

PANGAEA.
Data Publisher for Earth & Environmental Science

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

marum
Zentrum für Marine
Umweltwissenschaften

Universität Bremen

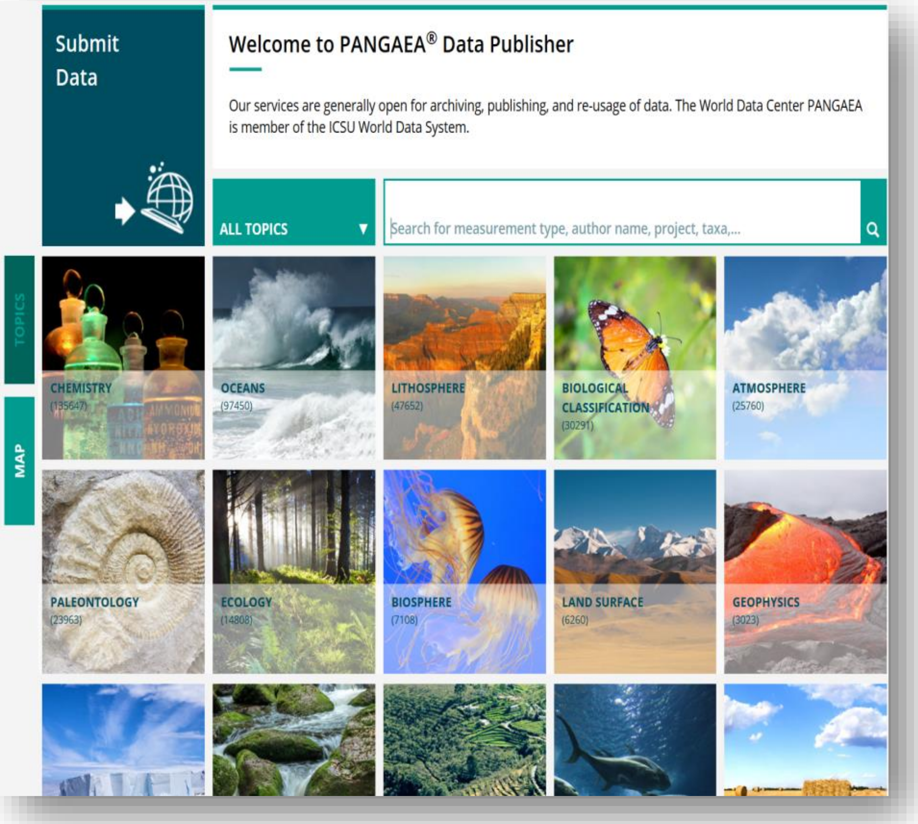
Presentation Outline

- PANGAEA Digital Data Repository
- Data Recommender System
  a. Metadata-based data recommendation
  b. Users interaction-based data recommendation
- Online Evaluation & Results
- Conclusions

PANGAEA Digital Data Repository

- Founded in 1993
- Jointly managed by AWI and MARUM.
- Datasets from researchers, projects, research centres and infrastructures (national & international) published with DOIs
- Data types, e.g., time series, spatial, images, audio, video.

> 386000 scientific datasets (30.08.2019)



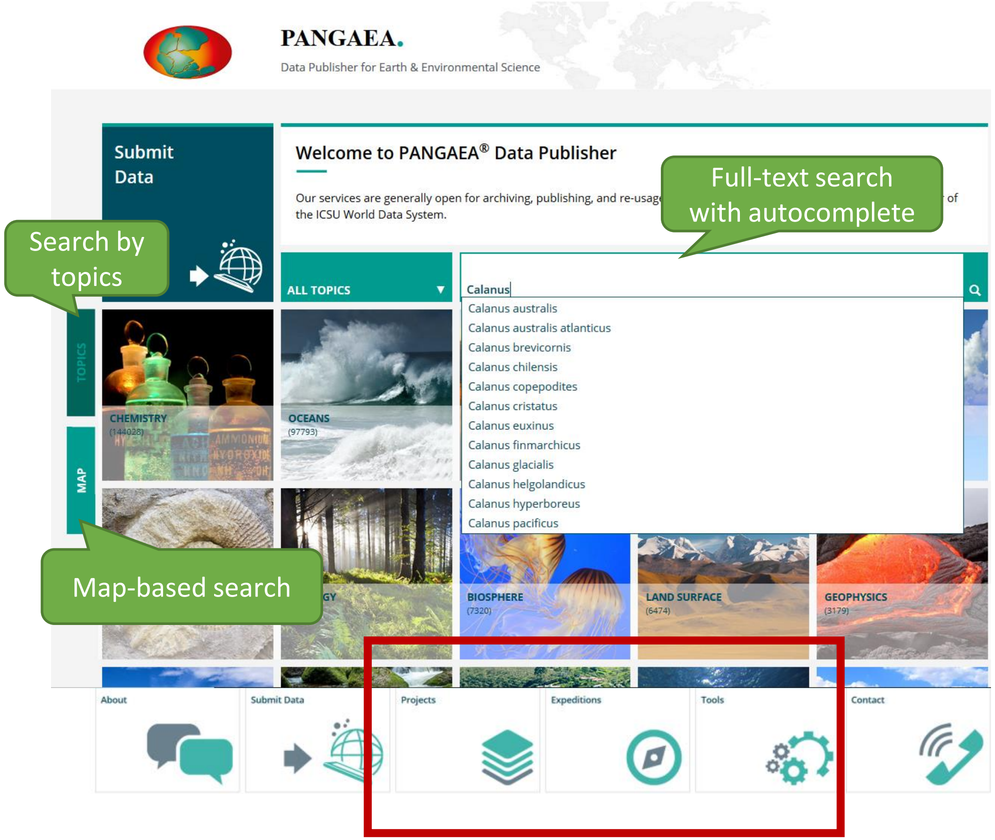
Welcome to PANGAEA® Data Publisher

Our services are generally open for archiving, publishing, and re-usage of data. The World Data Center PANGAEA is member of the ICSU World Data System.

PANGAEA Data Portal (https://pangaea.de/)

Data Search in PANGAEA

- PANGAEA offers tools/APIs for meta(data) access and discovery.



PANGAEA.
Data Publisher for Earth & Environmental Science

Submit Data

Welcome to PANGAEA® Data Publisher

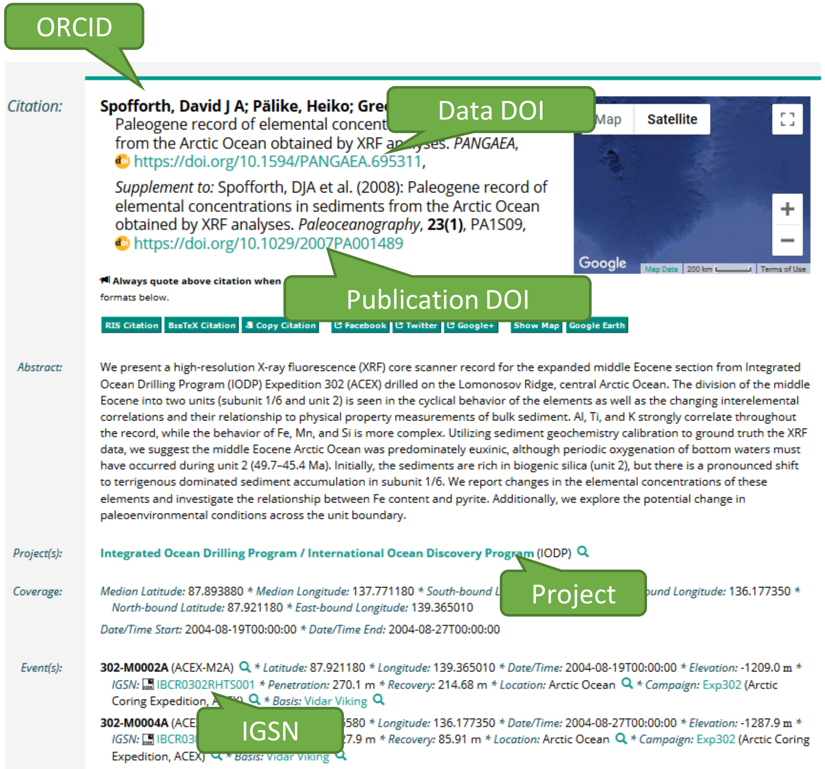
Our services are generally open for archiving, publishing, and re-usag … of the ICSU World Data System.

ALL TOPICS ▾

Calanus

Calanus australis
Calanus australis atlanticus
Calanus brevicornis
Calanus chilensis
Calanus copepodites
Calanus cristatus
Calanus euxinus
Calanus finmarchicus
Calanus glacialis
Calanus helgolandicus
Calanus hyperboreus
Calanus pacificus

Search functions in PANGAEA



Citation: **Spofforth, David J A; Pälike, Heiko; Gre…** Paleogene record of elemental concent… from the Arctic Ocean obtained by XRF an…yses. *PANGAEA*, https://doi.org/10.1594/PANGAEA.695311,

Supplement to: Spofforth, DJA et al. (2008): Paleogene record of elemental concentrations in sediments from the Arctic Ocean obtained by XRF analyses. *Paleoceanography*, **23(1)**, PA1S09, https://doi.org/10.1029/2007PA001489

Always quote above citation when … formats below.

RIS Citation | BibTeX Citation | Copy Citation | Facebook | Twitter | Google+ | Show Map | Google Earth

Abstract: We present a high-resolution X-ray fluorescence (XRF) core scanner record for the expanded middle Eocene section from Integrated Ocean Drilling Program (IODP) Expedition 302 (ACEX) drilled on the Lomonosov Ridge, central Arctic Ocean. The division of the middle Eocene into two units (subunit 1/6 and unit 2) is seen in the cyclical behavior of the elements as well as the changing interelemental correlations and their relationship to physical property measurements of bulk sediment. Al, Ti, and K strongly correlate throughout the record, while the behavior of Fe, Mn, and Si is more complex. Utilizing sediment geochemistry calibration to ground truth the XRF data, we suggest the middle Eocene Arctic Ocean was predominately euxinic, although periodic oxygenation of bottom waters must have occurred during unit 2 (49.7–45.4 Ma). Initially, the sediments are rich in biogenic silica (unit 2), but there is a pronounced shift to terrigenous dominated sediment accumulation in subunit 1/6. We report changes in the elemental concentrations of these elements and investigate the relationship between Fe content and pyrite. Additionally, we explore the potential change in paleoenvironmental conditions across the unit boundary.

Project(s): **Integrated Ocean Drilling Program / International Ocean Discovery Program** (IODP)

Coverage: *Median Latitude:* 87.893880 * *Median Longitude:* 137.771180 * *South-bound L…* …*und Longitude:* 136.177350 * *North-bound Latitude:* 87.921180 * *East-bound Longitude:* 139.365010

Date/Time Start: 2004-08-19T00:00:00 * *Date/Time End:* 2004-08-27T00:00:00

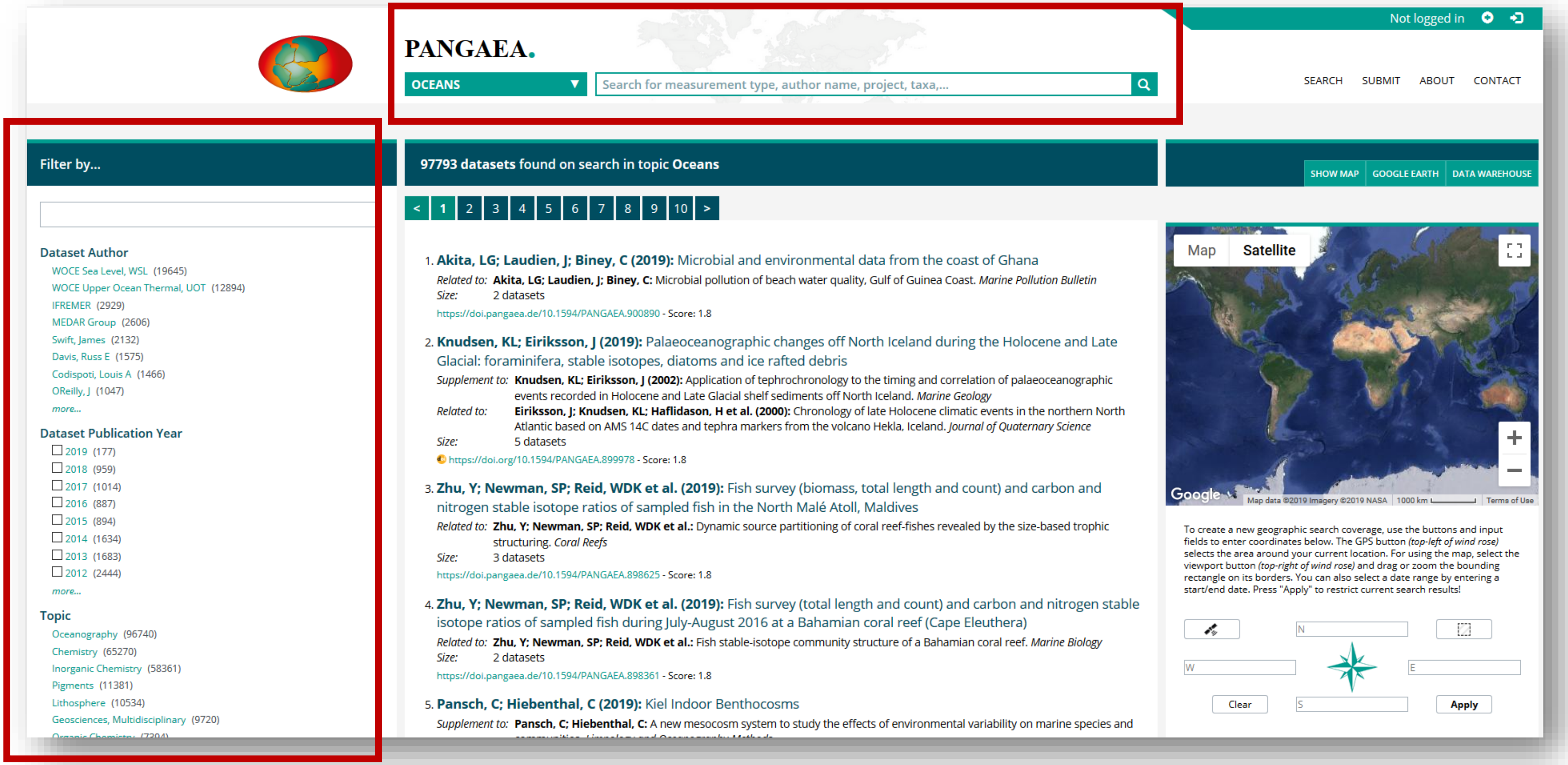
Event(s): **302-M0002A** (ACEX-M2A) * *Latitude:* 87.921180 * *Longitude:* 139.365010 * *Date/Time:* 2004-08-19T00:00:00 * *Elevation:* -1209.0 m * *IGSN:* IBCR0302RHTS001 * *Penetration:* 270.1 m * *Recovery:* 214.68 m * *Location:* Arctic Ocean * *Campaign:* Exp302 (Arctic Coring Expedition, A…) * *Basis:* Vidar Viking

302-M0004A (ACE… …580 * *Longitude:* 136.177350 * *Date/Time:* 2004-08-27T00:00:00 * *Elevation:* -1287.9 m * *IGSN:* IBCR03… …7.9 m * *Recovery:* 85.91 m * *Location:* Arctic Ocean * *Campaign:* Exp302 (Arctic Coring Expedition, ACEX) * *Basis:* Vidar Viking

Dataset and its related research objects

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

Data Search in PANGAEA



Filter by...

Dataset Author

- WOCE Sea Level, WSL (19645)
- WOCE Upper Ocean Thermal, UOT (12894)
- IFREMER (2929)
- MEDAR Group (2606)
- Swift, James (2132)
- Davis, Russ E (1575)
- Codispoti, Louis A (1466)
- OReilly, J (1047)
- *more...*

Dataset Publication Year

- □ 2019 (177)
- □ 2018 (959)
- □ 2017 (1014)
- □ 2016 (887)
- □ 2015 (894)
- □ 2014 (1634)
- □ 2013 (1683)
- □ 2012 (2444)
- *more...*

Topic

- Oceanography (96740)
- Chemistry (65270)
- Inorganic Chemistry (58361)
- Pigments (11381)
- Lithosphere (10534)
- Geosciences, Multidisciplinary (9720)
- Organic Chemistry (7304)

97793 datasets found on search in topic Oceans

1. **Akita, LG; Laudien, J; Biney, C (2019):** Microbial and environmental data from the coast of Ghana
   Related to: **Akita, LG; Laudien, J; Biney, C:** Microbial pollution of beach water quality, Gulf of Guinea Coast. *Marine Pollution Bulletin*
   Size: 2 datasets
   https://doi.pangaea.de/10.1594/PANGAEA.900890 - Score: 1.8

2. **Knudsen, KL; Eiriksson, J (2019):** Palaeoceanographic changes off North Iceland during the Holocene and Late Glacial: foraminifera, stable isotopes, diatoms and ice rafted debris
   Supplement to: **Knudsen, KL; Eiriksson, J (2002):** Application of tephrochronology to the timing and correlation of palaeoceanographic events recorded in Holocene and Late Glacial shelf sediments off North Iceland. *Marine Geology*
   Related to: **Eiriksson, J; Knudsen, KL; Haflidason, H et al. (2000):** Chronology of late Holocene climatic events in the northern North Atlantic based on AMS 14C dates and tephra markers from the volcano Hekla, Iceland. *Journal of Quaternary Science*
   Size: 5 datasets
   https://doi.org/10.1594/PANGAEA.899978 - Score: 1.8

3. **Zhu, Y; Newman, SP; Reid, WDK et al. (2019):** Fish survey (biomass, total length and count) and carbon and nitrogen stable isotope ratios of sampled fish in the North Malé Atoll, Maldives
   Related to: **Zhu, Y; Newman, SP; Reid, WDK et al.:** Dynamic source partitioning of coral reef-fishes revealed by the size-based trophic structuring. *Coral Reefs*
   Size: 3 datasets
   https://doi.pangaea.de/10.1594/PANGAEA.898625 - Score: 1.8

4. **Zhu, Y; Newman, SP; Reid, WDK et al. (2019):** Fish survey (total length and count) and carbon and nitrogen stable isotope ratios of sampled fish during July-August 2016 at a Bahamian coral reef (Cape Eleuthera)
   Related to: **Zhu, Y; Newman, SP; Reid, WDK et al.:** Fish stable-isotope community structure of a Bahamian coral reef. *Marine Biology*
   Size: 2 datasets
   https://doi.pangaea.de/10.1594/PANGAEA.898361 - Score: 1.8

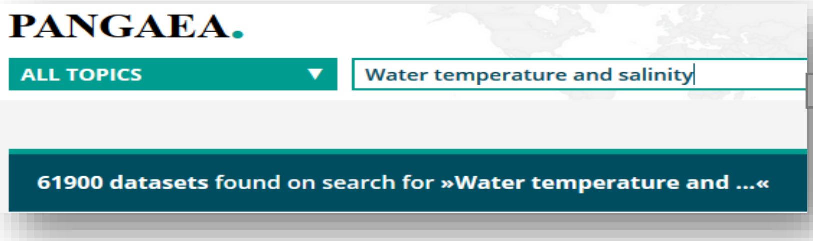
5. **Pansch, C; Hiebenthal, C (2019):** Kiel Indoor Benthocosms
   Supplement to: **Pansch, C; Hiebenthal, C:** A new mesocosm system to study the effects of environmental variability on marine species and

To create a new geographic search coverage, use the buttons and input fields to enter coordinates below. The GPS button *(top-left of wind rose)* selects the area around your current location. For using the map, select the viewport button *(top-right of wind rose)* and drag or zoom the bounding rectangle on its borders. You can also select a date range by entering a start/end date. Press "Apply" to restrict current search results!

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

Search Meets Discovery

- “Search is often struggling to deliver meaningful results, unless you’re very explicit and goal oriented…” (Bibblio, Search vs Discovery, 2015)
- How about users ..
  - may not know what/how to search
  - who are not aware of the range of available datasets
  - if presented with many datasets may not be able to choose the datasets-of-interest.

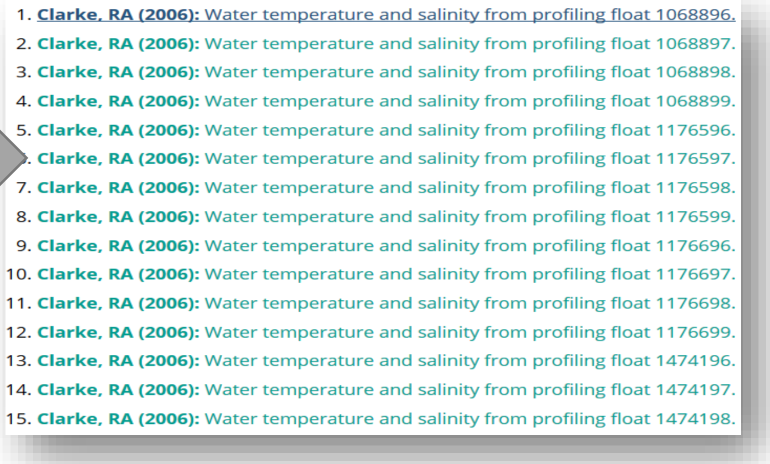


PANGAEA.

Water temperature and salinity

61900 datasets found on search for »Water temperature and ...«

Keyword search brings too many almost identical datasets; diversity is missing!



1. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1068896.
2. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1068897.
3. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1068898.
4. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1068899.
5. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176596.
6. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176597.
7. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176598.
8. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176599.
9. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176696.
10. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176697.
11. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176698.
12. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1176699.
13. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1474196.
14. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1474197.
15. **Clarke, RA (2006):** Water temperature and salinity from profiling float 1474198.

How can users discover relevant and ‘novel’ datasets on the portal?

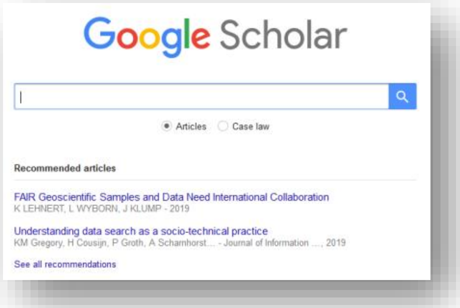
Recommender System

A recommender system is an information filtering system that provide users with personalized contents and services.

Meetup

amazon

Mr. DLib
Machine-readable Digital Library



Google Scholar

● Articles ○ Case law

Recommended articles

FAIR Geoscientific Samples and Data Need International Collaboration
K LEHNERT, L WYBORN, J KLUMP - 2019

Understanding data search as a socio-technical practice
KM Gregory, H Cousijn, P Groth, A Scharnhorst... - Journal of Information ..., 2019

See all recommendations

Quora

coursera

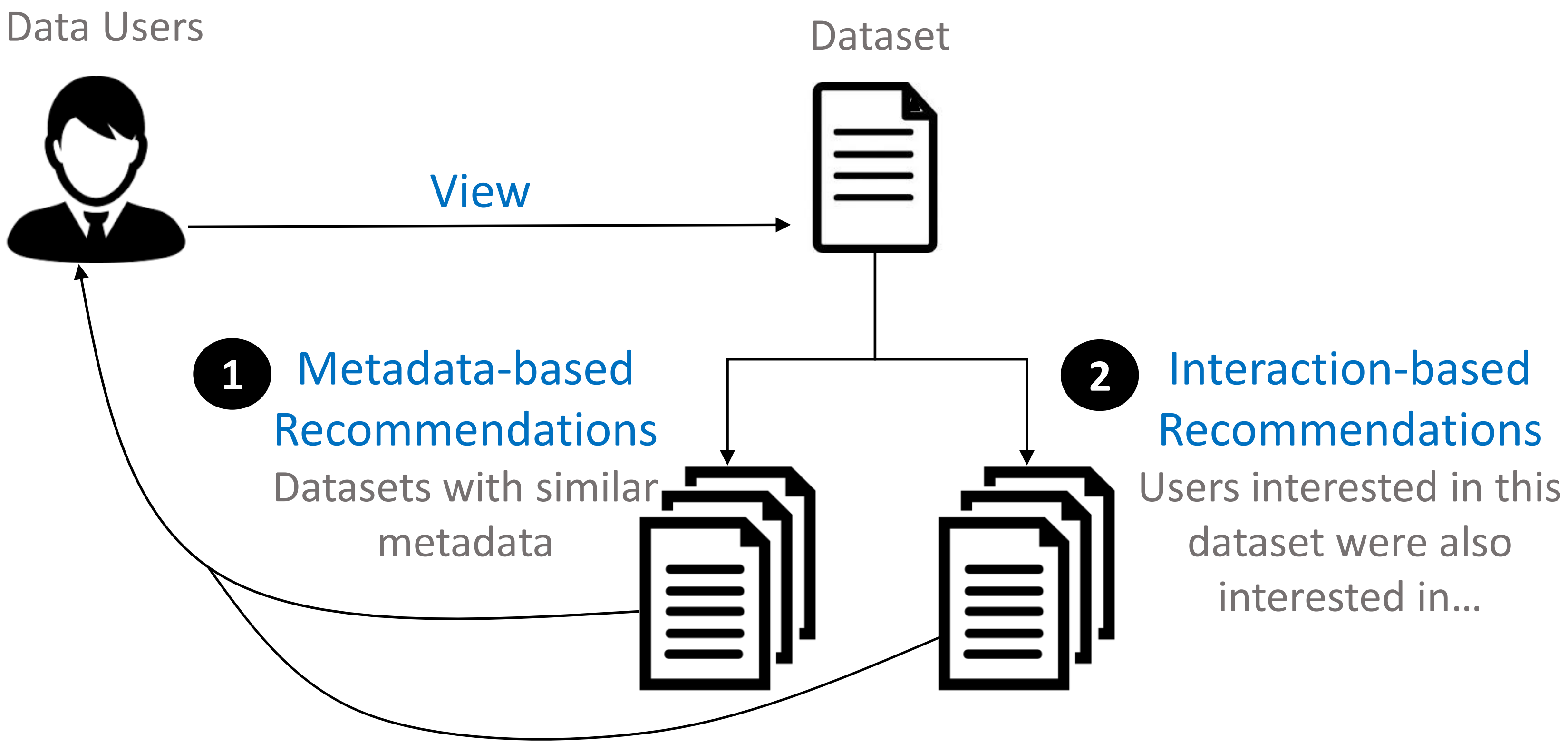
IMDb

LinkedIn

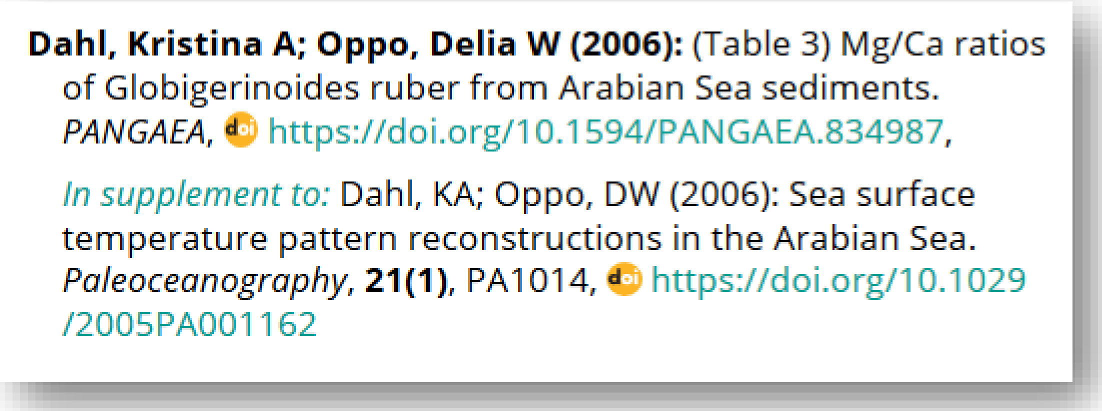
ebay

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

PANGAEA Data Recommender


Data Users
Dataset
View
1
Metadata-based Recommendations
Datasets with similar metadata
2
Interaction-based Recommendations
Users interested in this dataset were also interested in…

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.



Dahl, Kristina A; Oppo, Delia W (2006): (Table 3) Mg/Ca ratios of Globigerinoides ruber from Arabian Sea sediments. *PANGAEA*, https://doi.org/10.1594/PANGAEA.834987,

In supplement to: Dahl, KA; Oppo, DW (2006): Sea surface temperature pattern reconstructions in the Arabian Sea. *Paleoceanography*, **21(1)**, PA1014, https://doi.org/10.1029/2005PA001162

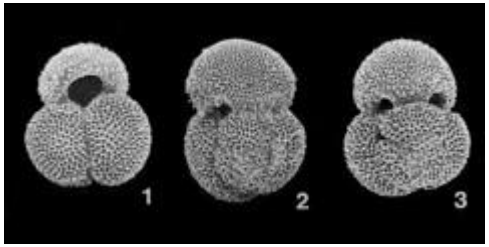
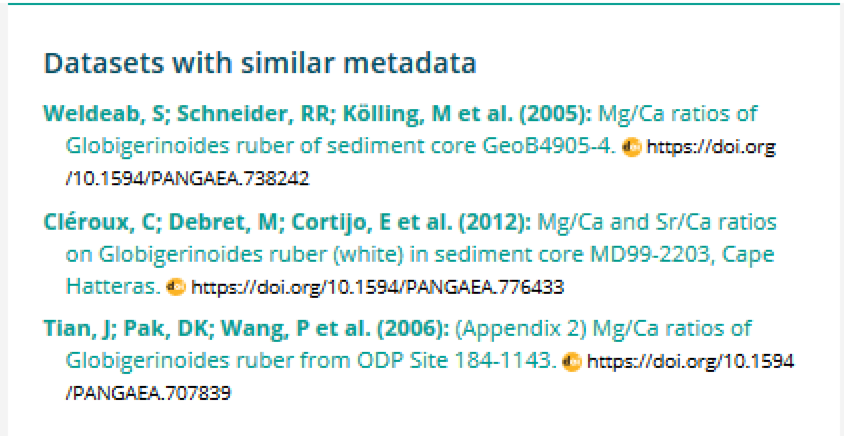


Image : World Register of Marine Species

More diverse recommendations.

Metadata-based Recommendation



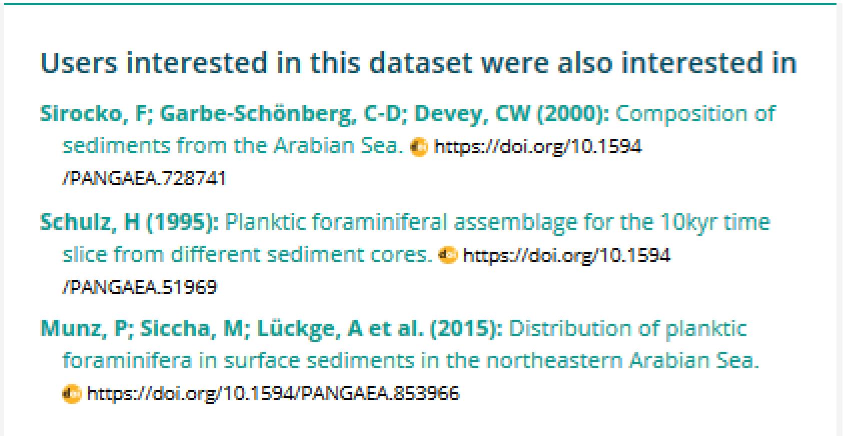
Datasets with similar metadata

Weldeab, S; Schneider, RR; Kölling, M et al. (2005): Mg/Ca ratios of Globigerinoides ruber of sediment core GeoB4905-4. https://doi.org/10.1594/PANGAEA.738242

Cléroux, C; Debret, M; Cortijo, E et al. (2012): Mg/Ca and Sr/Ca ratios on Globigerinoides ruber (white) in sediment core MD99-2203, Cape Hatteras. https://doi.org/10.1594/PANGAEA.776433

Tian, J; Pak, DK; Wang, P et al. (2006): (Appendix 2) Mg/Ca ratios of Globigerinoides ruber from ODP Site 184-1143. https://doi.org/10.1594/PANGAEA.707839

Interaction-based Recommendation



Users interested in this dataset were also interested in

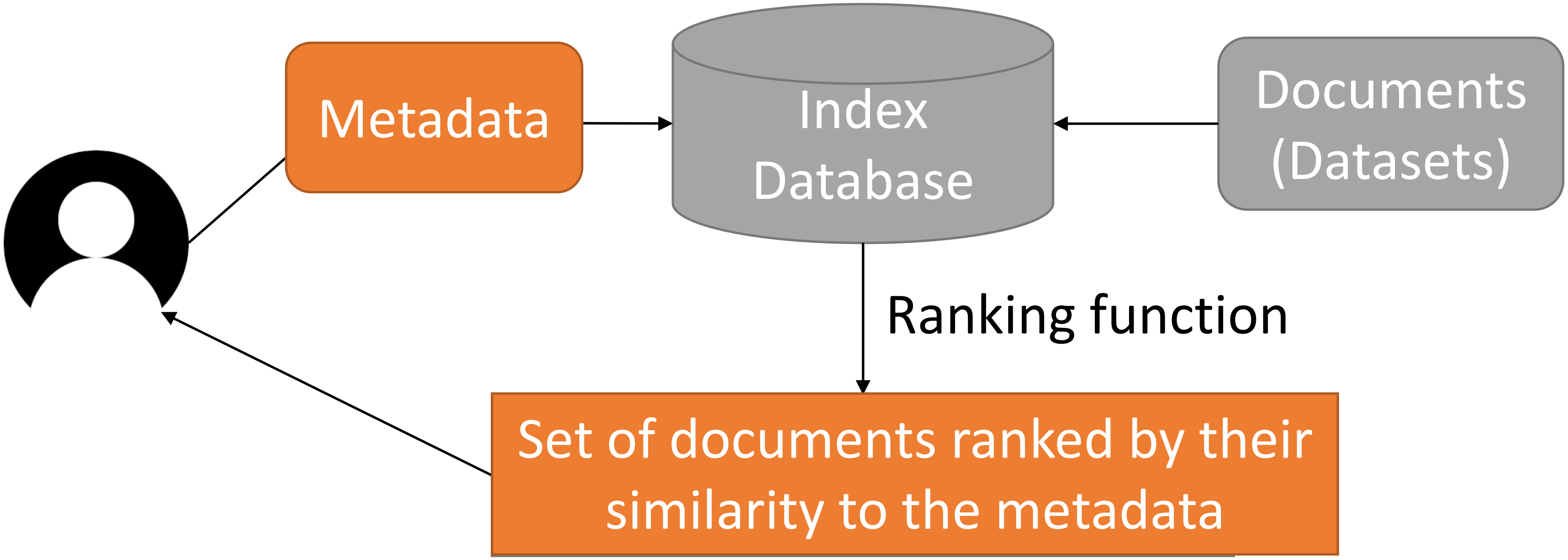
Sirocko, F; Garbe-Schönberg, C-D; Devey, CW (2000): Composition of sediments from the Arabian Sea. https://doi.org/10.1594/PANGAEA.728741

Schulz, H (1995): Planktic foraminiferal assemblage for the 10kyr time slice from different sediment cores. https://doi.org/10.1594/PANGAEA.51969

Munz, P; Siccha, M; Lückge, A et al. (2015): Distribution of planktic foraminifera in surface sediments in the northeastern Arabian Sea. https://doi.org/10.1594/PANGAEA.853966

Metadata-based Data Recommendation

- Leverages ElasticSearch More Like This (MLT) with boosting.
- MLT returns datasets that are similar to a provided data based on metadata elements, e.g., title, abstract, related publication, authors, topics, projects, devices, campaign, location, time.



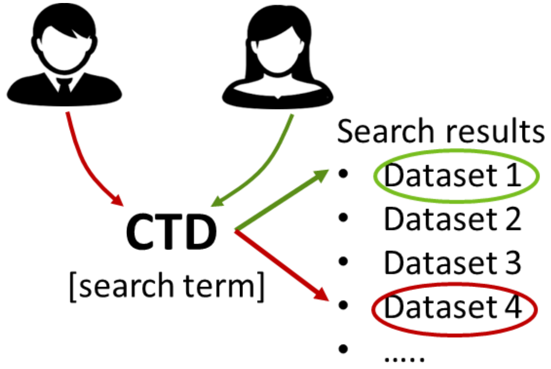
Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

Usage-based Data Recommendation

- Utilizes 3 **user interactions** (extracted from the server logs) such as *search interaction, joint download,* and *total download.*

Search interaction

Datasets examined after launching similar searches are likely to be similar.

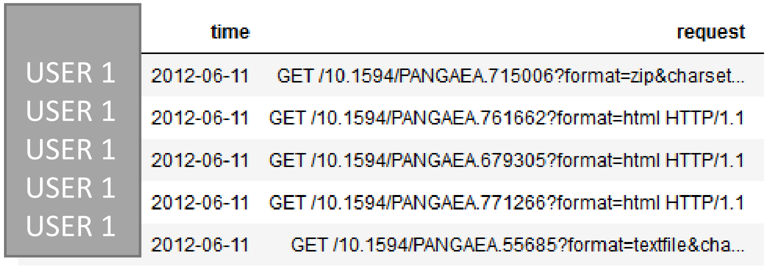


	T1	T2	Tn
D1			
D2			
D3			
D4			

Dataset vs. Search Term Matrix

Joint download

Jointly downloaded datasets are likely to be similar.



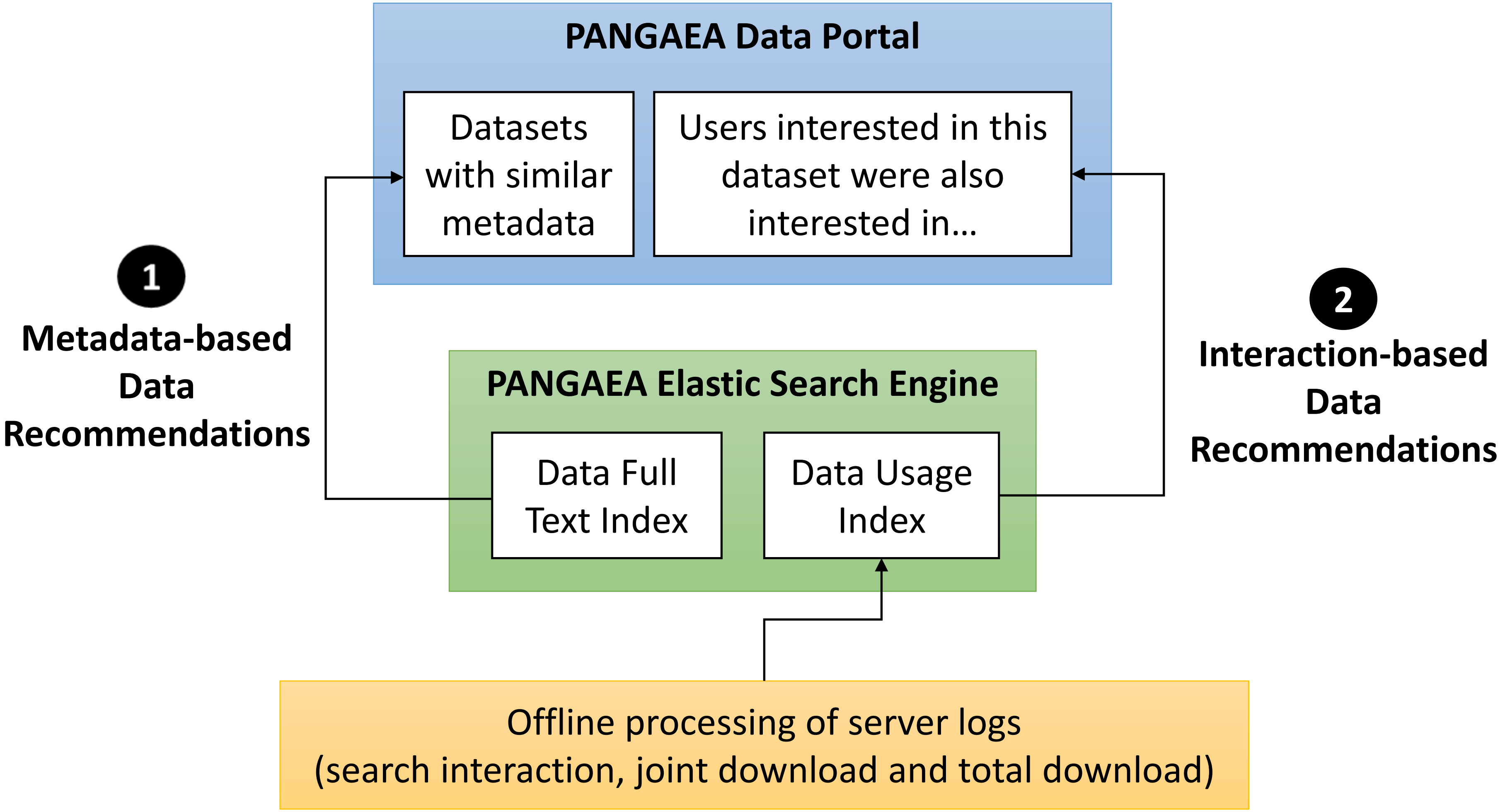
	time	request
USER 1	2012-06-11	GET /10.1594/PANGAEA.715006?format=zip&charset...
USER 1	2012-06-11	GET /10.1594/PANGAEA.761662?format=html HTTP/1.1
USER 1	2012-06-11	GET /10.1594/PANGAEA.679305?format=html HTTP/1.1
USER 1	2012-06-11	GET /10.1594/PANGAEA.771266?format=html HTTP/1.1
USER 1	2012-06-11	GET /10.1594/PANGAEA.55685?format=textfile&cha...

	U1	U2	U3	Un
D1				
D2				
D3				
Dn				

Dataset vs. User Matrix

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

System Implementation

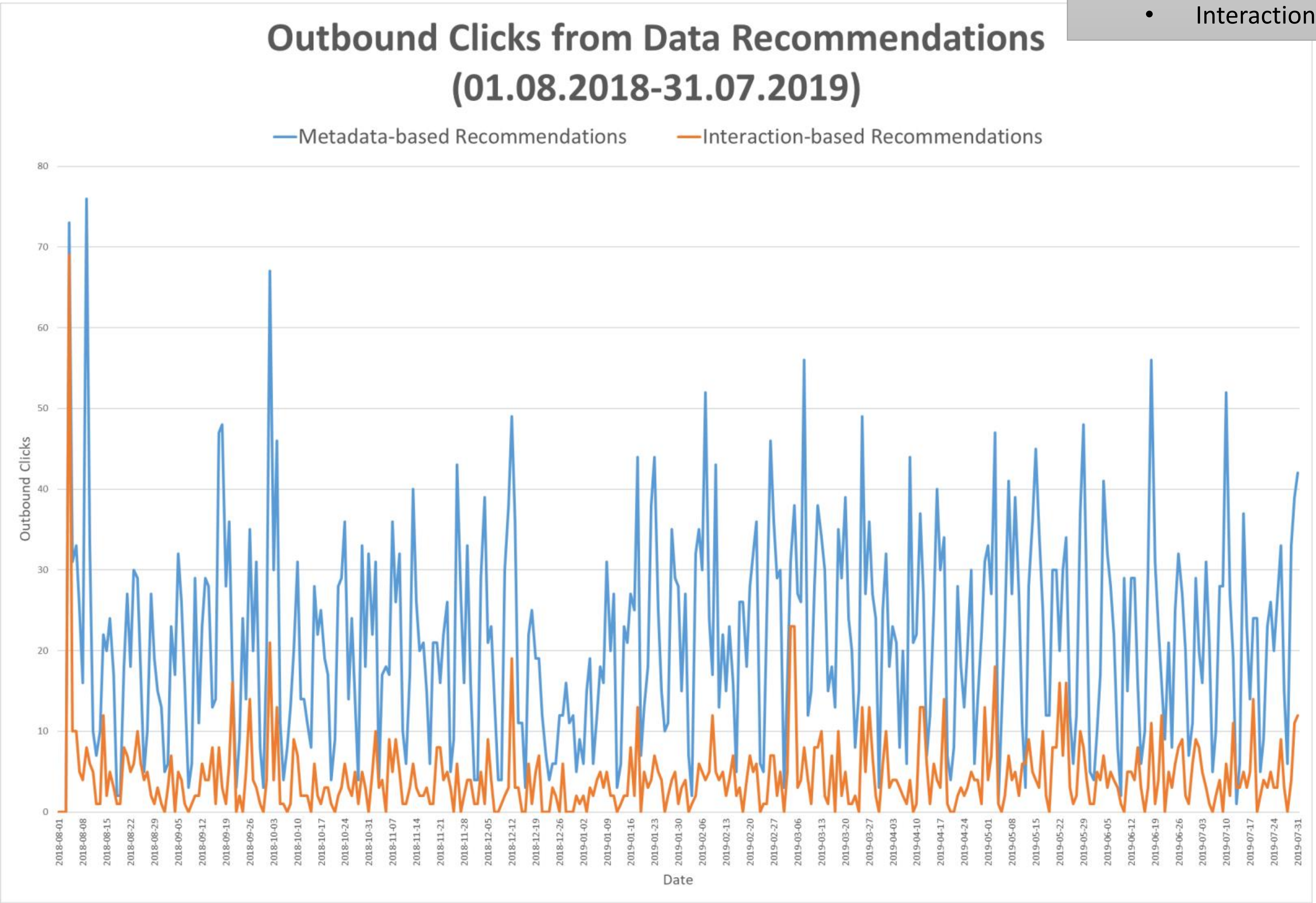


PANGAEA Data Portal
Datasets with similar metadata
Users interested in this dataset were also interested in…
1
Metadata-based Data Recommendations
2
Interaction-based Data Recommendations
PANGAEA Elastic Search Engine
Data Full Text Index
Data Usage Index
Offline processing of server logs
(search interaction, joint download and total download)

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

Online Evaluation

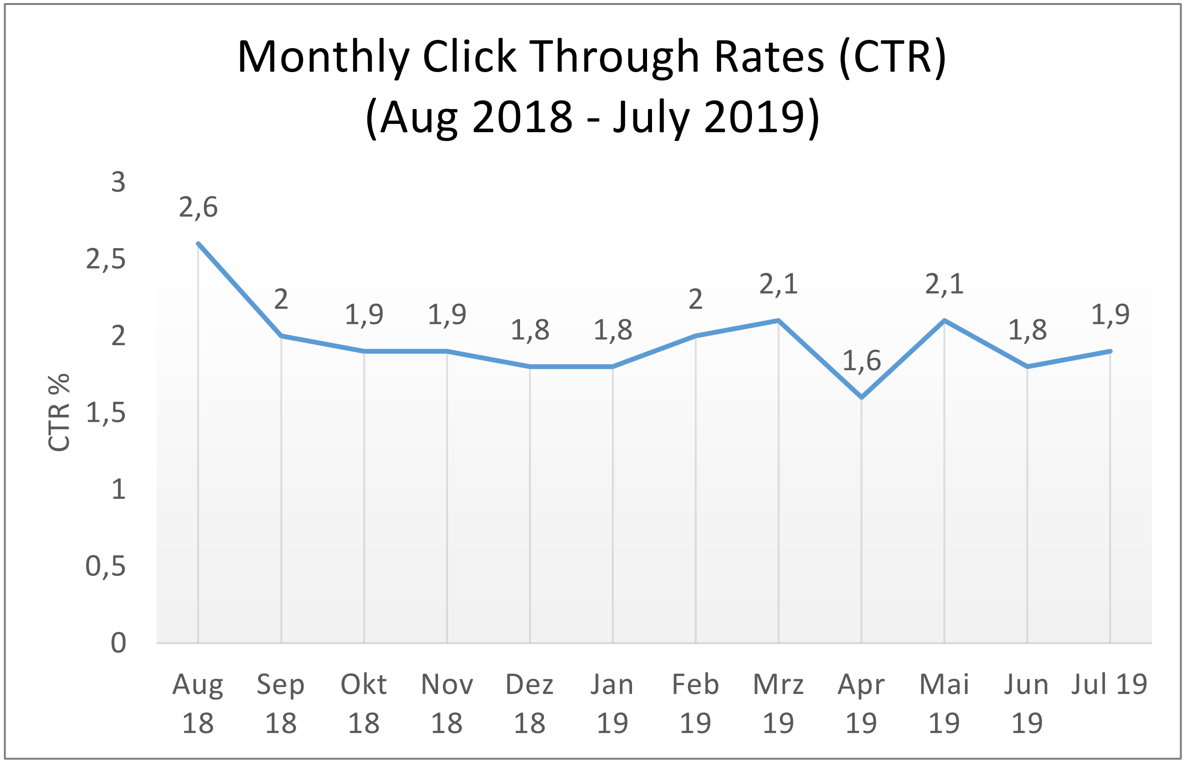
- Average clicks/day : 26
- % contribution to total events
  - Metadata : 83%
  - Interaction : 17%


Outbound Clicks from Data Recommendations (01.08.2018-31.07.2019)

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

Online Evaluation

- Click Through Rate (CTR) = Clicks ÷ Impressions



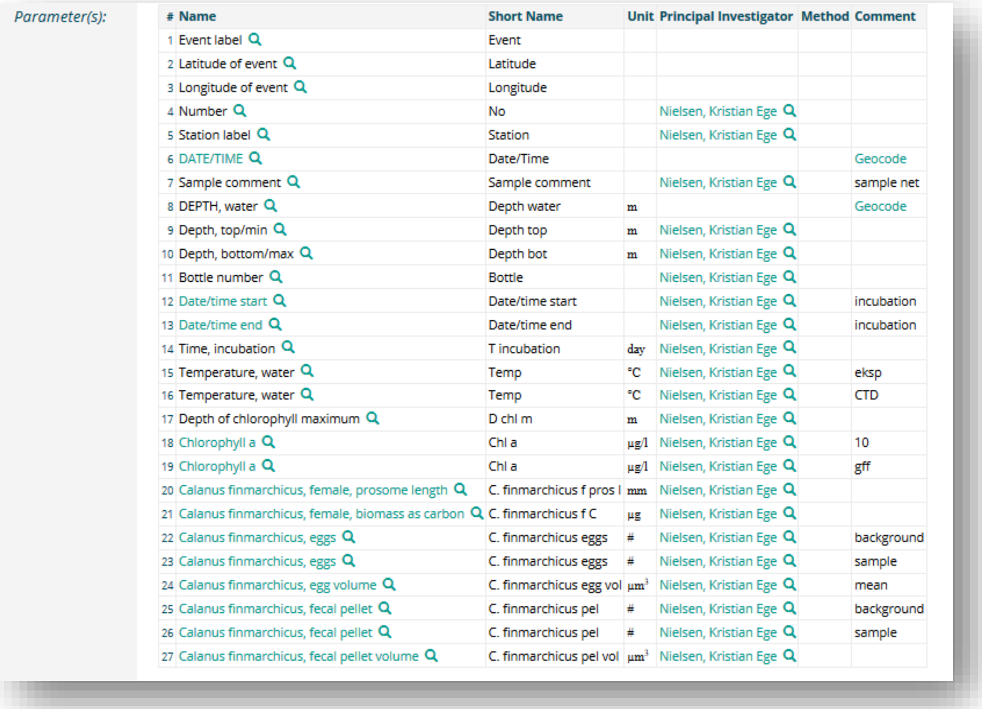
Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.

Conclusions

- Developed a recommender to improve data discovery, which presents users with two kinds of recommendations.
- Building a data recommender on top of the ElasticSearch enhances the scalability and maintainability of the recommender system.

Ongoing/planned work:

- More features – ontological concepts of parameters.
- Improve the presentation of recommendations.
- Extend online evaluation (conversion rate).



Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method	Comment
1	Event label	Event				
2	Latitude of event	Latitude				
3	Longitude of event	Longitude				
4	Number	No		Nielsen, Kristian Ege		
5	Station label	Station		Nielsen, Kristian Ege		
6	DATE/TIME	Date/Time				Geocode
7	Sample comment	Sample comment		Nielsen, Kristian Ege		sample net
8	DEPTH, water	Depth water	m			Geocode
9	Depth, top/min	Depth top	m	Nielsen, Kristian Ege		
10	Depth, bottom/max	Depth bot	m	Nielsen, Kristian Ege		
11	Bottle number	Bottle		Nielsen, Kristian Ege		
12	Date/time start	Date/time start		Nielsen, Kristian Ege		incubation
13	Date/time end	Date/time end		Nielsen, Kristian Ege		incubation
14	Time, incubation	T incubation	day	Nielsen, Kristian Ege		
15	Temperature, water	Temp	°C	Nielsen, Kristian Ege		eksp
16	Temperature, water	Temp	°C	Nielsen, Kristian Ege		CTD
17	Depth of chlorophyll maximum	D chl m	m	Nielsen, Kristian Ege		
18	Chlorophyll a	Chl a	µg/l	Nielsen, Kristian Ege		10
19	Chlorophyll a	Chl a	µg/l	Nielsen, Kristian Ege		gff
20	Calanus finmarchicus, female, prosome length	C. finmarchicus f pros l	mm	Nielsen, Kristian Ege		
21	Calanus finmarchicus, female, biomass as carbon	C. finmarchicus f C	µg	Nielsen, Kristian Ege		
22	Calanus finmarchicus, eggs	C. finmarchicus eggs	#	Nielsen, Kristian Ege		background
23	Calanus finmarchicus, eggs	C. finmarchicus eggs	#	Nielsen, Kristian Ege		sample
24	Calanus finmarchicus, egg volume	C. finmarchicus egg vol	μm^3	Nielsen, Kristian Ege		mean
25	Calanus finmarchicus, fecal pellet	C. finmarchicus pel	#	Nielsen, Kristian Ege		background
26	Calanus finmarchicus, fecal pellet	C. finmarchicus pel	#	Nielsen, Kristian Ege		sample
27	Calanus finmarchicus, fecal pellet volume	C. finmarchicus pel vol	μm^3	Nielsen, Kristian Ege		

Devaraju, A. et al., Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining, Geospatial Sensing Conference 2019.