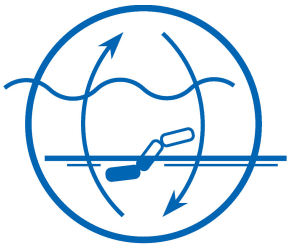
Interconnecting Arctic observatory data through machine-actionable knowledge representation: are ontologies fit for purpose?

Masters Thesis
submitted by
Kai Blumberg*



for the Marine Microbiology (Marmic) program
at the International Max-Planck Research School

Bremen, March 2018

*https://orcid.org/0000-0002-3410-4655

1st Reviewer: **Dr. Pier Luigi Buttigieg**

Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven

2nd Reviewer: **Dr. Pelin Yilmaz**

Max Planck Institute for Marine Microbiology, Bremen

STATEMENT

I herewith confirm that I have written this thesis unaided and that I used no other resources than those mentioned.

ERKLÄRUNG

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

(Place and Date / Ort und Datum)	(Signature / Unterschrift)

Contents

Summary

Well this about sums it up xP

Introduction

Ecological relevance of polar systems

Rapid effects of climate change on Polar systems

Arctic climate change With the a rapidly changing environmental conditions, the Arctic is very vulnerable.

Anthropogenic green house gas emissions are leading to increased climate change and weather extremes.

cite [1] for rapid pace of climate change and how the rate of movement of ecosystems south is unlike anything seen in earth's history making it really hard for species to keep up evolutionarily.

[2] Arctic to be free of ice within 20-50 years.

//discuss some Arctic climate change research**

Microbes and Biogeochemical cycles //maybe cut //maybe this can be fit in here? perhaps start with a microbial perspective to keep marmic happy?

> The prokaryotic and eukaryotic microorganisms that drive the pelagic ocean's biogeochemical cycles are currently facing an unprecedented set of comprehensive anthropogenic changes [3]

Polar observatories

//monitoring efforts **Polar ocean observatories and marine monitoring programs**

Polar marine monitoring initiatives such as FRAM … are working to gauge the effects of climate change on such rapidly changing environments.

AtlantOS observatory system //maybe mention this?

the Atlantic Ocean Observation Systems (AtlantOS) 1st AtlantOS Briefing Paper

FRAM & HAUSGARTEN awi polar observatory initiatives

At the forefront of climate change affected environments are polar habitats.

HAUSGARTEN intro: [4]

FRAM intro: [5]

Need for semantics in Environmental data

//Antje mentioned other fields already make better use of semantics than Environmental researcher could mention the **Monarch initiative**

PANGAEA

observational networks often upload their data to open access repositories such as the PANGAEA

Although vast quantities of environmental data are freely available to the scientific community, integrated analysis of such data is hindered by a lack of logical connections between different types of data.

** Make better use of the generated data** // why generate all this Arctic observational data when we can't get the most use of it. … transition to the need for linked data. COuld also have some other ideas to serve as the transition glue.

Observatories generate considerable volumes and varieties of data. The management and integration of such data remains a major obstacle, as the data are often not semantically interoperable. I.e. the data cannot be used in combination, because they are not annotated with a controlled vocabulary of interconnected terms which would allow for a computer to perform logical reasoning upon them.

FROM [6] paraphrase and harvest useful introductory material

> Research in ecology increasingly relies on the integration of small, focused studies, to produce larger datasets that allow for more powerful, synthetic analyses. The results of these synthetic analyses are critical in guiding decisions about how to sustainably manage our natural environment, so it is important for researchers to effectively discover relevant data, and appropriately integrate these within their analyses. However, ecological data encompasses an extremely broad range of data types, structures, and semantic concepts. Moreover, ecological data is widely distributed, with few well-established repositories or standard protocols for their archiving and retrieval. These factors make the discovery and integration of ecological data sets a highly labor-intensive task.

open science

[7] Open Data Means Better Science > Data provides the evidence for the published body of scientific knowledge, which is the foundation for all scientific progress. The more data is made openly available in a useful manner, the greater the level of transparency and reproducibility and hence the more efficient the scientific process becomes, to the benefit of society. This viewpoint is becoming mainstream among many funders, publishers, scientists, and other stakeholders in research, but barriers to achieving widespread publication of open data remain.

FAIR the FAIR data guiding principles (machine-focused findability, accessibility, interoperability reusability) [8]

AWI data is currently Findable and accessable at a high level for example within Pangaea files. Improvements would be to make the data findable and accessible. Improve Polar data re-usability with the cryo-MIXS exten-

sion paper in prep. Most importantly Interoperability, a formally controlled and machine accessible vocabulary, through ontologies, (ENVO, PATO, PCO, ECOCORE).

We need to transition to using **linked data** wiki

Such efforts could benefit from *linked data* a term referring to data which is published in a structured format which allows it to be linked to other data.

This is done by making use of standard web technologies.

Linked data makes use of Hypertext Transfer Protocol (HTTP) to give data objects a web address, as well as the Resource Description Framework (RDF) [9] a … to share information in a machine-readable format. This allows for

> In computing, linked data (often capitalized as Linked Data) is a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers.

semantic web

wiki

> The Semantic Web is an extension of the World Wide Web through standards by the World Wide Web Consortium (W3C). [10] The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). [11]
>
> According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries".[2] The term was coined by Tim Berners-Lee for a web of data that can be processed by machines[3]—that is, one in which much of the meaning is machine-readable.
>
> Linked data may also be open data, in which case it is usually described as linked open data (LOD).

linked open data

read and cite [12] about linked open data arguments presented by tim_berners-lee and on the wiki page on open data https://en.wikipedia.org/wiki/Open_data

from the wiki: citing [12] > Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

> Open data which is also linked data is usually termed linked open data.
>
> Open data may include non-textual material such as maps, genomes, connectomes, chemical compounds,

parlalles with open science wiki

> the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.

OPeNDAP

> OPeNDAP will be a fundamental component of systems which provide machine-to-machine interoperability with semantic meaning in a highly distributed environment of heterogeneous datasets.

Open-source Project for a Network Data Access Protocol There is a need for semantic interoperability …

Internet of things

build up on the semantic web will be the Internet of things, which will have a major inpact on environmental sciences in terms of sensor newtorks … as there will be be an influx of ocean sciences big data such as sensor networks. SWE SOS and SENSORML // maybe

UN decade of ocean science for sustainable development 2021-2030.

This is related to his work on ocean best practices (to generate such data): as there will be be an influx of ocean sciences big data such as sensor networks. SWE SOS and SENSORML look more into this. Ontologies and this kind of semantic work will be important for mobilize this large data generated by sensor networks, for ocean best practices decade of ocean science. My work will help prepare for this on slot of coming big data using the awi data case study.

Tara oceans

cite tara as …

hurwitzlab projects

mention the http://www.hurwitzlab.org/projects/ocean-cloud-commons/ ocean cloud commons as providing a querable version of Tara.

Ontologies and the OBO Foundry

Ontology, a human and machine readable semantic representation of domain knowledge …

An ontology is a hierarchically structured, machine and human readable representation of the knowledge used by experts to describe entities, and capture the relationships between them [13]. In informatics, ontologies exist in the form of a knowledge graph, where nodes represent entities, and edges represent logical relations linking entities together (i.e. axioms). Ontologies provide a digital semantic infrastructure upon which advanced querying, discovery and analysis of data can occur.

Ontologies are a methodology to systematically structure and connect data, allowing users to ask more complicated questions involving the synthesis of disparate data types which currently can not be combined.

knowledge graph

[14] citation for knowledge graph

for knowledge outreach

Knowledge graphs are becoming more popular and useful, need to bridge the gap between patchy but growing resources such as Wikipedia, and expert knowledge (locked away in text books), using an ontology helps to bridge this, it can be applied to querying Wikipedia data and for improved semantic representation make data FAIR. Ontology for an agreed upon term structure

//revise a bit from lab rotation: Because, no single knowledge graph can encompass the needs of interdisciplinary projects, work must be done in a coordinated fashion with other ontology researchers and developers. In order to interconnect ontologies representing scientific knowledge from different domains, the Open Biological and Biomedical Ontology (OBO) Foundry and Library was created [13]. The OBO Foundry and Library established a set of principles by which to develop and coordinate ontologies such that the scientific knowledge they represent and hence the data they link can interoperate. These ontologies share a common upper level in the hierarchy and use of the same types of logical connective operations to interlink their knowledge. Following these principles are a family of ontologies representing scientific knowledge from non-overlapping domains, which can be used in combination to describe natural phenomena in greater depth. OBO compliant ontologies make use of the Basic Formal Ontology (BFO) [15] [16] [17], to ensure they have a compatible hierarchical structure, and use logical relations from the Relations Ontology (RO) [18], to standardize the connections between their knowledge.

OBO compliant ontologies can be benefit observatory networks such as Hausgarten FRAM, by providing connections between data collected by researchers of different disciplines studying overlapping entities.

//example from my rotation add something like this. > For example sea ice physicists studying the reflectivity of various ice mass features, may have light intensity data that would help microbial ecologists studying photosynthetic bacteria in brine channels, to calculate the light dependent growth rates of such bacteria

Ontologies of interest

ENVO for representing environmental semantics.

ENVO papers: [19] [20]

The Environment Ontology (ENVO) represents expert knowledge about different types of environments[19][20]. ENVO is an OBO aligned ontology.

Environmental knowledge represented by ENVO is used to annotate data from a variety of life science disciplines including oceanography and polar research. [19][20]

Gene Ontology go paper: [21]

GO frequently used to interpret omic data [21]. It has been used to do genomewide RNA expression profile data to compare samples based on shared biological pathways. [22]

The combination of GO and ENVO is less frequently attempted. [23]

Paring GO with ENVO is a potential avenue for future study allowing researchers to ask questions such as > "What is the omic potential of microbes associated with particular environments?".

SDGIO

Policy and SDGIOs

Making Marine Life Count: A New Baseline for Policy [24] Just use a little bit from this as policy intro.

DOOS Consultative Draft (no DOI) for insight into functions that can be understood as ecosystem services of the deep, and thus linked to natural capital.

UN sustainability development goals in response to climate change

The effects of increased climate change and extreme weather events are hardest felt by indigenous people and the global precariat subsiding via land and ocean subsistence farming and fishing.

UN publication: TRANSFORMING OUR WORLD: THE 2030 AGENDA FOR SUSTAINABLE DEVELOPMENT no DOI reference for the sustainable development goals and targets.

The UN framework for SDG's have setup targets for improvements to many global issues such as UN SDG 14 for ocean health.

14.1

> By 2025, prevent and significantly reduce marine pollution of all kinds, in particular from land-based activities, including marine debris and nutrient pollution

link the nitrogen phosphorus data to the concept of those cycle being out of balance as doccumented in the Planetary Boundaries: Exploring the Safe Operating Space for Humanity paper. [25]

United Nations Environment Programme

SDGIO is an OBO compliant ontology

uses the same interoperable semantic standards to ENVO. Although UNEP PURLS cannot currently be queried.

Linking earth science data initiatives such ESIP Open knowledge network to the UN SDGIO's

There exist a variety of earth and life science initiatives attempting to capture and represent the knowledge associated with environmental data. …

The knowledge required to interface the concepts needed for the Sustainable development goals are represented in a machine operable form via the SDGIO sustainable development goals interface ontology.

Example Ontology usage

A communal catalogue reveals Earth's multiscale microbial diversity. //Uses EMPO a light-weight application ontology built on ENVO the Earth Microbiome Project Ontology [26] //good to have an example which demonstrates the utility of ENVO for an application ontology to provide utility.

//from my rotation rewrite example > Thesen et al.13. show how such a federated semantic approach can enhance handling of environmental and phenotype data, in order to ask increasingly complex questions such as "Which crop varieties are expected to do well in a particular location over the next century?". Thesen et al Emerging semantics to link phenotype and environment [27]

ontology management of big data

For example the HASNetO ontology [28] > has been in use to support the data management of a number of large-scale ecological monitoring activities (observations) and empirical experiments.

maybe also cite:

http://dx.doi.org/10.1016/j.margen.2017.02.006 piers paper with the italians.

role of data in 2015 - 2020 ESIP Strategic Plan

link to my log

Demonstration of ontobee ontology system use

currently systems are able to answer questions such as

> What compounds play a role as algae metabolites?

I can get data back, typical question in awi work, in order to answer this, Put as intro example. Have this be in the introduction.

easy enough to answer Make use of the CHEBI class: algal metabolite

purl

querying the ontobee sparql endpoint

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT DISTINCT ?purl (STR(?label) as ?label)
WHERE
{
  ?purl rdfs:subClassOf/owl:someValuesFrom obo:CHEBI_84735.
  ?purl rdfs:subClassOf/owl:onProperty obo:RO_0000087.
  ?purl rdfs:label ?label.
}
GROUP BY ?purl
LIMIT 10
```

This query gives us the purls and the labels of the first 20 classes which are subclasses of 'has role' some algal metabolite

using the restriction has role.

The group by ?purl is to ensure we don't get duplicates of purls which have duplicated labels such as http://purl.obolibrary.org/obo/CHEBI_15756 which has labels: `hexadecanoic acid` and `Hexadecanoic acid`

Returning the following results:

Table 1: Compounds serving as algal metabolites.

purl	reference doi
CHEBI_80716	aplysiatoxin 11(R)-HEPE(1-)
CHEBI_90820	all-cis-docosa-7,10,13,16-tetraenoic acid
CHEBI_53487	(7Z,10Z,13Z,16Z,19Z)-docosapentaenoic
CHEBI_53488	acid 3-mercaptopropionate
CHEBI_86386	2-hydroxypropanoic acid microthecin
CHEBI_78320	2-palmitoylglycerol 2-oxoglutarate(2-)
CHEBI_51835	Sucrose
CHEBI_75455	
CHEBI_16810	
CHEBI_17992	

structure of thesis into competency questions:

In order to leverage growing data and knowledge representation semantic infrastructure we test if a semantic knowledge web represented by an ontologies can be used in combination with AWI data to address competency questions such as:

Materials and Methods

Model polar datastore creation

used these ENVO releases of interest: Ecotone, Polar express, Hot tub time machine.

Datasets used in Datastore

1. Inorganic nutrients measured on water bottle samples at AWI HAUSGARTEN during POLARSTERN cruise MSM29. [29]
2. Physical oceanography and current meter data from mooring TD-2014-LT. [30]
3. Chlorophyll a measured on water bottle samples during POLARSTERN cruise ARK-XXIV/2. [31][32]
4. Global chlorophyll "a" concentrations for diatoms, haptophytes and prokaryotes obtained with the Diagnostic Pigment Analysis of HPLC data compiled from several databases and individual cruises. [33][34]
5. Biogenic particle flux at AWI HAUSGARTEN from mooring FEVI7. [35][36]
6. Snow height on sea ice and sea ice drift from autonomous measurements from buoy 2015S22, deployed during the Norwegian Young sea ICE cruise N-ICE 2015. [37][38]
7. Sea ice thickness at Ice Camp 1 on 2013-09-01 (GEM2IceTh_DiveHole_IceStation1). [39][40]
8. Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-1_10. [41][42]
9. Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-2_11. [43][42]
10. Unpublished metagenomic data from deep sea sediments from Hausgarten POLARSTERN Polarstern cruise PS85, encompassing both functional genomic data, and 16S taxonomic data, courtesy of Josephine Z. Rapp.

Tools used to build the datastore

semantic technologies make use of the specifications of the World Wide Web Consortium (W3C) [10]

SPARQL 1.1 Query Language and W3C Recommendation 21 March 2013 query language for RDF [44]

Python [45] Version 2.7.12

RDF 1.1 Concepts and Abstract Syntax W3C Recommendation 25 February 2014 [9]

RDF specifications turtle [11],

Anything To Triples (any23) a library, a web service and a command line tool that extracts structured data in RDF format from a variety of Web documents [46].

owl [47]

> The Web Ontology Language OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. OWL is developed as a vocabulary extension of RDF (the Resource Description Framework)

Protégé [48] [49]

Semantic Data Annotation

Semantic annotation of example data was conducted in the RDF serialization turtle, drawing upon its blank node feature to facilitate scripting owl code in RDF. Annotations make use ontology terms from the OBO Foundry [13]. Ontology terms can be search for using Ontobee A linked data server hosting ontologies and their terms. [50]

sparql query scripting

scripts to perform queries were written in python verion?

using the rdf-lib module

Queryies preformed against the ontobee endpoint http://sparql.hegroup.org/sparql/ a serive provied by the He Group [50]

The script makes use of a conjunctive graph object from the rdf-lib module, to emulate an RDF triple store.

material and methods used to answer competency questions

Interconnecting genomic and environmental data via ontology methods

genomic_data

Abyssal and Bathyal metagenomic data provided by Jose Rapp. Run using the metagenomicNGS assembly and annotation pipeline available from: here I don't have access to view it, but I have the script saved `/kblumberg_masters_thesis/working/genomic_workflow`

Samples 1 and 2 collected from depths of 1244m and 2403m which best correspond to 'marine bathyal zone biome'

Samples 3 and 4 collected from depths of 3531m and 5525m which best correspond to 'marine abyssal zone biome'

Neritic transcriptomic data provided by Dr. David Probandt, from from sandy sediment of a marine neritic benthic zone biome of about 8 m depth. Use the first 4 samples: labeled X1, X2, X3, X4.

Ontology guided data assembly for ecological analysis methods

Connecting information contained within ontology terms to the term authors methods

Connecting datasets and publications about an ontology term methods

Interconnecting stated and unstated knowledge via an ontology knowledge graph methods

Mobilizing ontology annotated data methods

Ontological representation of real world phenomena methods

Vocamp Virtual Glacial Hackathon methods

Results

The results section is organized by **Competency questions** seeking to evaluate fitness for purpose of interconnecting disperate data via well-structured ontologies.

experiments to test knowledge model against competency questions.

Interconnecting genomic and environmental data via ontology

We made use of the interoperate semantics of the Gene Ontology and the Environent Ontology, to mobilzie and query data …

make use of … we were able to moblize data to ask and answer a question such as:

> "What are the relative abundance frequencies of oxidation-reduction process genes in various types of marine biomes?"

To evaluate the use of such interlinking of semantics to make comparisons we examined if the results of the previous question differentiate marine sediments from bathyal and abyssal zone biomes from those of neritic as would be expected biologically?

Table 2: Selected results of relative genomic and transcriptomic abundances of oxidation-reduction process in various types of marine biomes highlighting differences between deep neretic samples.

label	marine abyssal zone biome	marine bathyal zone biome	marine neritic benthic zone biome
oxidation-reduction process	18.15	18.39	9.36
aerobic respiration	0.23	0.26	0.87
methanogenesis	0.11	0.12	0.06
ATP synthesis coupled electron transport	0.06	0.06	0.04
L-lysine catabolic process to acetate	0.06	0.07	0.01
respiratory electron transport chain	0.03	0.03	0.13
electron transport chain	0.02	0.02	0.05
photosynthetic electron transport in photosystem II	0.00	0.00	16.08
photosynthetic electron transport chain	0.00	0.00	1.38

//Discusson From a biological perspective yes the results make sense.

Deep samples had double the abundaces of non specific PFAM annotations to general oxidation-reduction reduction processes 18%, relative to neritic samples, which had three fold increases in aerobic respiration gene abundaces relative to the deep samples.

Deep samples had nearly double methanogenesis gene abundaces than neritic samples. Neritic samples had much greater relative respiratory electron transport chain abundances than deep samples.

neritic samples have elevated abundances of photosynthetic related genes, 16% photosystem II electron transport and 1.4% photosynthetic electron transport chain, contrasting with the 0.00% abundances of such genes in the deep benthic samples.

These results indicate as would be expected that neritic samples are relatively enriched in photosynthesis, aerobic respiration and respiratory electron transport chain related genes relative to the deep samples enriched in undifferentiated oxidation-reduction processes, and methanogenesis related genes.

Further exploring the use of interoperable GO and ENVO semantics to compare genomic abundances of samples annotated with different ENVO terms, we can ask a question such as:

> "What are the relative abundance frequencies of vitamin biosynthetic process genes in various types of marine biomes?"

Table 3: Relative abundance of vitamin biosynthetic process genes in various types of marine biomes.

label	marine abyssal zone biome	marine bathyal zone biome	marine neritic benthic zone biome
riboflavin biosynthetic process	0.25	0.25	0.07
cobalamin biosynthetic process	0.19	0.19	0.03
pantothenate biosynthetic process	0.13	0.12	0.04
thiamine biosynthetic process	0.10	0.11	0.04
pyridoxine biosynthetic process	0.10	0.09	0.02
vitamin B6 biosynthetic process	0.05	0.05	0.02
pyridoxal phosphate biosynthetic process	0.05	0.05	0.02
pyrroloquinoline quinone biosynthetic process	0.00	0.00	0.00
anaerobic cobalamin biosynthetic process	0.00	0.00	0.00

From this table we note that in the deep sample abyssal and bathyal, the relative gene abundance of riboflavin genes was ~3.5 times greater. As flavins have been implicated as electron donors in the reduction of insoluble ferric to soluble ferrous iron as well as the transport of ferrous to the cytoplasm [51][52], we investigated transition metal ion binding and transport subclasses where we found that ferrous ion binding is 0.03-0.04% abundance in deep vs 0.00% in neretic and ferrous iron transport gene abundance is double in deep than neritic 0.04% vs0.02% in neretic samples.

//Discussion OBO complient Ontologies are a living knolwedge model don't take an absolute stance, but rather an open knolwedge model *FIND CITATION IN [15] and can be used to represent hypothesis without stating them to be the absolute truth. An example of where incorporating expert knowledge into the OBO ontology semantic layer, would be to represent the knowledge that flavin production is likely linked to extracellular iron reduction.

Axioms such as … could be added to … class which would faciliate the search for data based on such knowledge. This knolwedged represented wihitn the semantic layer could faciliate the search for data about participants in a iron reduction process, helping to make the connection between increased riboflavin biosynthesis and increased ferrous ion binding and ferrous iron transport genes.

// add one one/2 more PCOA figures which differentiate the neritic and deep samples using the helliinger tranformation or somehting like that.

SECTION NOTE How well can ontology connect information ?data and people

Ontology guided data assembly for ecological analysis

New Version:

Utilizing semantics to draw together relevant knowledge. If there were for example a class such as

sea-ice associated phytoplankton community

//May want to restructure the classs to be like fungi-associated environment sea-ice associated phytoplankton community

//OR I THINK I SHOULD DO IT LIKE THIS!!! environment determined by a biofilm on a saline surface environment determined by a phytoplankton community associated with sea-ice.

definition:

```
A phytoplankton community which is adjacent to some sea ice.
```

with axioms:

`'subclass of' some 'phytoplankton community'`

`'has part' some 'chlorophyll a'`

`'located in' some ('seawater' and ('part of' some 'marine water body'))`

`'adjacent to' some 'sea ice'`

//We are able to leverage the semantics referenced by such a **sea-ice associated phytoplankton community** class, and

Assembling a list of all the classes which are referenced by the **sea-ice associated phytoplankton community** along with their subclasses, we get a list of semantic terms which we can query for data about. Imagine in the example we have additionally filtered for data in the same spatiotemporal location.

By leveraging the semantics included in a term such as **sea-ice associated phytoplankton community** along with the the ontology knowledge graph, we are able to retrieve and assemble relevant data opon which to perform an ecological analysis. By doing a principal component analysis on this data we can investigate which of these many environmental variables have the greatest loading on the analysis.

`/home/kai/Desktop/grad_school/marmic/master_thesis/kblumberg_masters_thesis/datastore/competency_questions/assemble_data_for_ ecological_analysis`

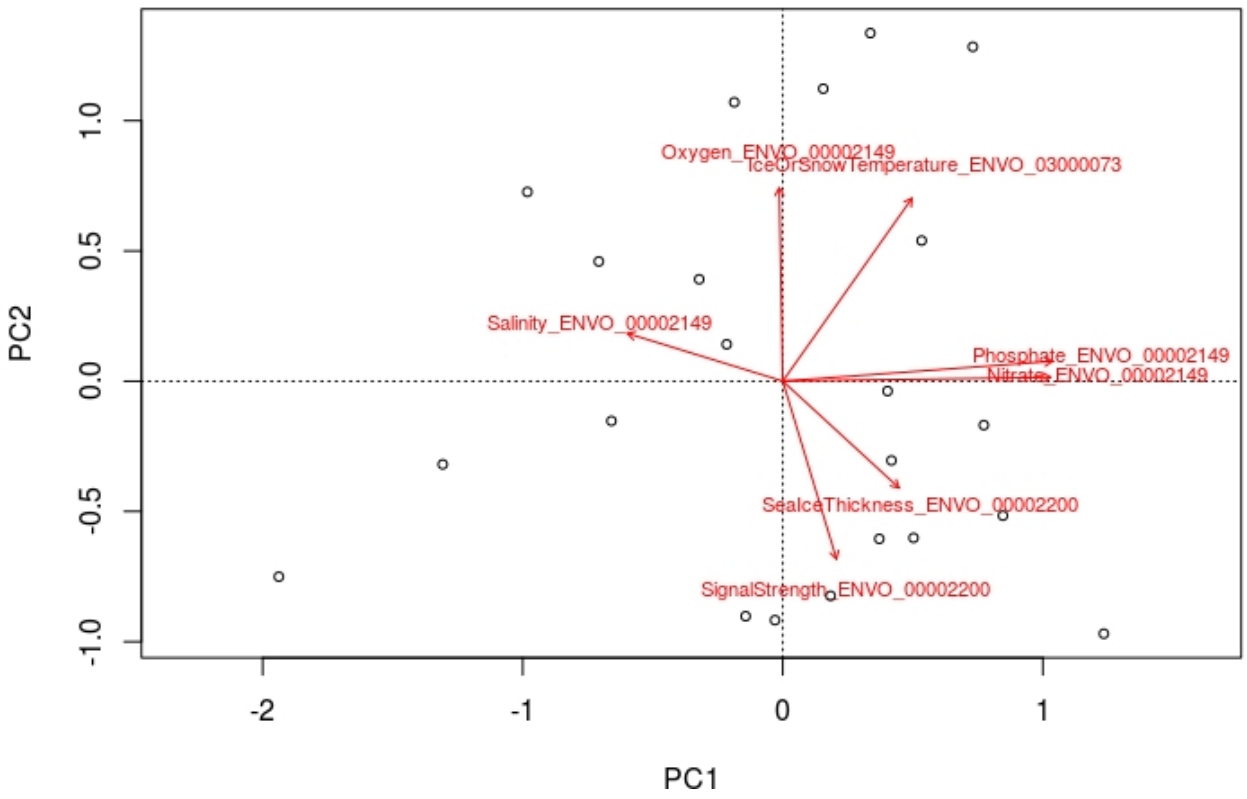


Figure 1: PCA on assembly of data about participants in sea ice formation processes.

material/methods: take columns: influence_snow_depth.csvSignalStrength, inorganic_nutrients.csvNitrate, physical_oceanography.csvOxygen, inorganic_nutrients.csvPhosphate, physical_oceanography.csvSalinity, ice_algal_chlorophyll_myi.csvIceOrSnowTemperature, influence_snow_depth.csvSeaIceThickness

//old disscussion stuff for Anya meeting:

My workflow assembles useful and expected data such as: minimum and maximum sea ice depth, sea ice temperature, the degree of illumination of sea ice, sea ice texture, and thickness of snow on sea ice,

Also retrieved are some unexpected but potentially useful data such as: sea water salinity, sea water chlorophyll and areal chlorophyll a concentration, sea water phosphate and nitrate concentrations.

When assessing the potential of some seawater to freeze, data about the water's salinity valuable as per the relationship between salinity and the freezing point of water.

Nutrient data to asses the role of nutrient limitation on bloom termination, post sea ice retreat.

Also potentially valuable to asses such sea water for it's potential to freeze are data about nitrate, phosphate and chlorophyll, as they are indicators of the biotic activities in such seawater which are most likely linked to the extent to which the ice is freezing. This leads to the generation of more hypothesis about relationships between for example sea ice and nutrient concentrations or chlorophyll which can be harvested into the knowledge graph. Examples of which include the occurrence of blooms associated with the melting of sea ice, or the effects of

sea ice melting on on water body stratification, which could potentially inhere in the nutrient data. Codifying such relationships into the semantic layer and using that layer to annotate and mobilize data provides a way illuminate the connections between otherwise disparate data sets, for example using nitrate and phosphate data in combination with other data such as temperature, to help report on the potential freezing or melting processes.

Connecting information contained within ontology terms to the term authors

Question: > "How well do ontologies connect authors of terms to the information they helped to encode?"

ontology	% terms with created by	% terms with term editor
envo	14.5	4.2
envoPolar	17.2	31.4

see my thesis here

The Internet is enabling collaborative dissemination of knowledge and data

id who's knowledge is captured

who conbributed this knowed to the system. 1) need ppl to record knowed into ontology, 2) ontologies need to microcredit contribution 3) microcredit connect to ORCID. fit ont micro cred knowl to unamb ids of person connecting to lving system, that system queryable.

in methods examined micro creding (put concise version of methods for each competency question)

in results

in discussion not all parts are generated by individuals some are autogenerated. pier has id range doesn't put orcid on terms. hop to editor file avg pson won't do that so, Make the case for microcrediting.

where did knowledge come from dbx ref and author,

choose an ontology like envo find out what proportion of terms have ORCID there are other old ways, but orcid will most likely be kept current, talk about other ways of crediting, search term createdby term editor. not everyone has an orcid, for this work focus on it because const updated website doens't change… to make ontologies fit for purpose persisent pulled by api.

answer if we can trace back knowledge to initator of knokew, but only able to do so for orcids aware of editor files but impracticalbe to find this for each ontology , better have ocids which are universally pullable.

Ontologies being semantic representations of expert knowledge should empower users to connect knowledge but also facilitate networking among scientists.

Hence as part of the evaluation of the fitness for purpose of ontologies for interconnecting interdisciplinary data, we evaluated the utility of ontologies and semantic querying to retrieve author information about the creator of an ontology term.

Connecting datasets and publications about an ontology term.

Retrieving publications associated with datasets about parts of an ontology term.

Can we use ontologies to find papers referencing data about a term?

> "What are all the papers which reference any data set, which is about a part of a marine biome?"

The following returns a variey of papers which have been referenced by datasets which are about parts of marine biomes. In this example two datasets which are both annotated as part of a marine water body along with their associated publication DOI's [53][54][55][56][57][58][59][60][32][61][62][63][64][40]

Table 5: DOI's of publications obtained querying for references of datasets which are about part of a marine biome.

data set	reference doi	reference title
global chlorophyll a	10.1016/j.dsr.2011.01.008	An evaluation of the application of CHEMTAX to Antarctic coastal pigment data [53]
	10.3402/polar.v34.23349	Summertime plankton ecology in Fram Strait-a compilation of long- and short-term observations [32]
	10.1029/2003EO380001	Unique data repository facilitates ocean color satellite validation [62]
	10.1029/2005JC003207	Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll [64]
influence snow depth	10.1002/2016JC012325	Influence of snow depth and surface flooding on light transmission through Antarctic pack ice [40]

Interconnecting stated and unstated knowledge via an ontology knowledge graph

Assessing the interconnectivity between stated and unstated knowledge in an ontology knowledge graph

[14] citation for knowledge graph

> "Are ontology knowledge graphs sufficiently well connected to be able to lead researchers to new knowledge via unstated linkages to identified knowledge?"

take the assumption that the ontology graph lead scientists to other kinds of data, look at envo polar subset as graph, from any node what's the average degree. import subset inmpot to cryoscape and calc avg degree of nodes

program cytoscape. find nodes in envo polar. Calc degree dist. Can this lead scients to other data, ice gets can look at uses in ontology on ontobee. start with 3/4 glaciers terms how many things are 5 steps away. can import ontologies into view and look for things x degrees of seperation. interpret intrconnecctivity. interconnect knowledge about things,

//vertex and edge betweenness are (roughly) defined by the number of geodesics (shortest paths) going through a vertex or an edge. //The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

//closeness centrality: sum of the length of the shortest paths between the node and all other nodes in the graph (how central a node is in a graph) //clustering coefficient: measure of the degree to which nodes in a graph tend to cluster together. edges between neighbors which do exits relative to all possible ones which could exist. 0=unconnected 1=maximally connected to all neighbors

//network diameter: is the largest distance between two nodes

//average shortest path length, also known as the characteristic path length, gives the expected distance between two connected nodes.

//average number of neighbors indicates the average connectivity of a node in the network

//average connectivity is a measure for the expected number of vertices that have to be removed to separate a randomly chosen pair of vertices

//network density hows how densely the network is populated with edges

//multi-edge node pairs indicates how often neighboring nodes are linked by more than one edge.

Table 6: network parameters calculated from the graph of the envoPolar subset of ENVO.

network parameter	value
number of nodes	265
number of edges	402
clustering coefficient	0.047
connected components	8
network diameter	7
average shortest path length	2.190
average connectivity (number of neighbors)	2.875
network density	0.0
number of self-loops	0
multi-edge node pairs	20

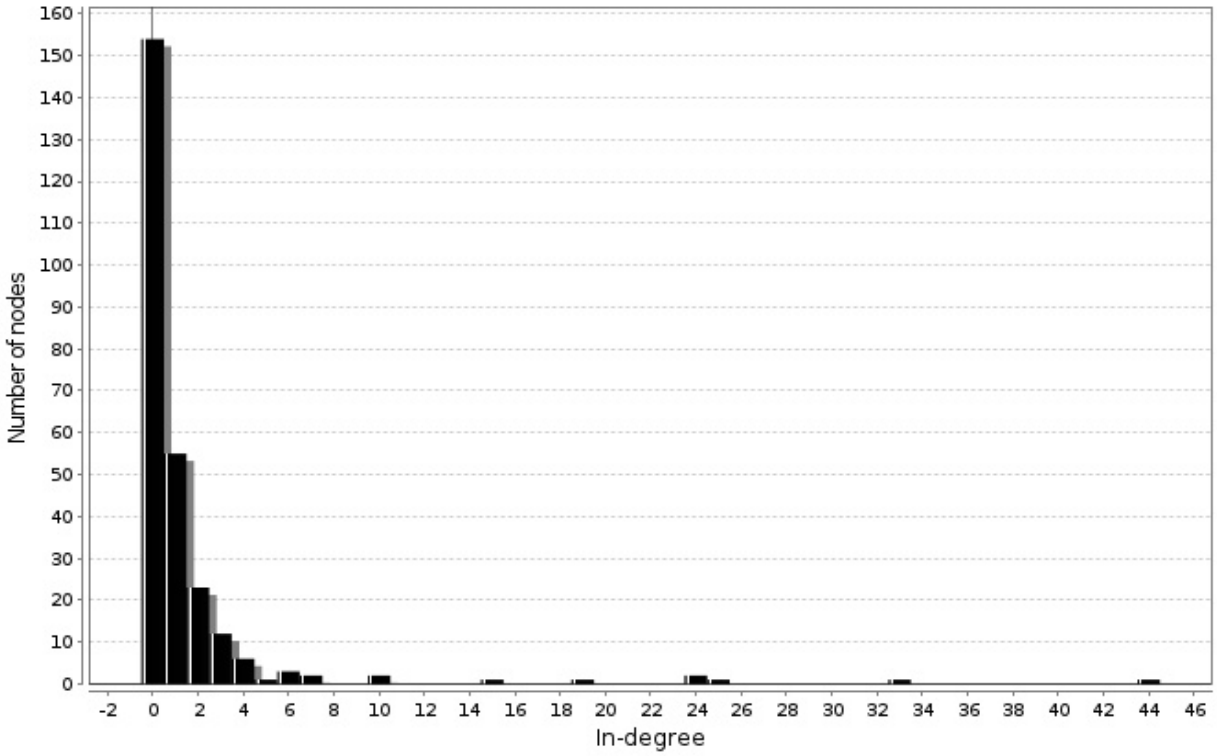


Figure 2: In degree distribution of the envoPolar subset analyzed as a graph.

//Discussion Yes classes are interelated in part because of process cauasal algebra in RO. capture parts using BFO. using RO to a good degree

discuss if you don't use upper level models these scripts wouldn't work.

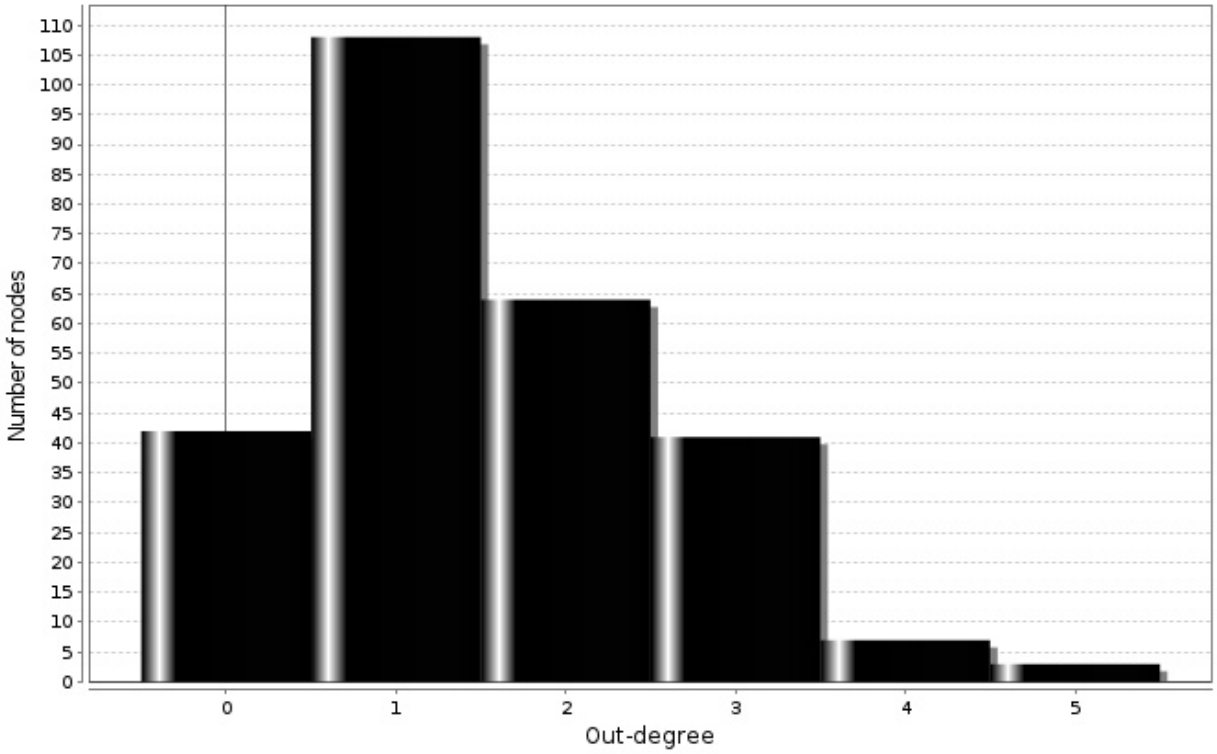


Figure 3: Out degree distribution of the envoPolar subset analyzed as a graph.

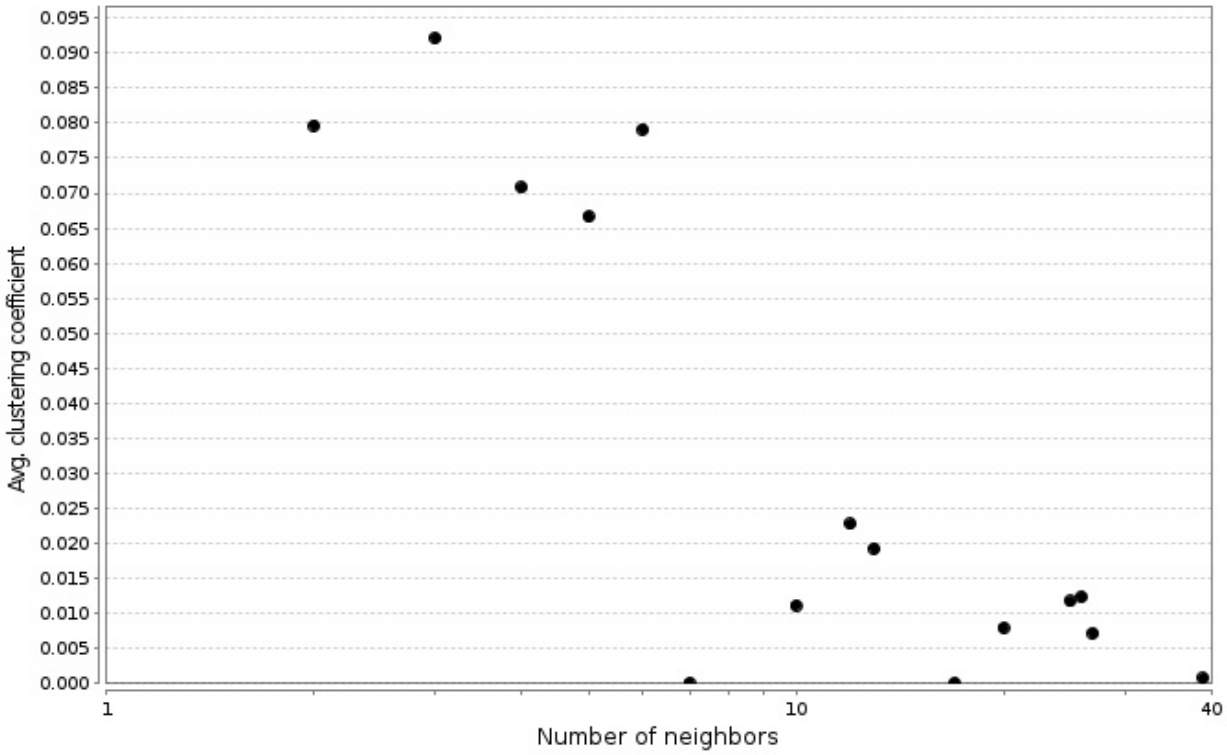


Figure 4: Average clustering coefficient of the envoPolar subset analyzed as a graph.

Mobilizing ontology annotated data

> What level of querying expertise is required to access Arctic observatory data annotated with obo complient owl axiomatic structures?

have 3 levels of expertise basic intermediate advanced. have the basic one just be data about X . The first querying case.

present it with histograms / bar charts see my notebook. do it with the two cases:

Querying exclusive AND annotations

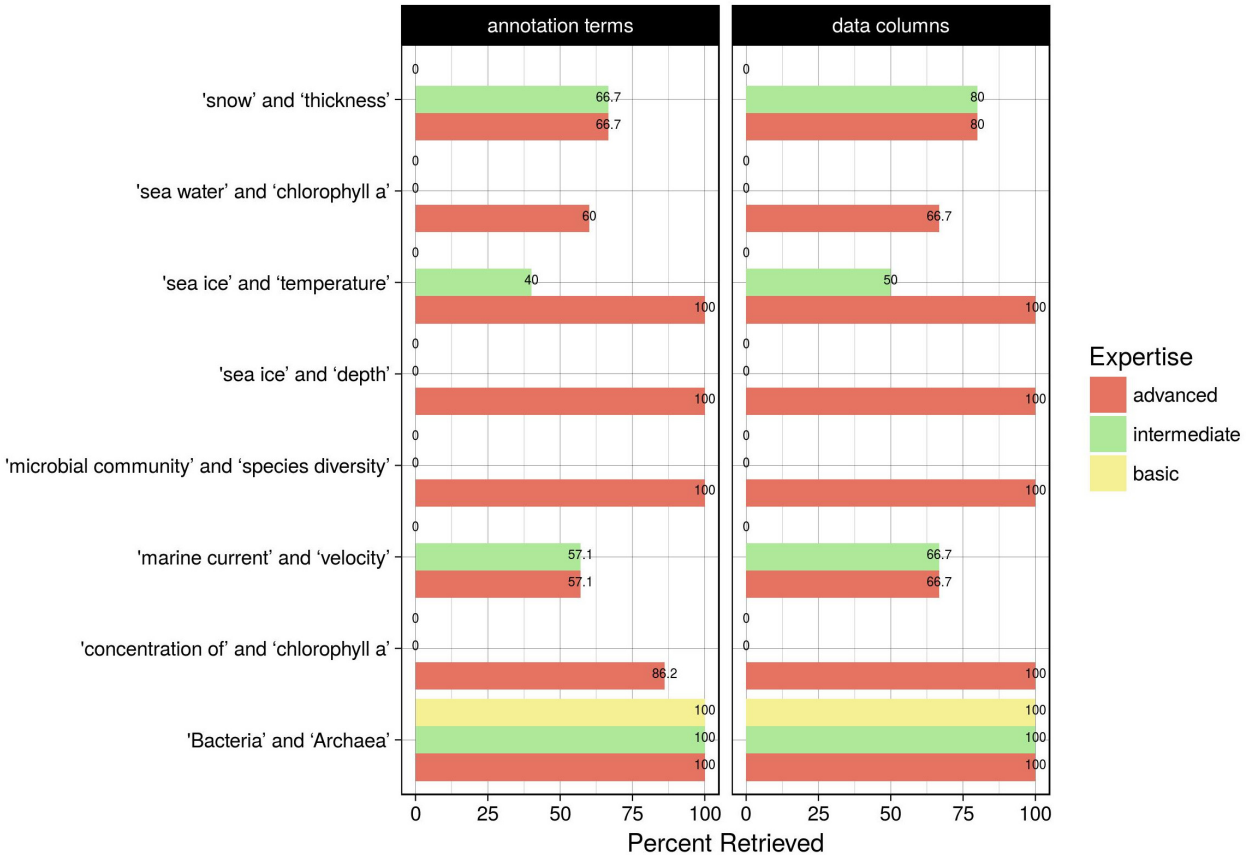


Figure 5: Analysis of querying expertise required to obtain data matrix columns and annotations when querying for data about subclasses of a term AND another term.

Querying parts of annotation

test if there is a similar bug to the the Querying exclusive AND annotations case.

perhaps it would be better to use or `query_for_parts_associated_with_input_class` because these cases would be easier scenarios to fit into the table.

//Discussion

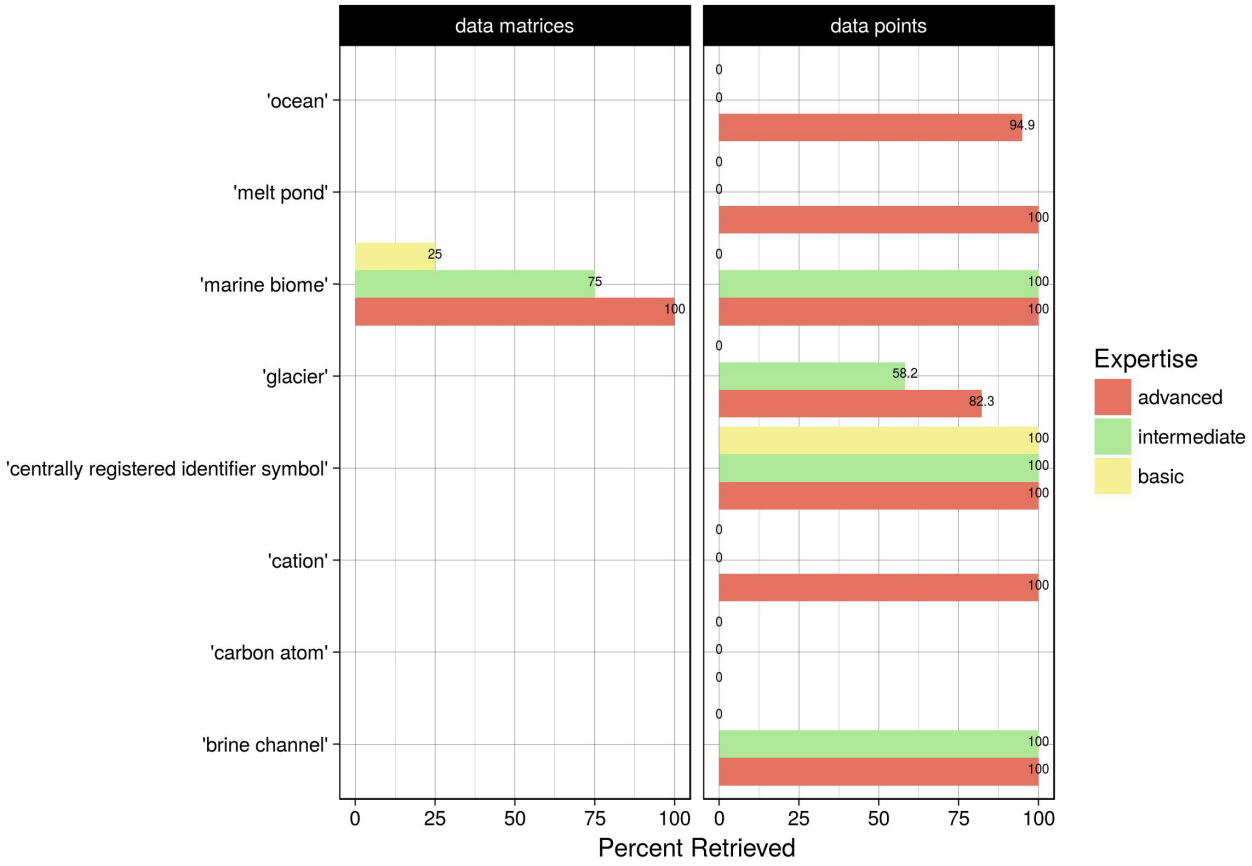


Figure 6: Analysis of querying expertise required to obtain data matrices and data points when querying for data about associated with parts of a term.

//wrap these in an example like the concentration of chlorophyll a in seawater. talking about how the post-compositional annotation model can help to facilitate the retrieval of data at various levels of granularity.

Ontological representation of real world phenomena

PCO contributions & Plankton Ecology

give some interesting examples of the PCO term contributions I wanted to make

In the discussion secion for this:

Address: How well does an ontology represent real world phenomena, find example of how people talk about it. Find key features of these dynamics and show how much my ontology captures. use some of my bloom pco ideas to do this. specifically the story like seaice melt triggers bloom.

Example of Post Compositional Data Annotation with Ontology Terms

change this competency question example to be about how to annotate data which is about a **marine environment determined by a diatom community** or a **marine environment determined by a diatom community bloom** instead of being about I intend to create these classes.

Vocamp Virtual Glacial Hackathon

vocamp:

> VoCamp is a series of informal events where people can spend some dedicated time creating lightweight vocabularies/ontologies for the Semantic Web/Web of Data.

Virtual-Hackahon-on-Glacier-topic

//to be held on Feb. 2nd. I should have an example of moving snow and ice related data ready to demonstrate by then.

Hackathon competency question

if ppl create semantic polar model how much would that be upset when other onto dev also work on it, stability of system if they are to fragile to change, its not good. Did the hackathon fundematally change how were modeling glaciers. Show envo graph and contributions from hackathon. were the producs of the hackathon extensions or … is this going to

defend in discussion to evaluate ccs the permutatois is immense needs colabs and human groups. Defend myself here.

would this completely change how it's modeled .

Talk about the hackathon as the method, the results are revision, evaluation output of that : did glacier hackathion table

does envo not rep glacier semantics, answer no were ahead of the game. Nothing suggested reqires differt technologies is this solution robust and extensible to cope with future semantic develop or alt semantic dve paradigmns,

joined other ppl see best state of knowlede, mention ppl syrie, ruth lewis, look at outcomes, were in the running robust and ther are efforts to align thigns, but our thing is probably interoperable.

Pier doesn't like ths. > To what extend are polar semantics encoded into the environment ontology accepted by domain experts such as glaciologists.

I was able to test this question during the vocamp glacial hackathon, organized by … . During the "hackathon" a variety of scientists and domain experts participated in a collaborative semantics research session with the objective of common vocabulary/ontology for glaciers and related concepts to be used and made interoperable between various existing ontologies.

tuned into and deliberated expert knowledge

33 terms formation processes standing stock ice removal from glaciers were the focus of the work. Have a look at the gloceries for the citations for where ruth got these terms.

Pier didn't like this talble it fornow. **table of how many and which of those terms are semantically represented in ENVO?** perhaps a column for number of links to constituents? Pier says yes.

Final envo representation relative to vocamp consensus diagram.

Polar terms	included in ENVO (Y/N or purl)	% axiomatic links captured in ENVO

In semantic research the term "unpacking" refers to the disambiguation of terms, clarifying and separating out ambiguous or overlapping terminology to arrive at a single set of terms and definitions.

could also make a table about how much of ENVO is represented the BS diagram from the east coast group, and make a case about how we didn't do so much of their BS.

Overall many of the polar terms added to ENVO in the course of this work were generally accepted in a consensus of the domain experts. Work sourced from this hackathon was added to the environment ontology in the … release.

to the supplemental for this add the list of participants, and perhaps the final diagram.

classify ice caps … as glaciers (things not just on land?) We'll fix this in envo. Use the DOI for the hackathon repostory to reference this

Discussion

//Discussion In my masters thesis work I have devised a semantic data annotation and querying schema. It allows for the phenomena inhering in data, to be represented and searched in the same way as ontology classes. Annotating data to be semantically inter-operable with existing ontologies, allows us to ask questions of interdisciplinary data, making use of the connections between phenomena encoded within ontologies.

//Discussion? In my masters thesis work I have been writing scripts to assemble and query a demonstration datastore comprised of semantically annotated AWI data. As a part of my proposed work, I would create a human and machine-readable web accessible endpoint to host a variety of AWI data, as well as a the semantic search tools to facilitate querying it.

Model polar datastore

annotated with ontology terms

queryable using semantic web technologies.

makes use of post-compositional style data annotation terms.

Querying Semantically Annotated Data using polar semantics to annotate AWI Polar data in a machine-readable way. This allows for knowledge to be captured in a data querying

Creating Classes vs post compositional annotation for data annotation

Interconnecting genomic and environmental data via ontology discussion

Ontology guided data assembly for ecological analysis discussion

Connecting information contained within ontology terms to the term authors discussion

Connecting datasets and publications about an ontology term discussion

Interconnecting stated and unstated knowledge via an ontology knowledge graph discussion

Mobilizing ontology annotated data discussion

What strategies for using ontology semantics be used to post compositionally annotate data aid with in mobilization of such data.

Ontology development vs using existing semantics to mobilize data.

Data annotation and mobilization

look for data about “concentration of chlorophyll a in seawater”

v.s. ‘concentration of’ and ‘chlorophyll a’ and ‘seawater’

outlook for semantic system to automate the annotation of data for submission to an interoperable datastore and the querying of such a datastore.

1) Talk about pre vs post composition of ontology terms using concentration of chlorophyll in seawater as an example. talk about the way we modeled the datastore in owl would allow for a query either way (one precomposed concentraion of chlorophyll in seawater vs the post composition version) Show the query case required to retrieve all the information contained within this this axiom

//Maybe don’t present this as a new model for annotation but instead just example datastore using owl style annotations which support querying. //Discussion In this work we present a novel semantic data annotation model. Semantics have been used to represent data … //TODO FIND REFS. In this model data annotations are composed of terms from the OBO Foundry. Data annotations are written in The RDF turtle specification, and structured as nested owl classes. Annotating the data as owl classes ensures parity to the OBO ontologies. This enables us to perform sparql queries on the annotated data in the same manor as would be done to query OBO Foundry ontologies.

In order to emulated owl code written in RDF, we chose the turtle RDF format for its ability to nest blank nodes within strings of triples.

//ADD THE is about property in the data model, it could also be cool to have a vue figure which explains the workflow.

The creation of ontology classes involves the composition of axioms, the links between classes, which are assembled from other preexisting ontology classes and relational properties. In ontology development this is refereed to as precomposition, which has the effect of taking a set of ontology classes and properties and joining them together in a specific way and assigning this assemblage to be a novel class.

The proposed semantic data annotation model allows for this process to be done in reverse. This is not necessary when an appropriate term for annotation already exists, however, in cases where the appropriate annotation term is lacking, it can be created from a combination of other terms. This practice, referred to as “post composition”, enables a user to annotate their data with axioms that comprise a non existent ontology term. By writing the data annotations as owl classes, they are functionally equivalent to existing ontology classes, in terms of their ability to be searched for using a sparql query.

This allows for the phenomena inhering in data, to be represented in a machine readable semantic layer prior to their incorporation as ontology terms.

The model makes use of owl equivalence classes, to structure the annotation as the intersection (and) and or union (or) of post compositionally annotated classes.

Thus the proposed data annotation model will allow for users, who are not ontologists, to post compositionally annotate their data. //ADD section about how I'll write a tool to automate this in the outlook.

Ontological representation of real world phenomena discussion

Vocamp Virtual Glacial Hackathon discussion

An example of the semantic clarification that took place during the "hackathon" was coming to a consensus definition of *ablation*. Ontologies take an agnostic stance when representing knowledge which has multiple definitions or which pertains to competing hypothesis. In the *ablation* example the NOAA National Weather Service Glossary 2009 [65] stipulates the restriction that only melting and evaporation processes contribute to ablation. The Cogley et al. IACS-UNESCO Glacier Mass Balance 2011 [66] definition however, refers to all processes which reduce the mass of a glacier. Specifically noting the inclusion of calving processes as significantly contributing to ablation processes.

In order to incorporate such discrepancies into a semantic knowledge graph a variety of approaches can be taken in parallel. A general *ablation* class can be created to include all the possible ice loss processes included in the various definitions of *ablation*. If users are attempting to mobilize data about a specific combination of ice loss process classes, they may post-compose a semantic annotation which includes the specific processes of interest as axioms. A post-compositional annotation describing data specifically about *ablation* due to melting ice and ice calving could for example be:

> 'ice loss process' and ('formed as result of' some ('icemelt' or 'ice calving process'))

If pre-composition is desired, in for example a case where a combination of specific ablation processes are commonly refereed to together as a set, a new term with a descriptive label could be created. A pre-compositional invocation of the example mentioned above would to create a descriptive term such as *calving and icemelt derived ablation*. Having a descriptive human readable label would facilitate the term's use for people such as domain experts or data stewards who are annotating data or describing a specific process. From a linked data perspective, both the pre-compositional and post-compositional annotations of the phenomena in question would make use of the same axiom (above), hence in terms of machine-readability and machine-searchability would be equivalent. This would facilitate the interoperation of data annotated both manually for example with a term such as *calving and icemelt derived ablation* and automatically for example by a semi-automated routine for post-compositionally annotating data, making use of existing terms.

Ontology interoperation with existing web semantics resources

UNEP SDGIO

Despite operating within a semantically which is interoperable with the OBO Foundry the UNEP ontology is currently non queryable. Future work needs to be done to improve the way SDGIO purls are hosted via UNEP so

that they can be querable. This would allow for the the incorporation of data mobililzed via semantics to the UN SDGs to help achieve their objectives.

aligning obo with Sweet

aligning with DBPEDIA

Conclusion

This work has demonstrated that semantics can be used to mobilize polar data.

Outlook

consistent data structures for published data

outlook/discussion: a semantical data annotation system can work but the data needs to be consistently structured, have a common standard. This isn't too much to ask for, examples like neon national ecological observatory network, tara or osd have fixed standards for data and or metadata.

Demonstrate to research groups such as AWI the importance of consistently structuring data

Could maybe mention the new FAIR tools which are coming to evaluate if data is truly FAIR in terms of interoperability.

Polar knowledge application ontology

Tilman Satelite Data

Paper: Diatom Phenology in the Southern Ocean: Mean Patterns, Trends and the Role of Climate Oscillations. [67] //Associated with the plankton ecology project using Tillman satellite chlorophyll data and the plankton bloom ontology classes. **maybe move this to outlook for how this could be used in a system which draws from larger datasets (like my email to Anya explained)** Plus I could also talk about this paper as a motivation for the PCO terms, harvesting expert Domain knowledge for example from Anya's section doing the full cycle of the scientific with semantic questions.

also like what is outlined in the **PhD project proposal for a Helmholtz Information & Data Science School** proposal

form a hypothesis, test etc.

example of expert awi knowledge to harvest:

Harvest anya's expert knowledge into ontologies to capturing phenomena such as the "wineglass effect" distribution of mesoscale eddies, and the spacial relationships to carbon fluxes and deep sea export. Also link knowledge about the effects of cyclones, zooplankton migrations, Zooplanton traits (through work on the phenotype and trait ontology PATO).

add the stuff from the email to Anya.

I believe the use of ontologies and semantics data annotation could serve as a valuable tool to address broad biological questions, such as those in the Raes et al. 2017 paper, about which mechanism, temperature or productivity is responsible for marine microbial diversity.

An outlook for the goals presented in this work would be to semantically annotate a wide variety of interdisciplinary AWI datasets in order render such data machine-readable and query-able. This creates the possibility to ask deeper questions of large data sets to address fundamental biological questions such as: "Does microbial diversity coincide with temperature or with primary productivity sourced from nitrogen fixation?"

Such questions could be asked of semantically annotated and machine-readable genomic datasets, which contain basic metadata. Such data could be sourced from anywhere, in house AWI data or already published data, from a variety environmental locations. Working with a data publication service such as PANGAEA to host such data in an open machine-readable web accessible format would allow for complex queries and questions to be asked.

For example to address the aforementioned question, we would perform a query to gather all datasets which include temperature, functional genomic and taxonomic information. From this ecological analysis could be conducted such as testing if temperature tends to correlate with microbial diversity, or with samples enriched in nitrogen fixation genes. The intentional interoperability between the Environment Ontology and the Gene Ontology would facilitate a query for the latter.

FROM heibrids application

Despite the existence of large quantities of polar-relavant data generated by institutions such as AWI, such data is typically not published in machine-readable formats. Needed are methods to make disperate data work interoperably. Using highly-structured semantics provided by ontologies, data can be annotated, linked in machiene readable open data formats and mobilized in semantic web queries. Proposed is a project to utlize ontology semantics to interconnect polar data to facilitate the interconnection and mobilization of polar and genomic data.

Ontological semantic research, unpacking, categorizing and capturing polar knolwedge from experts at AWI would factor prominantly into the project. Semantic contributions would be made to the Environent Ontology, and related ontologies interoperable with the Gene Ontology.

Annotation of polar data using enhanced ontology semantics would be conducted to mobilize data at a fine level of granularity. Allowing for questions such as "What metabolic ecosystem services are provided by microbial

communities of sea ice?" Text-mining approaches would aslo be evaluated to faciliate the semantic capture of relevant data sets.

Creation of community standards for polar linked data

original MIxS paper: [68]

//talk about my contributions to the cryoMIxS project. Including work from my lab rotation. //no

//talk about the annotaion for alreal chlorophyll a, plus some of the other terms used to annotate some of the data in the datastore, will work toward the semantic axiomatization and definitions of terms which will be included in the cryoMIxS paper. //no

//talk about the need to Create community standards for polar linked data. and how this is being adddressed in the cryoMIxS extension paper.

Semantics as AWI Public Outreach

AWI Education & Communication

Contributions to semantic models such as those discussed in this work serve to improve AWI public outreach efforts to educate and communicate polar research outputs to the public. Dissemination of AWI knowledge has been demonstrated in this work via the contributions made to the open source encyclopedia Wikipedia. This was achieved by aligning the dpbedia ontology glacial semantics to those of ENVO, which were contributed during this work.

AWI DBPEDIA contributions Contributing semantic knowledge to the website Wikipedia in the form of an improved heirarchaly structure but aligning with ENVO.

Implement and talk about dpbedia contributions, hopefully they'll let me edit. My intention is to align dbpedia glacial semantics to those in ENVO, should be relatively quick and easy once I can edit.

[12] (citation for dpbedia)

Perparing future data for semantic querying, additional work would involve creating community genomic sequence submission standards, cryoMIxS builiding on the existing MIxS standards.

linking ontology mobilized data to United Nations' Sustainable Development Goals

Finally, an evaluation of the fitness for purpose of semantically captured knolwedge and data would be conducted, to address questions relevant to the United Nations' Sustainable Development Goals, Development Targets, and Essential Ocean Variables.

References

1. **Naafs BDA, Castro JM, Gea GAD, Quijano ML, Schmidt DN *et al.*** Gradual and sustained carbon dioxide release during aptian oceanic anoxic event 1a. *Nature Geoscience* 2016;9:135–139.

2. **Wang M, Overland JE**. A sea ice free summer arctic within 30 years? *Geophysical Research Letters* 2009;36:n/a–n/a.

3. **Hutchins DA, Fu F**. Microorganisms and ocean global change. *Nature Microbiology* 2017;2:17058.

4. **Soltwedel T, Bauerfeind E, Bergmann M, Budaeva N, Hoste E *et al.*** HAUSGARTEN: Multidisciplinary investigations at a deep-sea, long-term observatory in the arctic ocean. *Oceanography* 2005;18:46–61.

5. **Soltwedel T, Schauer U, Boebel O, Nothig E-M, Bracher A *et al.*** FRAM - FRontiers in arctic marine monitoring visions for permanent observations in a gateway to the arctic ocean. In: *2013 MTS/IEEE OCEANS - bergen*. IEEE. Epub ahead of print June 2013. DOI: 10.1109/oceans-bergen.2013.6608008.

6. **Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D *et al.*** An ontology for describing and synthesizing ecological observation data. *Ecological Informatics* 2007;2:279–296.

7. **Molloy JC**. The open knowledge foundation: Open data means better science. *PLoS Biology* 2011;9:e1001195.

8. **Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M *et al.*** The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 2016;3:160018.

9. **Richard Cyganiak, DERI, NUI Galway, David Wood, 3 Round Stones, Markus Lanthaler *et al.*** RDF 1.1 Concepts and Abstract Syntax. *RDF 1.1 Concepts and Abstract Syntax*. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ (2014, accessed 4 February 2018).

10. **Tim Berners-Lee**. World Wide Web Consortium (W3C). https://www.w3.org/ (accessed 4 February 2018).

11. **David Beckett, Tim Berners-Lee, W3C, Eric Prud'hommeaux, Gavin Carothers *et al.*** RDF 1.1 Turtle. https://www.w3.org/TR/turtle/ (2014, accessed 4 February 2018).

12. **Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R *et al.*** DBpedia: A nucleus for a web of open data. In: *The semantic web*. Springer Berlin Heidelberg. pp. 722–735.

13. **Smith B, Michael Ashburner, Rosse C, Bard J, Bug W *et al.*** The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 2007;25:1251–1255.

14. **Paulheim H**. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 2016;8:489–508.

15. **Arp R, Smith B, Spear AD**. *Building ontologies with basic formal ontology*. The MIT Press. Epub ahead of print August 2015. DOI: 10.7551/mitpress/9780262527811.001.0001.

16. Basic Formal Ontology (BFO) Home. http://basic-formal-ontology.org/ (accessed 4 February 2018).

17. Basic Formal Ontology (BFO). https://github.com/BFO-ontology/BFO (accessed 4 February 2018).

18. Oborel/obo-relations. *GitHub*. https://github.com/oborel/obo-relations (accessed 4 February 2018).

19. **Buttigieg P, Morrison N, Smith B, Mungall CJ, and SEL**. The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics* 2013;4:43.

20. **Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL *et al.*** The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*;7. Epub ahead of print September 2016. DOI: 10.1186/s13326-016-0097-6.

21. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H *et al.*** Gene ontology: Tool for the unification of biology. *Nature Genetics* 2000;25:25–29.

22. **Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL *et al.*** Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005;102:15545–15550.

23. **Henschel A, Anwar MZ, Manohar V**. Comprehensive meta-analysis of ontology annotated 16S rRNA profiles identifies beta diversity clusters of environmental bacterial communities. *PLOS Computational Biology* 2015;11:e1004468.

24. **Williams MJ, Ausubel J, Poiner I, Garcia SM, Baker DJ *et al.*** Making marine life count: A new baseline for policy. *PLoS Biology* 2010;8:e1000531.

25. **Rockström J, Steffen W, Noone K, Persson, Chapin FSI *et al.*** Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society*;14. Epub ahead of print 2009. DOI: 10.5751/es-03180-140232.

26. A communal catalogue reveals earth's multiscale microbial diversity. *Nature*. Epub ahead of print November 2017. DOI: 10.1038/nature24621.

27. **Thessen AE, Bunker DE, Buttigieg PL, Cooper LD, Dahdul WM *et al.*** Emerging semantics to link phenotype and environment. *PeerJ* 2015;3:e1470.

28. **Pinheiro P, Mcguinness D, O. Santos H**. Human-aware sensor network ontology: Semantic support for empirical data collection.

29. **Bauerfeind E, Kattner G, Ludwichowski K-U, Nöthig E-M, Sandhop N**. Inorganic nutrients measured on water bottle samples at AWI HAUSGARTEN during POLARSTERN cruise MSM29. Epub ahead of print 2014. DOI: 10.1594/PANGAEA.834685.

30. **Bauerfeind E, von Appen W-J, Soltwedel T, Lochthofen N**. Physical oceanography and current meter data from mooring TD-2014-LT. Epub ahead of print 2016. DOI: 10.1594/PANGAEA.861860.

31. **Nöthig E-M, Bauerfeind E, Metfies K, Simon S, Lorenzen C**. Chlorophyll a measured on water bottle samples during POLARSTERN cruise ARK-XXIV/2. Data Set; PANGAEA. Epub ahead of print 2015. DOI: 10.1594/PANGAEA.855799.

32. **Nöthig E-M, Bracher A, Engel A, Metfies K, Niehoff B *et al.*** Summertime plankton ecology in fram straita compilation of long- and short-term observations. *Polar Research* 2015;34:23349.

33. **Soppa MA, Peeken I, Bracher A**. Global chlorophyll "a" concentrations for diatoms, haptophytes and prokaryotes obtained with the Diagnostic Pigment Analysis of HPLC data compiled from several databases and individual cruises. Data Set; PANGAEA. Epub ahead of print 2017. DOI: 10.1594/PANGAEA.875879.

34. **Losa SN, Soppa MA, Dinter T, Wolanin A, Brewin RJW *et al.*** Synergistic exploitation of hyper- and multi-spectral precursor sentinel measurements to determine phytoplankton functional types (SynSenPFT). *Frontiers in Marine Science*;4. Epub ahead of print July 2017. DOI: 10.3389/fmars.2017.00203.

35. **Bauerfeind E, Nöthig E-M, Beszczynska A, Fahl K, Kaleschke L *et al.*** Biogenic particle flux at AWI HAUSGARTEN from mooring FEVI7. Data Set; PANGAEA. Epub ahead of print 2009. DOI: 10.1594/PANGAEA.714844.

36. **Bauerfeind E, Nöthig E-M, Beszczynska A, Fahl K, Kaleschke L *et al.*** Particle sedimentation patterns in the eastern fram strait during 20002005: Results from the arctic long-term observatory HAUSGARTEN. *Deep Sea Research Part I: Oceanographic Research Papers* 2009;56:1471–1487.

37. **Nicolaus M, Itkin P, Spreen G**. Snow height on sea ice and sea ice drift from autonomous measurements from buoy 2015S22, deployed during the Norwegian Young sea ICE cruise N-ICE 2015. Data Set; Alfred Wegener Institute, Helmholtz Center for Polar; Marine Research, Bremerhaven; PANGAEA. Epub ahead of print 2015. DOI: 10.1594/PANGAEA.846861.

38. **Nicolaus M, Hoppmann M, Arndt S, Hendricks S, Katlein C *et al.*** Snow height and air temperature on sea ice from Snow Buoy measurements. Epub ahead of print 2017. DOI: 10.1594/PANGAEA.875638.

39. **Ricker R, Krumpen T, Schiller M**. Sea ice thickness at Ice Camp 1 on 2013-09-01 (GEM2IceTh_DiveHole_IceStation1). Data Set; PANGAEA. Epub ahead of print 2017. DOI: 10.1594/PANGAEA.870689.

40. **Arndt S, Meiners KM, Ricker R, Krumpen T, Katlein C *et al.*** Influence of snow depth and surface flooding on light transmission through antarctic pack ice. *Journal of Geophysical Research: Oceans* 2017;122:2108–2119.

41. **Lange BA, Michel C, Beckers J, Casey JA, Flores H *et al.*** Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-1_10. Data Set; PANGAEA. Epub ahead of print 2015. DOI: 10.1594/PANGAEA.842359.

42. **Lange BA, Michel C, Beckers JF, Casey JA, Flores H *et al.*** Comparing springtime ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice from the lincoln sea. *PLOS ONE* 2015;10:e0122418.

43. **Lange BA, Michel C, Beckers J, Casey JA, Flores H *et al.*** Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-2_11. Data Set; PANGAEA. Epub ahead of print 2015. DOI: 10.1594/PANGAEA.842363.

44. **Steve Harris, Garlik, a part of Experian, Andy Seaborne, The Apache Software Foundation**. SPARQL 1.1 Query Language. *SPARQL 1.1 Query Language*. https://www.w3.org/TR/sparql11-query/ (2013).

45. Welcome to Python.Org. *Python.org*. https://www.python.org/ (accessed 4 February 2018).

46. Apache Any23 – Apache Any23 - Introduction. http://any23.apache.org/ (accessed 4 February 2018).

47. **Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness *et al.*** OWL Web Ontology Language Reference. https://www.w3.org/TR/owl-ref/ (2004, accessed 4 February 2018).

48. **Musen MA**. The protégé project. *AI Matters* 2015;1:4–12.

49. Protégé. https://protege.stanford.edu/ (accessed 4 February 2018).

50. **Ong E, Xiang Z, Zhao B, Liu Y, Lin Y *et al.*** Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res* 2017;45:D347–D352.

51. **Crossley RA, Gaskin DJH, Holmes K, Mulholland F, Wells JM *et al.*** Riboflavin biosynthesis is associated with assimilatory ferric reduction and iron acquisition by campylobacter jejuni. *Applied and Environmental Microbiology* 2007;73:7819–7825.

52. **Fuller SJ, McMillan DGG, Renz MB, Schmidt M, Burke IT *et al.*** Extracellular electron transport-mediated fe(III) reduction by a community of alkaliphilic bacteria that use flavins as electron shuttles. *Applied and Environmental Microbiology* 2013;80:128–137.

53. **Kozlowski WA, Deutschman D, Garibotti I, Trees C, Vernet M**. An evaluation of the application of CHEMTAX to antarctic coastal pigment data. *Deep Sea Research Part I: Oceanographic Research Papers* 2011;58:350–364.

54. **Franklin DJ, Poulton AJ, Steinke M, Young J, Peeken I *et al.*** Dimethylsulphide, DMSP-lyase activity and microplankton community structure inside and outside of the mauritanian upwelling. *Progress in Oceanography* 2009;83:134–142.

55. **Zindler C, Peeken I, Marandino CA, Bange HW**. Environmental control on the variability of DMS and DMSP in the mauritanian upwelling region. *Biogeosciences* 2012;9:1041–1051.

56. **Soppa M, Hirata T, Silva B, Dinter T, Peeken I *et al.*** Global retrieval of diatom abundance based on phytoplankton pigments and satellite data. *Remote Sensing* 2014;6:10089–10106.

57. **Cheah W, Taylor BB, Wiegmann S, Raimund S, Krahmann G *et al.*** Photophysiological state of natural phytoplankton communities in the south china sea and sulu sea. *Biogeosciences Discussions* 2013;10:12115–12153.

58. **Trimborn S, Hoppe CJ, Taylor BB, Bracher A, Hassler C**. Physiological characteristics of open ocean and coastal phytoplankton communities of western antarctic peninsula and drake passage waters. *Deep Sea Research Part I: Oceanographic Research Papers* 2015;98:115–124.

59. **Sauzède R, Claustre H, Jamet C, Uitz J, Ras J *et al.*** Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications. *Journal of Geophysical Research: Oceans* 2015;120:451–470.

60. **Zindler C, Bracher A, Marandino CA, Taylor B, Torrecilla E *et al.*** Sulphur compounds, methane, and phytoplankton: Interactions along a northsouth transit in the western pacific ocean. *Biogeosciences* 2013;10:3297–3311.

61. **Peloquin J, Swan C, Gruber N, Vogt M, Claustre H *et al.*** The MAREDAT global database of high performance liquid chromatography marine pigment measurements. *Earth System Science Data* 2013;5:109–123.

62. **Werdell PJ, Bailey S, Fargion G, Pietras C, Knobelspiesse K *et al.*** Unique data repository facilitates ocean color satellite validation. *Eos, Transactions American Geophysical Union* 2003;84:377.

63. **Bracher A, Taylor MH, Taylor B, Dinter T, Röttgers R *et al.*** Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations. *Ocean Science* 2015;11:139–158.

64. **Uitz J, Claustre H, Morel A, Hooker SB**. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal of Geophysical Research*;111. Epub ahead of print 2006. DOI: 10.1029/2005jc003207.

65. **NWS Internet Services Team**. Glossary - NOAA's National Weather Service. *National Weather Service Glossary*. http://w1.weather.gov/glossary/ (2009).

66. **Cogley J, Hock R, Rasmussen L, Arendt A, Bauder A *et al.*** Glossary of Glacier Mass Balance and Related Terms. http://unesdoc.unesco.org/images/0019/001925/192525e.pdf (2011).

67. **Soppa M, Völker C, Bracher A**. Diatom phenology in the southern ocean: Mean patterns, trends and the role of climate oscillations. *Remote Sensing* 2016;8:420.

68. **Yilmaz P, Kottmann R, Field D, Knight R, Cole JR *et al.*** Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* 2011;29:415–420.

Appendices

Model polar datastore annotations

// add github/…/datastore tables here

`inorganic_nutrients.csv`

Data is about a: 'marine water body'

column	post compositional annotation:
Event	'centrally registered identifier' and 'is about' some ('observing process' or 'specimen collection process')
Date	temporal instant
Time	
Latitude	latitude coordinate measurement datum
Longitude	longitude coordinate measurement datum
Elevation	(elevation or depth) and ('inheres in' some 'marine water body')
Water Depth	(elevation or depth) and ('inheres in' some 'marine water body')
Nitrate	'concentration of']('http://purl.obolibrary.org/obo/PATO_0000033) and ('inheres in' some (nitrate and ('part of' some 'sea water')))
Nitrite	'concentration of' and ('inheres in' some (nitrite and ('part of' some 'sea water')))
Silicate	'concentration of' and ('inheres in' some ('silicate(4-)' and ('part of' some 'sea water')))
Phosphate	'concentration of' and ('inheres in' some (phosphate and ('part of' some 'sea water')))
Ammonium	'concentration of' and ('inheres in' some (ammonium and ('part of' some 'sea water')))

`physical_oceanography.csv`

Data is about a: 'marine current'

column	post compositional annotation:
Date	temporal instant
Time	
Gear Identification Number	('centrally registered identifier' and 'manufactured product')
Water Depth	(elevation or depth) and ('inheres in' some 'marine water body')

column	post compositional annotation:
Pressure	pressure and (‘inheres in’ some ‘sea water’)
Temperature	temperature and (‘inheres in’ some ‘sea water’)
Salinity	osmolarity and (‘inheres in’ some (‘salt’ and (‘part of’ some ‘sea water’)))
Horizontal Current Velocity	velocity and (‘inheres in’ some ‘marine current’ and ‘has quality’ some ‘horizontal’)
Current Direction	direction
East-west Current Velocity	velocity and (‘inheres in’ some ‘marine current’)
North-south Current Velocity	velocity and (‘inheres in’ some ‘marine current’)
Oxygen	‘concentration of’ and (‘inheres in’ some (dioxygen and (‘part of’ some ‘sea water’)))

`chlorophyll_a.csv`

Data set is about: ‘chlorophyll a’ and (‘part of’ some ‘marine water body’)

column	post compositional annotation:
Event	‘centrally registered identifier’ and ‘is about’ some (‘observing process’ or ‘specimen collection process’)
Date Time	temporal instant
Latitude	latitude coordinate measurement datum
Longitude	longitude coordinate measurement datum
Elevation	(elevation or depth) and (‘inheres in’ some ‘marine water body’)
Water Depth	(elevation or depth) and (‘inheres in’ some ‘marine water body’)
Chlorophyll A	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some ‘sea water’)))

`global_chlorophyll_a.csv`

Data is about: ‘chlorophyll a’ and (‘part of’ some ‘marine water body’)

column	post compositional annotation:
Ordinal Number	‘categorical label’ and ‘is about’ some specimen
Date Time	temporal instant
Latitude	latitude coordinate measurement datum
Longitude	longitude coordinate measurement datum
Water Depth	(elevation or depth) and (‘inheres in’ some ‘marine water body’)
Total Chloro-phyll A	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some ‘sea water’)))
Diatom Chloro-phyll A	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some Bacillariophyta)))
Haptophyte Chloro-phyll A	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some Coccolithales)))
Prokaryote Chloro-phyll A	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some Bacteria)))
Database Cross Reference	hasDbXref

`biogenic_particle_flux.csv`

Data is about: ‘material transport process’ and (‘has input’ some ‘marine snow’)

column	post compositional annotation:
Water Depth	(elevation or depth) and (‘inheres in’ some ‘marine water body’)

column	post compositional annotation:
Date Time	‘zero-dimensional temporal region’
Date Time End	‘zero-dimensional temporal region’
Duration	‘one-dimensional temporal region’
Sample Label	‘categorical label’ and ‘is about’ some specimen
Seston Flux	flux and (‘inheres in’ some ‘marine snow’)
Calcium Carbonate Flux	flux and (‘inheres in’ some (‘part of’ some ‘marine snow’) and (‘composed primarily of’ some (‘calcium carbonate’ and ‘part of’ some ‘organic molecular entity’)))
Particulate Organic Carbon Flux	flux and (‘inheres in’ some (‘part of’ some ‘marine snow’) and (‘composed primarily of’ some (‘carbon atom’ and ‘part of’ some ‘organic molecular entity’)))
Particulate Organic Nitrogen Flux	flux and (‘inheres in’ some (‘part of’ some ‘marine snow’) and (‘composed primarily of’ some (‘nitrogen atom’ and ‘part of’ some ‘organic molecular entity’)))
Particulate Silicon Flux	flux and (‘inheres in’ some (‘part of’ some ‘marine snow’) and (‘composed primarily of’ some (‘silicon atom’ and ‘part of’ some ‘organic molecular entity’)))

`snow_height.csv`

Data is about: ‘first year ice’

column	single term annotation	need to create class	post compositional annotation:
Date			‘temporal instant’
Time			
Latitude			‘latitude coordinate measurement datum’
Longitude			‘longitude coordinate measurement datum’
Snow Height Sensor 1	snow thickness	Part of MIxSPR	thickness and (‘inheres in’ some ‘snow’)
Snow Height Sensor 2	snow thickness	Part of MIxSPR	thickness and (‘inheres in’ some ‘snow’)
Snow Height Sensor 3	snow thickness	Part of MIxSPR	thickness and (‘inheres in’ some ‘snow’)
Snow Height Sensor 4	snow thickness	Part of MIxSPR	thickness and (‘inheres in’ some ‘snow’)
Snow Height Mean	snow thickness	Part of MIxSPR	‘expected value’ and ‘is about’ min 2 (‘data item’ and ‘is about’ some (thickness and ‘inheres in’ some ‘snow’))
Atmospheric Pressure	atmospheric pressure	Part of MIxSPR	pressure and (‘inheres in’ some atmosphere)
Air Temperature			temperature of air

column	Single term annotation	need to create class	post compositional annotation:
Ice Temperature			temperature and (‘inheres in’ some ‘sea ice’)

`influence_snow_depth.csv`

Data is about: ‘physical quality’ and (‘inheres in’ some (‘marine water body’ and (‘adjacent to’ some ‘sea ice’)))

column	Single term annotation	need to create class	post compositional annotation:
Date Time			temporal instant
Latitude			latitude coordinate measurement datum
Longitude			longitude coordinate measurement datum
Relative Distance X			distance and (‘inheres in’ some ‘sea ice’)
Relative Distance Y			distance and (‘inheres in’ some ‘sea ice’)
Sea Ice Thickness	sea ice thickness	yes part of cryoMIxSPR	thickness and (‘inheres in’ some ‘sea ice’)
Signal Strength			‘degree of illumination’ and (‘inheres in’ some ‘sea ice’)
Database Cross Reference			hasDbXref

`ice_algal_chlorophyll_myi.csv`

Data is about: (‘chlorophyll a’ and ‘multiyear ice’)

column	post compositional annotation:
Identification	‘categorical label’ and ‘is about’ some specimen
Site	site
Sea Ice Type	‘categorical label’ and ‘is about’ some ‘multiyear ice’
Ice Or Snow Depth	depth and (‘inheres in’ some (‘multiyear ice’ or ‘snow’))
Minimum Ice Or Snow Depth	depth and (‘inheres in’ some (‘multiyear ice’ or ‘snow’))
Maximum Ice Or Snow Depth	depth and (‘inheres in’ some (‘multiyear ice’ or ‘snow’))
Chlorophyll A Concentration	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some ‘sea water’)))
Areal Chloro- phyll A	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some (‘sea water’ and ‘part of’ some ‘liquid planetary surface’))))
Salinity	osmolarity and (‘inheres in’ some (‘salt’ and (‘part of’ some meltwater)))
Ice Or Snow Temperature	temperature and (‘inheres in’ some (‘multiyear ice’ or ‘snow’))
Brine Volume	‘volume’ and (‘inheres in’ some ‘brine’)
Texture	‘morphology’ and (‘inheres in’ some ‘multiyear ice’)
Sea Ice Type Portion	‘fiat object part’ and ‘part of’ some ‘multiyear ice’

column	post compositional annotation:

`ice_algal_chlorophyll_fyi.csv`

Data is about: (‘chlorophyll a’ and ‘first year ice’)

column	post compositional annotation:
Identification	‘categorical label’ and ‘is about’ some specimen
Site	site
Sea Ice Type	‘categorical label’ and ‘is about’ some ‘first year ice’
Ice Or Snow Depth	depth and (‘inheres in’ some (‘first year ice’ or ‘snow’))
Minimum Ice Or Snow Depth	depth and (‘inheres in’ some (‘first year ice’ or ‘snow’))
Maximum Ice Or Snow Depth	depth and (‘inheres in’ some (‘first year ice’ or ‘snow’))
Chlorophyll A Concentration	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some ‘sea water’)))
Areal Chloro- phyll A	‘concentration of’ and (‘inheres in’ some (‘chlorophyll a’ and (‘part of’ some (‘sea water’ and ‘part of’ some ‘liquid planetary surface’))))
Salinity	osmolarity and (‘inheres in’ some (‘salt’ and (‘part of’ some meltwater)))
Texture	‘morphology’ and (‘inheres in’ some ‘first year ice’)
Sea Ice Type Portion	‘fiat object part’ and ‘part of’ some ‘first year ice’

column	post compositional annotation:

`molecular_function_bathyal.csv`

data is about some: 'molecular_function' and ('part of' some ('microbial community' and ('part of' some ('deep marine sediment' and ('part of' some 'marine bathyal zone biome')))))

column	post compositional annotation:
Molecular Function	'categorical label' and 'is about' some 'molecular_function'
Sample 1 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))
Sample 2 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))

`molecular_function_abyssal.csv`

data is about some: 'molecular_function' and ('part of' some ('microbial community' and ('part of' some ('deep marine sediment' and ('part of' some 'marine abyssal zone biome')))))

column	post compositional annotation:
Molecular Function	'categorical label' and 'is about' some 'molecular_function'
Sample 1 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))
Sample 2 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))

column	post compositional annotation:

`molecular_function_neritic.csv`

data is about some: 'molecular_function' and ('part of' some ('microbial community' and ('part of' some ('sandy sediment' and ('part of' some 'marine neritic benthic zone biome')))))

column	post compositional annotation:
Molecular Function	'categorical label' and 'is about' some 'molecular_function'
Sample 1 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))
Sample 2 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))
Sample 3 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))
Sample 4 Molecular Function Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'molecular_function')))

`cellular_components_bathyal.csv`

data is about some: 'cellular_component' and ('part of' some ('microbial community' and ('part of' some ('deep marine sediment' and ('part of' some 'marine bathyal zone biome')))))

column	post compositional annotation:
Cellular Components	'categorical label' and 'is about' some 'cellular_component'

column	post compositional annotation:
Sample 1 Cellular Compo-nents Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘cellular_component’)))
Sample 2 Cellular Compo-nents Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘cellular_component’)))

`cellular_components_abyssal.csv`

data is about some: ‘cellular_component’ and (‘part of’ some (‘microbial community’ and (‘part of’ some (‘deep marine sediment’ and (‘part of’ some ‘marine abyssal zone biome’)))))

column	post compositional annotation:
Cellular Components	‘categorical label’ and ‘is about’ some ‘cellular_component’
Sample 1 Cellular Compo-nents Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘cellular_component’)))
Sample 2 Cellular Compo-nents Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘cellular_component’)))

`cellular_components_neritic.csv`

data is about some: ‘cellular_component’ and (‘part of’ some (‘microbial community’ and (‘part of’ some (‘sandy sediment’ and (‘part of’ some ‘marine neritic benthic zone biome’)))))

column	post compositional annotation:
Cellular Components	‘categorical label’ and ‘is about’ some ‘cellular_component’

column	post compositional annotation:
Sample 1 Cellular Components Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'cellular_component')))
Sample 2 Cellular Components Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'cellular_component')))
Sample 3 Cellular Components Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'cellular_component')))
Sample 4 Cellular Components Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'cellular_component')))

`biological_process_bathyal.csv`

data is about some: 'biological_process' and ('part of' some ('microbial community' and ('part of' some ('deep marine sediment' and ('part of' some 'marine bathyal zone biome')))))

column	post compositional annotation:
Biological Process	'categorical label' and 'is about' some 'biological_process'
Sample 1 Biological Process Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'biological_process')))

column	post compositional annotation:
Sample 2 Biological Process Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘biological_process’)))

`biological_process_abyssal.csv`

data is about some: ‘biological_process’ and (‘part of’ some (‘microbial community’ and (‘part of’ some (‘deep marine sediment’ and (‘part of’ some ‘marine abyssal zone biome’)))))

column	post compositional annotation:
Biological Process	‘categorical label’ and ‘is about’ some ‘biological_process’
Sample 1 Biological Process Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘biological_process’)))
Sample 2 Biological Process Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘biological_process’)))

`biological_process_neritic.csv`

data is about some: ‘biological_process’ and (‘part of’ some (‘microbial community’ and (‘part of’ some (‘sandy sediment’ and (‘part of’ some ‘marine neritic benthic zone biome’)))))

column	post compositional annotation:
Biological Process	‘categorical label’ and ‘is about’ some ‘biological_process’
Sample 1 Biological Process Count	‘discrete random variable’ and (‘is about’ some (‘categorical label’ and (‘is about’ some ‘biological_process’)))

column	post compositional annotation:
Sample 2 Biological Process Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'biological_process')))
Sample 3 Biological Process Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'biological_process')))
Sample 4 Biological Process Count	'discrete random variable' and ('is about' some ('categorical label' and ('is about' some 'biological_process')))

`microbial_taxonomy_bathyal.csv`

*in the annotation have classes be an 'operational taxonomic unit matrix' (or part thereof) instead of a 'data matrix'. Make sure to add query cases for these in the query script.

data is about some: 'community species diversity' and ('inheres in' some ('microbial community' and ('part of' some ('deep marine sediment' and ('part of' some 'marine bathyal zone biome')))))

column	post compositional annotation:
Sample 1 Operational Taxonomic Unit Count	'discrete random variable' and ('is about' some ('community species diversity' and ('inheres in' some 'microbial community')))
Sample 2 Operational Taxonomic Unit Count	'discrete random variable' and ('is about' some ('community species diversity' and ('inheres in' some 'microbial community')))
Domain	(Bacteria or Archaea or Eukaryota)

column	post compositional annotation:
Phylum	phylum
Class	class
Order	order
Family	family
Genus	genus

`microbial_taxonomy_abyssal.csv`

*in the annotation have classes be an 'operational taxonomic unit matrix' (or part thereof) instead of a 'data matrix'. Make sure to add query cases for these in the query script.

data is about some: 'community species diversity' and ('inheres in' some ('microbial community' and ('part of' some ('deep marine sediment' and ('part of' some 'marine abyssal zone biome')))))

column	post compositional annotation:
Sample 1 Operational Taxonomic Unit Count	'discrete random variable' and ('is about' some ('community species diversity' and ('inheres in' some 'microbial community')))
Sample 2 Operational Taxonomic Unit Count	'discrete random variable' and ('is about' some ('community species diversity' and ('inheres in' some 'microbial community')))
Domain	(Bacteria or Archaea or Eukaryota)
Phylum	phylum
Class	class
Order	order
Family	family
Genus	genus

Competency Question supplemental

Interconnecting genomic and environmental data via ontology supplemental

//What cellular components differentiate deep marine biomes supplemental

Table 28: Full results of the relative genomic and transcriptomic abundances of oxidation-reduction process in various types of marine biomes.

label	marine abyssal zone biome	marine bathyal zone biome	marine neritic benthic zone biome
oxidation-reduction process	18.15	18.39	9.36
aerobic respiration	0.23	0.26	0.87
methanogenesis	0.11	0.12	0.06
ATP synthesis coupled electron transport	0.06	0.06	0.04
L-lysine catabolic process to acetate	0.06	0.07	0.01
respiratory electron transport chain	0.03	0.03	0.13
mitochondrial electron transport, NADH to ubiquinone	0.02	0.02	0.01
electron transport chain	0.02	0.02	0.05
fatty acid beta-oxidation using acyl-CoA dehydrogenase	0.02	0.02	0.01
anaerobic electron transport chain	0.01	0.01	0.00
glycogen biosynthetic process	0.00	0.00	0.01
aerobic electron transport chain	0.00	0.00	0.00
methanogenesis, from acetate	0.00	0.00	0.00
anaerobic glutamate catabolic process	0.00	0.00	0.00
fatty acid beta-oxidation	0.00	0.00	0.00
photosynthetic electron transport in photosystem II	0.00	0.00	16.08
heme oxidation	0.00	0.00	0.00
photosynthetic electron transport chain	0.00	0.00	1.38
mitochondrial electron transport, ubiquinol to cytochrome c	0.00	0.00	0.00

Question like: “What cellular components differentiate deep marine biomes?”

Example answer guided by querying data annotated with ENVO and GO ontology terms we find that subclasses of periplasmic space, are seperated in abyssal and bathyal deep marine biomes.

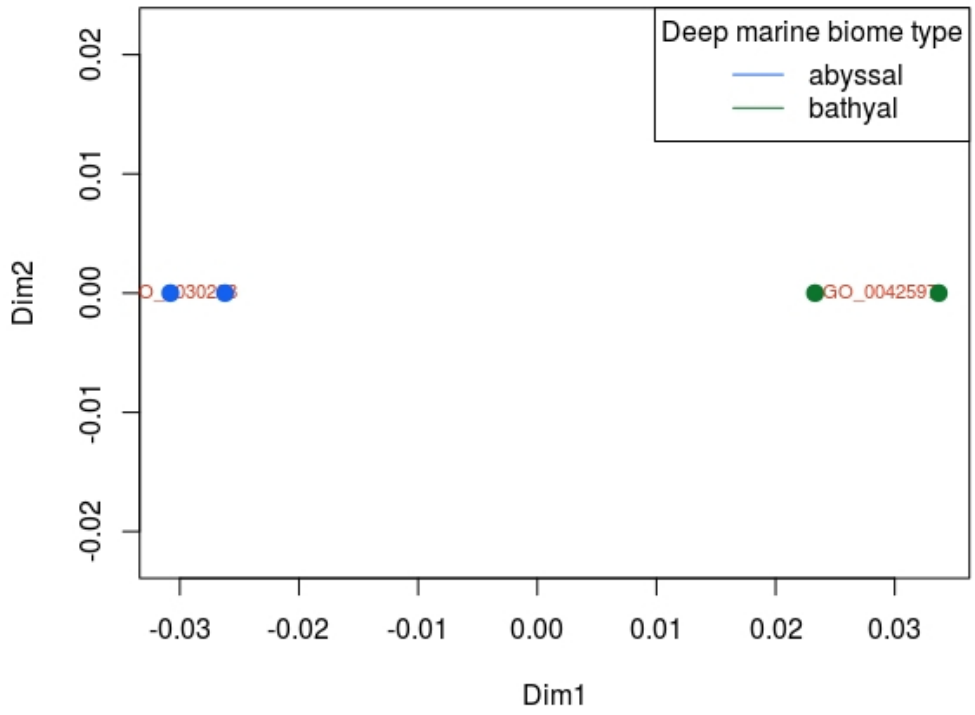


Figure 7: Example of ENVO deep marine biome classes differentiated in a principal coordinates analysis using GO cellular components subclasses of periplasmic space.

Differentiating terms: periplasmic space outer membrane-bounded periplasmic space

see `/home/kai/Desktop/grad_school/marmic/master_thesis/kblumberg_masters_thesis/datastore/competency_questions/connecting_GO_and_ENVO/analysis_cellular_comp_periplasm`

Ontology guided data assembly for ecological analysis supplemental

Connecting information contained within ontology terms to the term authors supplemental

Connecting datasets and publications about an ontology term supplemental

data annotation	reference doi	reference title
global chlorophyll a ENVO_00001999	10.1016/j.dsr.2011.01.008	An evaluation of the application of CHEMTAX to Antarctic coastal pigment data [53]

data annotation	reference doi	reference title
	10.1016/j.pocean.2009.07.011	Dimethylsulphide, DMSP-lyase activity and microplankton community structure inside and outside of the Mauritanian upwelling [54]
	10.5194/bg-9-1041-2012	Environmental control on the variability of DMS and DMSP in the Mauritanian upwelling region [55]
	10.3390/rs61010089	Global Retrieval of Diatom|[61] Abundance Based on Phytoplankton Pigments and Satellite Data [56]
	10.5194/bgd-10-12115-2013	Photophysiological state of natural phytoplankton communities in the South China Sea and Sulu Sea [57]
	10.1016/j.dsr.2014.12.010	Physiological characteristics of open ocean and coastal phytoplankton communities of Western Antarctic Peninsula and Drake Passage waters [58]
	10.1002/2014JC010355	Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications [59]

data annotation	reference doi	reference title
	10.5194/bg-10-3297-2013	Sulphur compounds,methane, and phytoplankton: Interactions along a north-south transit in the western Pacific Ocean [60]
	10.3402/polar.v34.23349	Summertime plankton ecology in Fram Strait-a compilation of long- and short-term observations [32]
	10.5194/essd-5-109-2013	The MAREDAT global database of high performance liquid chromatography marine pigment measurements [61]
	10.1029/2003EO380001	Unique data repository facilitates ocean color satellite validation [62]
	10.5194/os-11-139-2015	Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations [63]
	10.1029/2005JC003207	Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll [64]
influence snow depth ENVO_00001999	10.1002/2016JC012325	Influence of snow depth and surface flooding on light transmission through Antarctic pack ice [40]

Interconnecting stated and unstated knowledge via an ontology knowledge graph supplemental

Mobilizing ontology annotated data supplemental

Ontological representation of real world phenomena supplemental

//maybe add more of the pull request terms here?

Vocamp Virtual Glacial Hackathon supplemental

Python Script Descriptions Maybe include?