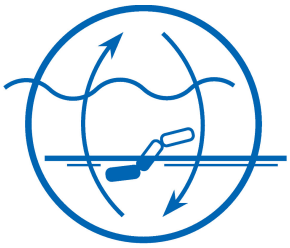
Interconnecting Arctic observatory data through machine-actionable knowledge representation: are ontologies fit for purpose?

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

a 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.

To illustrate how well structured ontology terms can be used to interconnect inderdiciplanry data, a model polar AWI polar datastore was created. Making use of interoperable ontology terms from the OBO Foundry, the data sets were annotated at both the column and csv file levels with ontology terms. Allowing for the datastore to be queryable

annotated with ontology terms

queryable using semantic web technologies.

makes use of post-compositional style data annotation terms.

Introduction

the information revolution has made the world increasingly volatile, uncertain, complex, and ambiguous [1]. Consequentailly

As a consequence, the biggest challenges faced by modern societies are not problems that can be solved; instead, they take the form of dilemmas that must be managed (Johansen 2007).[1]

Technological advances (e.g., DNA barcoding, automated species identification, artificial intelligence–assisted remote sensing) will revolutionize the collection of information for all domains of biodiversity knowledge, massively accelerating the rate of data capture.[1]

Data scarceness, limited description of patterns and processes, and gaps in theory are characteristic of all domains of ecology and evolution[1]

FROM [2] 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.

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 [3] 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.

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

//discuss some Arctic climate change research**

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

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: [5]

FRAM intro: [6]

** 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.

//transition between interdisciplinary observatory data and the need to have FAIR data.

FAIR the FAIR data guiding principles (machine-focused findability, accessibility, interoperability reusability) [7]

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

//just quick blurg To the effect of making science more open, challanges remain. **open science**

[8] 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.

//Part of the way to implemting FAIR data is to publish data in open and searable repositories such as Pangaea or BCO-DMO.

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.

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) http://www.whoi.edu/profile/bcodmo/

data publisher for a variety of NSF grants. Provides text based methods to search for data, but it's kindof hard to find stuff.

Alghouth publication of data with publishers is important it, more can be done. data needs to be interoperable I in FAIR.

to make the data interoperable we use ontologies

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 [9]. 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.

//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 [9]. 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) [10] [11] [12], to ensure they have a compatible hierarchical structure, and use logical relations from the Relations Ontology (RO) [13], 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: [14] [15]

The Environment Ontology (ENVO) represents expert knowledge about different types of environments[14][15]. 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. [14][15]

Gene Ontology go paper: [16]

GO frequently used to interpret omic data [16]. It has been used to do genomewide RNA expression profile data to compare samples based on shared biological pathways. [17]

The combination of GO and ENVO is less frequently attempted. [18]

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?".

ontologies use semantic web technologies. //semantic web JUST A TINY BIT NOT TOO MUCH!!! ## Semantics web

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) [19] 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). [20] The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). [21]

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.

//intro RDF is a framework for representing information in the semantic web, which is an extension of the world wide web //intro semantic technologies make use of the specifications of the World Wide Web Consortium (W3C) [20]

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 [22] 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 [22] > 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.

USE OF ontogies in other field semantics for other fields

Using ontogies to interconnect disparate data and information, to enable computational interrogation of models to reveal underlying relationships, exists within the medical domain. For example the Monarch Initiative, uses an

ontology-based strategy to integrate genotype–phenotype data from various sources and species, enabling users to explore phenotypic and genotypic relationships across species [23].

Example Ontology usage for environmental science

//fit this in somewhere that ontologies have the promise to help to integrate data of disperate sources such as omics data but it isn't quite there yet. //Either put this before or after the ontology section, [24] piers paper with the italians.

There is a need for the integration of integrating omics data to aid with present and future modelling initiatives. They conclude that such integration is not yet ready but are on the way. [24]

A communal catalogue reveals Earth's multiscale microbial diversity. //Uses EMPO a light-weight application ontology built on ENVO the Earth Microbiome Project Ontology [25] //good to have an example which demonstrates the utility of ENVO for an application ontology to provide utility.

//maybe cut this. //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 [26]

coming down the road

cite pier's paper [27] which mentions

> For example, a query such as "find all metagenomes collected from insects found in soil" requires data from many sources to be linked and freely available. Ontologies are essential to resolve queries such as this effectively, because data from different projects are often annotated with different levels of precision. For example, ENVO's environmental material hierarchy would allow this query to return results for samples collected in ENVO:loam, knowing that it is a subclass of ENVO:soil.

Say I've implemented a prelimary verion of this.

Efforts to link environmental and genomic data are underway by groups such as the hurwitzlab who have development for example ocean cloud commons

hurwitzlab projects

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

> The Ocean Cloud Commons (OCC) is a cloud-based resource and repository that allows researchers to query the Tara Oceans Expedition Data in the cloud

cite tara as …

Ontologies such as the Human-Aware Sensor Network Ontology (HASNetO) have been used to support the data management of a number of large-scale ecological monitoring activities [28]. Coming is the UN decade of ocean science for sustainable development 2021-2030, which will bring an an influx of data from proposed earth and ocean monitoring activities. These efforts will employ sensor networks which will generate vast quantities of data of heterogeneous types. Networks of sensors deployed to perform environmental monitoring activities will comprise a future "sensor web", intended to improve the ability to detect, monitor, and predict weather climate and the onset of natural hazards [29]. The ongoing efforts of the Open Geospatial Consortium (OGC) have seen the creation and development of community consensus standards for the sensor web [30]. Building upon semantic web technologies, the OGC has created the Sensor Web Enablement (SWE) standards which can be used as building blocks for the sensor web [31]. SWE standards enable developers to make all types of sensors, transducers and sensor data repositories discoverable, accessible and usable via the Web.

//sos part of SWE Sensor Observation Services (SOS) https://en.wikipedia.org/wiki/Sensor_Observation_Service http://www.opengeospatial.org/standards/sos

> is a web service to query real-time sensor data and sensor data time series and is part of the Sensor Web.
>
> The SOS standard is applicable to use cases in which sensor data needs to be managed in an interoperable way. This standard defines a Web service interface which allows querying observations, sensor metadata, as well as representations of observed features. Further, this standard defines means to register new sensors and to remove existing ones. Also, it defines operations to insert new sensor observations.

It also allows easy integration into existing Spatial Data Infrastructures or Geographic Information Systems.

Semantic Sensor Web (SSW) builds on SWE and extends them with Semantic Web technologies to provide enhanced descriptions and access to sensor data. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04557983

> Ontologies and other semantic technologies can be key enabling technologies for sensor networks because they will improve semantic interoperability and integration, as well as facilitate reasoning, classification and other types of assurance and automation not included in the Open Geospatial Consortium (OGC) standards.

SSN ontology https://doi.org/10.1016/j.websem.2012.05.003

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.

To prepare for future influx of sensor and observatory data as well as genomic data. making use of use of topical AWI polar data as example data. We evaluate if ontologies are fit for purpose to interconnect disperate environment and genomic datasets. evaluinting … (maybe don't need this) As well as evaluating the role of ontologies to facilitate interconnection of knowledge between authors, publications and datasets.

We evaluated all of this testing how ontologies perform against competency questions.

Materials and Methods

Model polar datastore creation

Example polar datastore

A polar data focused example datastore was assembled from freely available Alfred Wegener Institute for Polar and Marine Research datasets hosted by the data publisher PANGAEA [32], as well as unpublished genomic and transcriptomic datasets. Selected data sets were primary sourced from FRAM [6], and Hausgarten [5] observatory projects and programs, along with unpublished metagenomic and metatranscriptomic data provide by scientists for the Max Plank Institute for Marine Microbiology. A list of data sets included in the data store is provided in the supplementary section.

Datastore creation

The example datastore was created by converting comma separated value (csv) files into the Resource Description Framework (RDF) [19] specification format turtle [21], consisting of subject predicate object triples, using the Apache Anything To Triples (any23) command line *rover* tool [33]. Formating the example data using an RDF specification opens such data in a machine-readable manor, allowing for SPARQL Protocol and RDF Query Language (SPARQL) [34] queries to be performed.

Semantic Data Annotation

Semantic annotation of the example data was conducted in the RDF serialization turtle, drawing upon its blank node feature to facilitate scripting Web Ontology Language (OWL) [35] code in RDF. Annotations make use ontology terms from the Open Biomedical Ontologies (OBO) Foundry and Library [9], and follow the principals of best practices in ontology development as outlined in the book *BUILDING ONTOLOGIES WITH BASIC FORMAL ONTOLOGY* [10]. Ontology terms were accesses using the Ontobee linked data server for ontologies and their terms [36]. Ontology term annotation for the example polar datastore made frequent use of ENVO terms from the EnvoPolar subset which was primarily developed during the Ecotone, Polar express, Hot tub time machine ENVO releases.

Python scripts

Data SPARQL querying

Scripts were written in Python (Version 2.7.12) [37] and make use of the rdflib module (Version: 4.2.2) an RDF parsing python library [38] to parse the example RDF datastore, as well as assemble and execute SPARQL queries against the example datastore, or local RDF turtle serialization versions of local copies of ontologies. Additional SPARQL queries were preformed against the ontobee programmatic SPARQL endpoint available from http://sparql.hegroup.org/sparql/, provided by the He Group [36]. A list of the python scripts developed and utilized in the work is available in the supplementary Python scripts section.

Interconnecting genomic and environmental data via ontology methods

Metagenomic and metatranscriptomic data processing

Metagenomes and metatranscriptomes were processed using the following workflow. BBDuk from the BBTools suite was used for read quality control [39]. Ribosomal RNA filtering was conducted using the SortMeRNA tool [40]. Taxonomic classification of 16S ribosomal RNA from the metagenomic samples was done using the SINA alignment tool [41]. Open Read Frames (ORFs) were predicted from assembled metagenomes and metatranscriptomes using the FragGeneScan tool [42]. Predicted ORFs were annotated with protein families database (Pfam) [43] domains using the Ultra-fast Protein domain Classification tool (UProC) [44]. PFAM domains were mapped to GO terms using the Pfam2GO annotations page available from here [45].

Workflow

Retrieval of GO relative abundance data from ontology term annotated data was conducted using the `query_GO_annotation_of_data_files_csv_annotations_columns.py` script. Which made use of a list of benthic marine biome subclasses terms and a list of GO terms, which are subclasses of GO term of interest. Both subclasses lists were assembled using the `query_for_subclasses_of_input_purl.py` script.

Relative abundance calculations

Relative abundances of metagenomic and metatranscriptomic samples were calculated by querying for subclasses of a GO term of interest, querying for all GO terms in the corresponding GO hierarchy, biological process or molecular function, then dividing the relative abundances of individual terms by the sum of the total abundances of biological process or molecular function terms.

Principal coordinate analyses

Principal coordinate analysis (PCoA) on a data matrices of relative GO term abundances in ENVO annotated samples was conducted in R [46], using the vegan package [47]. Prior to PCoA analysis data was standardized using a hellinger tranformation and converted into a dissimilarity matrix using the Bray–Curtis method.

Ontology guided data assembly for ecological analysis methods

Subclasses of ontology terms included in the axioms of the hypothetical **environment determined by a phytoplankton community associated with sea-ice** term, were assembled using the `query_for_subclasses_of_input_purl.py` script, and concatended together in BASH. Data annotated as being about this assembled list of subclasses terms was retrieved from the datastore using the `query_annotation_of_data_files_data_or_columns_about_input.py` querying script.

A principal component analysis (PCA) was performed on the retrieved data using the vegan package [47]. Prior to PCA analysis, data was z-score standardized. For a descrition of the hypothetical term as well as a list of data

matrix columns utilized in the principal component analysis see the ontology guided data assembly for ecological analysis supplemental section.

Connecting information contained within ontology terms to the term authors methods

Percentages of Envo and EnvoPolar ontology terms annotated with a term editor or created_by relation referencing an Open Researcher and Contributor ID (ORCID) were calculated using the following workflow. Local turtle specification versions of the Envo and EnvoPolar ontologies were exported using Protégé [48][49]. Python scrips were used to SPARQL query the local turtle ontology versions, for total numbers of ENVO terms and numbers of terms annotated with a term editor or created_by relation an ORCID. Percentages of ontology terms with editor or created_by annotations were calculated from retrieved counts.

Connecting datasets and publications about an ontology term methods

The following was used to retrieve digital object identifiers (DOIs) of publications assocaited with datasets about parts of a marine biome. The python script `query_for_parts_associated_with_input_class.py` was used to query for ontology classes which are parts of or have parts which are a marine biome.

The resulting list of ontology which are parts assocaited with marine biomes were then queried against the local datastore for all data from data matrix columns which are annotated with a has database cross reference hasDbXref using the `query_data_set_references.py` python script. Results were post processed in R and libreoffice to create **Table 5**.

Reflections upon the VoCamp Glacier Ontology Hackathon methods

Simulations of semantic models proposed during the Feb 2, 2018 VoCamp Glacier Ontology Hackathon, utilizing or not utilizing the Basic Formal Ontology or Relations Ontology upper level semantic models, were evaluated for the extent to which annotated data would be retrieved by querying the example polar datastore. Retrieval of classes participating in subclasses of sea ice formation processes utilizing upper level semantic models, was conducted as follows. Subclasses of sea ice formation process terms were assembled using the `query_for_subclasses_of_input_purl.py` script. Subproperties of has participant were assembled using the `query_for_subproperties_of_input_purl.py` script. The results of classes which participate in in sea ice formation processes, **Table 6**, were discovered using the `query_for_classes_linked_by_input_classes_and_input_properties.py` script with the assembled subclasses of sea ice formation process and subproperties of has participant. The number of data items corresponding to classes obtained as a result of the previous script were retrieved using the `query_annotation_of_data_files_data_or_columns_about_input.py` script and counted based on the results.

The simulation for the retrieval of data lacking the Basic Fromal Ontology upper level semantic model was conducted using the same methods as above minus the query for subclasses of sea ice formation process, with only the term itself passed to the `query_for_classes_linked_by_input_classes_and_input_properties.py` script. The simulation for the retrieval of data lacking the Relations Ontology upper level semantic model was conducted without the use of the `query_for_classes_linked_by_input_classes_and_input_properties.py` script, directly passing the subclasses of sea ice formation process from the `query_for_subclasses_of_input_purl.py` script to the `query_annotation_of_data_files_data_or_columns_about_input.py` script.

Interconnecting stated and unstated knowledge via an ontology knowledge graph methods

The following was used to create a network from the Environent Ontology polar subset. The envoPolar subset from the ENVO v2017-08-22 Planetary ecology release, was used and is available from here. The EnvoPolar owl file was exported to the RDF specification turtle using the software Protégé [48][49]. Python scripts querying for all class and property terms in the envoPolar subset were used as inputs for the `query_for_classes_linked_by_input_classes_and_input_properties.py` script to obtain the connections between all classes in the ontology. The resulting output consisting of supject classes, properties linking subject classes to target classes, and target classes. Which was used to create a network in the program cytoscape [50].

Network parameters of the envoPolar subset of ENVO were calculated using the cytoscape network analyzer, treating the graph as directed. Figures for the distribution of shortest path lengths, average clustering coefficient, in degree distribution, out degree distribution, and betweenness centrality were generated in cytoscape.

Calculations of mean and median values for the in-degree, out-degree and shortest path length distributions were conducted in python using the statistics library Version: 1.0.3.5 [51] from distribution data output by the cytoscape network analysis.

Mobilizing ontology annotated data methods

The query expertise simulation for mobilizing data annotated with an exclusive AND intersection of two ontology terms was conducted as follows. Eight combinations of terms were used to query for data. For each combination, the `query_for_subclasses_of_input_purl.py` script was used to query for subclasses of the first term in the intersection combination. Modified versions of the `query_for_data_about_exclusive_and.py` script were run using the list of subclasses generated from the first term in the intersection combination, along with the second term. Retrieving data matrix annotations and columns annotated as being about the intersection of the subclasses of the first term with the second term.

The querying expertise simulation for mobilizing data annotated as being about parts associated with an ontology term was conducted as follows. Eight terms were queried for data annotated with parts associated therewith. For each term, the `query_for_parts_associated_with_input_class.py` script was

used to query for terms corresponding to associated parts of the input term. Modified versions of the `query_for_data_about_exclusive_and.py` script were run using the list of associated parts derived from the first script. Retrieving data matrices and data points annotated as being about parts associated with the input term.

For both experiments three modified versions of the data querying script were run to simulate data retrieval of of users of basic, intermediate and advanced querying proficiency. Percentages of retrieved data matrix columns, annotations, points and data matrices were calculated compared with the the results derived from an unmodified version of the `query_for_data_about_exclusive_and.py` or `query_for_data_about_exclusive_and.py` scripts which retrieved 100% of available data. A list of querying cases used to differentiate the levels of querying expertise is available in the mobilizing ontology annotated data supplemental section.

Ontological representation of plankton ecology related phenomena methods

Plankton ecology related ontology terms were prepared for the Environment Ontology (ENVO) [14][15], the Phenotypic Quality Ontology (PATO) [52], the Ecology Core Ontology ECOCORE [53] and the Population and Community Ontology (PCO) [54]. Following the principals of best practices in ontology development as outlined in the book *BUILDING ONTOLOGIES WITH BASIC FORMAL ONTOLOGY* [10].

Results

Interconnecting genomic and environmental data via ontology results

Using interoperable ontology semantics genomic and environmental data were mobilized to perform comparative analysis. The results of querying the example datastore for the relative genomic and transcriptomic abundance of sequences matching oxidation-reduction process genes, in various marine benthic biomes, are shown in **Table 2**.

Table 1: Selected results of relative genomic and transcriptomic abundances of oxidation-reduction processes in various types of marine benthic 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

label	marine abyssal zone biome	marine bathyal zone biome	marine neritic benthic zone biome
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

Analysis of these results show deep biome samples abyssal and bathyal, had double the relative abundances of non specific annotations to general oxidation-reduction reduction processes ~18%, relative to neritic samples. In contrast, neritic samples had three fold increases in aerobic respiration gene abundances relative to the deep samples.

Deep samples had nearly double methanogenesis gene abundances than those of neritic samples, while neritic samples had much greater relative respiratory electron transport chain abundances than deep samples.

Neritic samples had 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.

Comparisons of vitamin biosynthetic process genes of which are summarized in **Table 3**.

Table 2: Relative abundance of vitamin biosynthetic process genes in various types of marine benthic 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, we note the general trend that relative gene

abundance of vitamin biosynthetic process genes are higher than the relative transcriptomic abundance of neritic samples. For example the relative gene abundance of riboflavin genes was approximately 3.5 times greater in deep biome sample genomes than neritic sample transcriptomes.

To investigate other genomic features which may help to explain the differences in riboflavin abundance, we investigated transition metal ion binding and transport subclasses, 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 [55][56]. For full results of relative abundances of metal ion binding, and metal ion transport subclasses see supplement **Tables 10**, **11**. Querying for subclasses of metal ion binding we found that ferrous ion binding gene abundance is 0.02-0.03% in deep samples vs 0.00% in neretic. Furthermore deep sample ferrous iron transport gene abundance is double that of neritic sample, 0.04% vs 0.02%.

Results of our investigation into acid biosynthetic process differentiating marine benthis samples are as follows. We examined the results returned from a query for subclasses of the GO term organic acid biosynthetic process, in **Figure 1** a principal coordinate analysis conducted on relative gene and transcript abundance datasets annotated as various types of marine benthic biomes. These results indicate that there is a slight differentiation between the neritic and deep biome, abyssal and bathyal samples based on abundance of organic acid biosynthetic process genes and transcripts. From **Figure 1** we noted GO terms which have a more distinctive pull on the pcoa analysis, such as GO_0006535 cysteine biosynthetic process from serine, and GO_0006529 asparagine biosynthetic process.

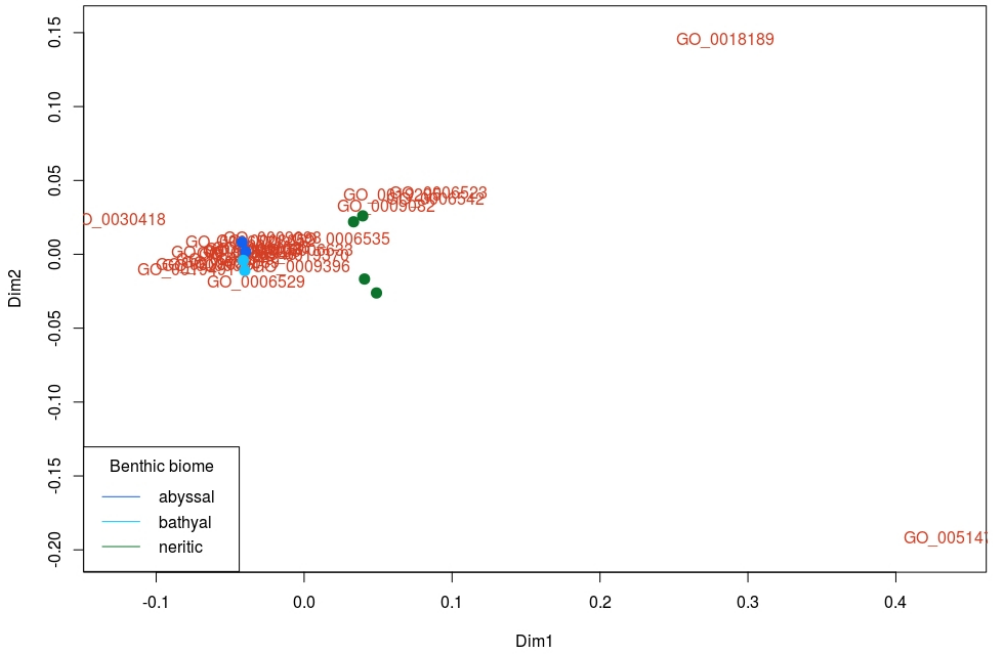


Figure 1: Principal coordinate analyses plot of relative genomic abundance of subclasses of organic acid biosynthetic processes in various marine benthic biomes.

Making use of the Gene Ontology hierarchy see supplement **Figure 9**, we noted that both of these terms are subclasses are types of cellular amino acid biosynthetic process, To investigate differentiating gene and transcript

abundance at a finer resolution another pcoa analysis was performed on the subclasses of cellular amino acid biosynthetic process. **Figure 2** shows a more clear separation between neritic and deep biome samples based on relative abundance of cellular amino acid biosynthetic process genes and transcripts.

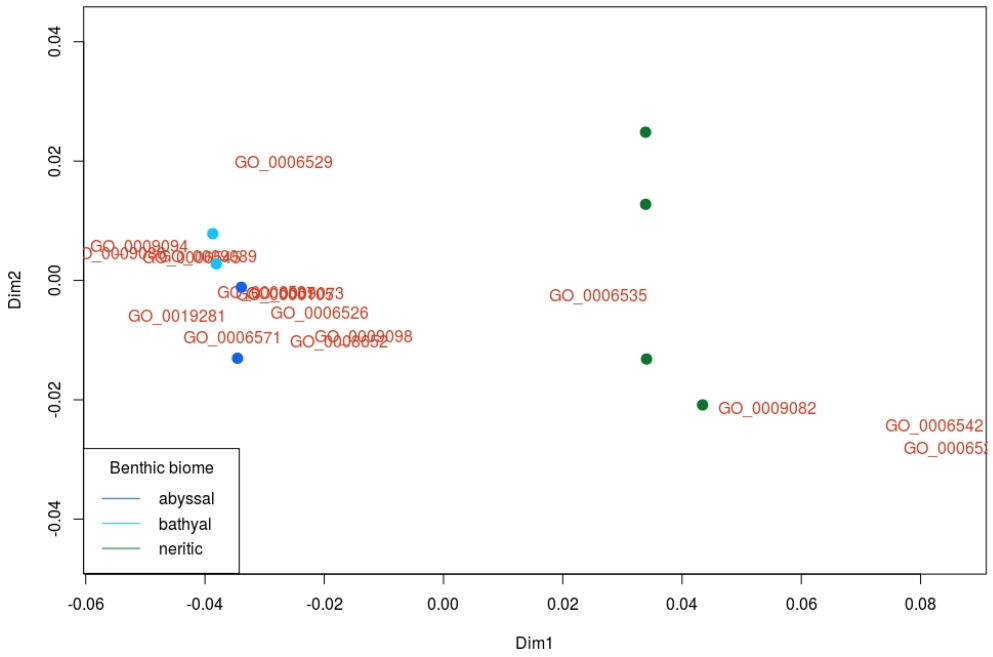


Figure 2: Principal coordinate analyses plot of relative genomic abundance of subclasses of cellular amino acid biosynthetic processes in various marine benthic biomes.

Examining the GO hierarchy for subclasses of cellular amino acid biosynthetic process, see supplement **Figure 9**, we can drill down into more specific terms further differentiate the benthic biome annotated samples.

Noting from the analysis that the term GO_0006535 cysteine biosynthetic process from serine has a lot of pull along pcoa dimension 1, we futher investigated the differences between between subclasses of serine family amino acid biosynthetic processes results shown in **Figure 3**. From this pcoa analysis of subclasses of serine family amino acid biosynthetic process we see a clear differentiation of neritic vs deep benthic biome samples. From this plot we learn that the relative gene abundance of GO_0006545 glycine biosynthetic process is more abundant in the deep samples, while GO_0006535 cysteine biosynthetic process from serine is more abundant in the neretic samples.

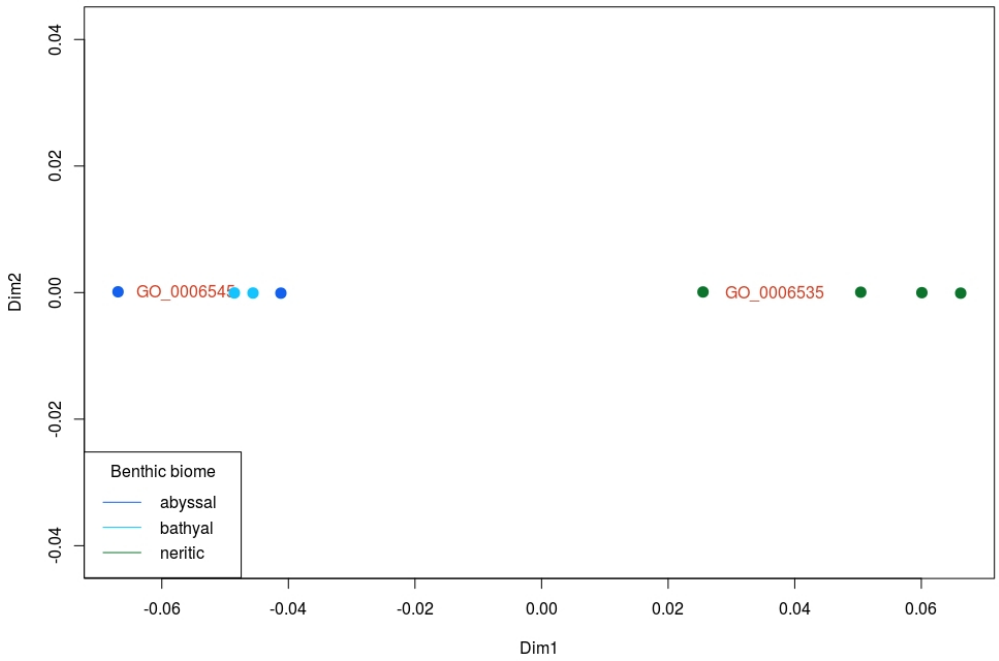


Figure 3: Principal coordinate analyses plot of relative genomic abundance of subclasses of serine family amino acid biosynthetic processes in various marine benthic biomes.

Ontology guided data assembly for ecological analysis results

We demonstrated how ontology terms can be used to facilitate the retrieval of data about a phenomena of interest, such as that which is represented by the hypothetical term **environment determined by a phytoplankton community associated with sea-ice**. By leveraging data annotated with terms which are included as axioms to perform an ecological analysis. The hypothetical term was defined as:

```
An environmental system which has its properties and dynamics determined by a
    phytoplankton community which is associated with sea-ice.
```

This hypothetical term would also include the following subclass axioms:

```
'environmental system determined by a community'

'determined by' some 'phytoplankton community'

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

'adjacent to' some 'sea ice'
```

Assembling a list of all the classes (and subclasses thereof) which are referenced by the axioms of this hypothetical term, we queried our example datastore for data about these terms. The assembled data comprises a dummy example as the the data are from different spatiotemporal locations. Next we performed a principal component analysis on this data to investigate which of these many environmental variables have the greatest loading on the

analysis.

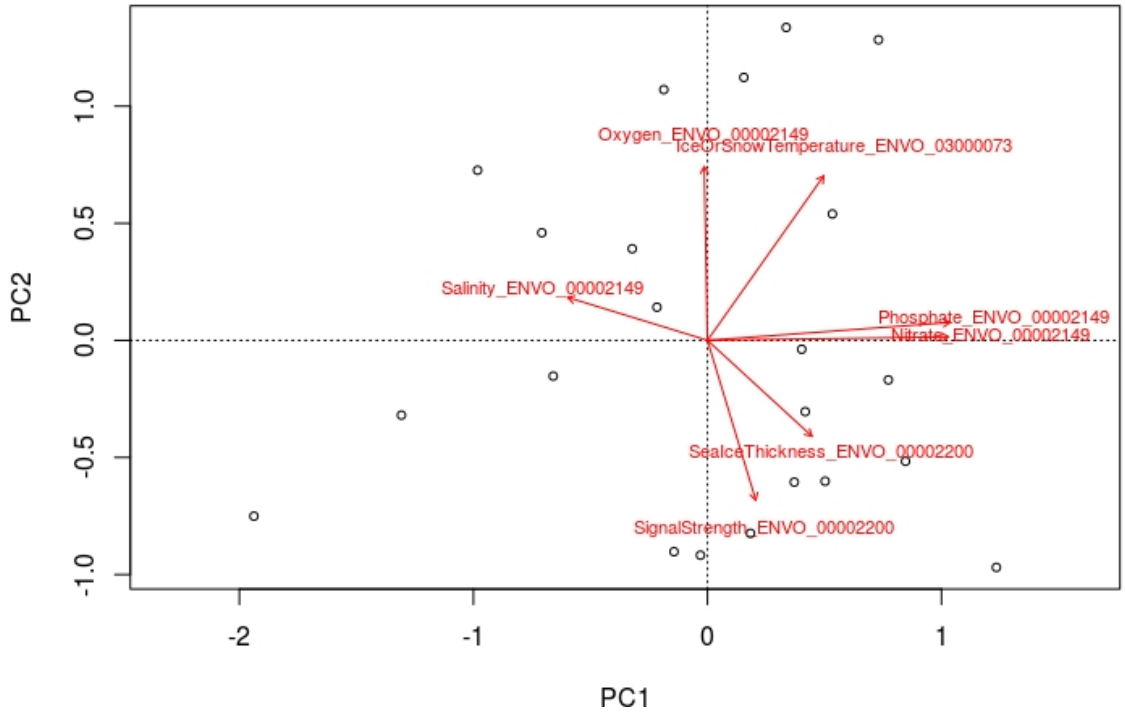


Figure 4: PCA on assembly of data about terms included as axioms of a hypothetical **environment determined by a phytoplankton community associated with sea-ice** ontology term.

Figure x, a hypothetical principal component analysis shows the effects of the various environmental variables, assembled due to their inclusion in axioms of the term **environment determined by a phytoplankton community associated with sea-ice**, as red arrows.

Included in figure x are the Environment Ontology terms which were were referenced as axioms of the hypothetical **environment determined by a phytoplankton community associated with sea-ice** term and with which annotated data was retrieved. For example "SignalStrength_ENVO_00002200" represents a column which is labeled "Signal Strength" which is about about a 'degree of illumination' which 'inheres in' some 'sea ice' ENVO_00002200.

Example results of this hypothetical analysis are that seawater phosphate and nitrate concentrations have the same effects on the dynamics of an environment determined by a sea-ice associated phytoplankton community.

The phosphate and nitrate seawater concentration explanatory variables are at a near 90 degree angle from the sea ice signal strength explanatory variables, suggesting these factors are not correlated. Where as sea ice signal strength is potentially inversely correlated with sea water oxygen concentration, as theses explanatory variables point in nearly opposite directions.

Connecting information contained within ontology terms to the term authors results

Evaluating the extent to which ontologies serve to interconnect people who contribute knowledge to the knowledge they have contributed. Queries were performed to calculate the proportions of Envo and EnvoPolar terms which are annotated with a term editor or oboInOwl created_by relation which reference an Open Researcher and Contributor ID (ORCID), a unique identifier for scientific and other academic authors and contributors [57]. The results of which are summarized in the following table:

Table 3: Percentage of Envo and EnvoPolar terms annotated with a "term editor" or "created_by" relation.

ontology	% terms with created by	% terms with term editor
Envo	14.5	4.2
EnvoPolar	17.2	31.4

Examining these results we find that in the full Environment Ontology only 4.2% of terms have a term editor annotation, contrasting with the 31.4% of terms from the Environment Ontology polar subset.

Terms related by a oboInOwl created_by relation only account for 14.5% of Environment Ontology terms, wheres they are found in 17.2% of Environment Ontology polar subset terms.

Altogether only approximately 20% of Envo terms are annotated with a term editor or created_by relation, contrasting with the nearly 50% referenced terms from the Environment Ontology polar subset.

Connecting datasets and publications about an ontology term results

We evaluated if ontologies could serve to connect users to publications about datasets annotated with ontology terms. In the example datastore there are two datasets which are annotated with a terms which satisfy the condition of being part of a marine biome: *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.* [58][59], and *Influence of snow depth and surface flooding on light transmission through Antarctic pack ice, supplementary data.* [60][61]. Both of which are about a marine water body.

Returned are the Digital object identifier (DOI) persistent uniform resource locators for the 14 publications which make use of these two example AWI datasets. For a full list see supplement **Table 12**. Selected example results of publications their digital object identifiers as well as the dataset from which the publications were retrieved are shown in **Table 5**. Retrieved publications about variety of marine water body related topics, including using chlorophyll pigments to determine phytoplankton taxomony, plankton ecology, vertical distributions of phytoplankton communities and light transmission through pack-ice.

Table 4: Selected examples of digital object identifiers 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 [62]
	10.3402/polar.v34.23349	Summertime plankton ecology in Fram Strait-a compilation of long- and short-term observations [63]
	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 [61]

Reflections upon the VoCamp Glacier Ontology Hackathon results

Semantic models for glacial-related terminology utilizing or not utilizing standardized upper levels semantic models, were proposed during the Feb 2, 2018 VoCamp Glacier Ontology Hackathon. We addressed the necessity of using both the Basic Formal Ontology and Relations ontology upper level semantic models in order to interconnect the semantically annotated contents of the example datastore.

Data annotated with unspecified ontology terms were discovered by querying the OBO knowledge graph for classes related to an input class via an input property. In this example, we queried for classes which participate in a sea ice formation process. Shown in **Table 6**, are the results of searching the knowledge graph for subclasses of sea ice formation processes (shown in the first column), related classes (third column), properties linking the subclasses to the related classes (second column), and the number of data items retrieved from the datastore annotated as being about the related class.

Table 5: Number of data items in datastore about classes which participate in sea ice formation processes

input class	property	related class	number of data items
sea ice formation process	has input	sea water	13
sea ice formation process	has output	sea ice	6
nilas formation process	has input	new ice	0
nilas formation process	has output	nilas	0
young ice formation process	has output	young ice	0
first year ice formation process	has output	first year ice	8
second year ice formation	has output	second year ice	0
multiyear ice formation process	has output	multiyear ice	8

The results of a simulation conducted to evaluate the extent to which the Basic Formal Ontology and Relations Ontology upper level semantic models are necessary in the retrieval of annotated data, are presented in **Table 7**. Lacking the Basic Formal Ontology upper level semantic model, only 54.3% of the data items were retrieved. Lacking the Relations Ontology upper level semantic model 0% of the relevant data could be retrieved.

Table 6: Percentages of data items in datastore about classes which participate in sea ice formation processes retrieved when lacking upper level semantic models.

lacking upper level semantic model	% data items retrieved
Basic Formal Ontology	54.3
Relations Ontology	0

Interconnecting stated and unstated knowledge via an ontology knowledge graph results

Treating connectivity within a network created from an ontology knowledge graph as a proxy for the extent to which ontologies connect researchers to new unspecified knowledge. We analyzed the envoPolar subset as a network. The resulting network property statistics are summarized as follows. The average degree, the number of edges corresponding to each node, is 1.517. The distributions of in-degree values, edges pointing into a node, and out-degree, edges leading away from a node are shown in supplementary **Figures 10**, and **11**. The average in-degree distribution shows a positive skew with a median of 0 relative to the mean degree of 1.517, with a very wide range of in-degree values from 0 to 44. The average out-degree distribution, also shows a positive skew with a median of 1 relative to the mean degree of 1.517, however, the out-degree values only range from 0 to 5. The analysis of additional network parameters are summarized in **Table 8**.

Table 7: Network parameters calculated from the graph of the envoPolar subset of ENVO.

network parameter	value
number of nodes	265
number of edges	402
average node degree	1.517
clustering coefficient	0.047
connected components	8
network diameter	7
mean shortest path length	2.246
average connectivity (number of neighbors)	2.875
network density	0.0
number of self-loops	0
multi-edge node pairs	20

The graph of the envoPolar subset include 265 classes represented as nodes with a total 402 connections (edges) interconnecting them. It is made up of 8 components, clusters of internally but not externally connected nodes and edges. The network diameter, the maximum distance path between two nodes, is 7. The network density, measuring how densely the network is populated with edges is 0.0. There are no self-loops, nodes with edges connecting back to themselves. There are 20 multi-edge node pairs, which measure how often neighboring nodes are linked by more than one edge.

Analysis of the distribution of shortest path lengths, the expected distance between two connected nodes is as follows. The mean shortest path length, is 2.246, and the distribution of shortest path lengths, **Figure 5**, shows a positive skew with the median shortest path length being 2.0. A value less than that of the mean.

The average connectivity, or average number of neighbors, indicating the expected number of vertices that would need to be removed to separate any randomly chosen pair of vertices is 2.875. The clustering coefficient, a

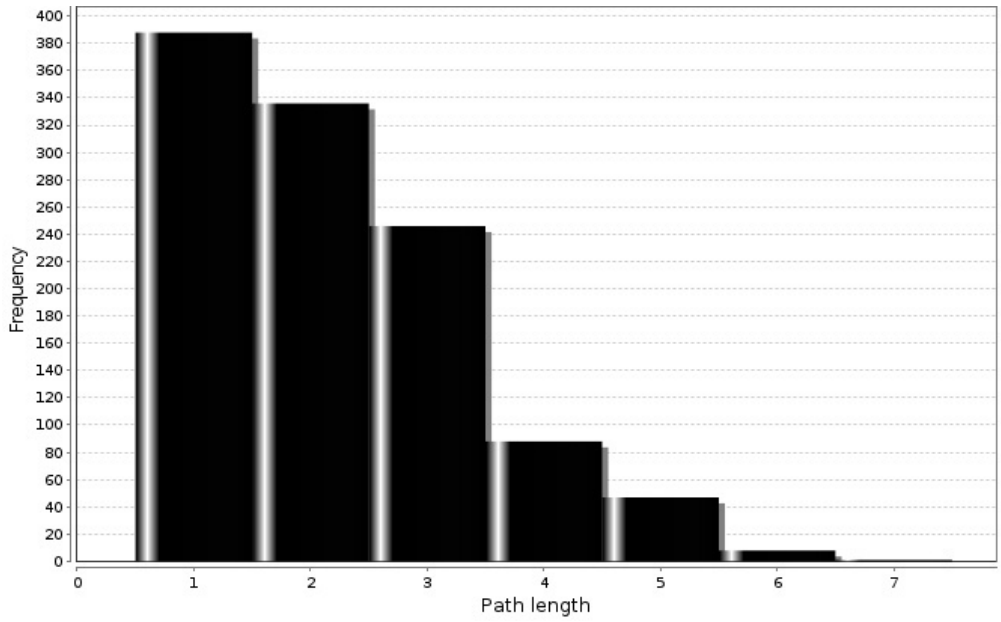


Figure 5: Distribution of shortest path lengths of the envoPolar subset analyzed as a network.

measure of the extent to which nodes in a graph tend to cluster together bounded on a scale from 0 to 1, zero being unconnected and 1 being completely connected is 0.047. Plotting the average cluster coefficients as a function of number of neighbors, see **Figure 6**, we observe two distinct clusters of nodes. Nodes either have a high average clustering coefficient and a small number of neighbors, or they have a low clustering coefficient and a large number of neighbors.

Finally we analyzed the the betweenness centrality of nodes in the envoPolar network, see supplementary **Figure 12**. Betweenness centrality of a node is a value ranging between 0 and 1, which reflects the amount of control this node exerts over the interactions of other nodes. Plotting betweenness centrality as as a function of number of neighboring nodes, we find only 3 nodes with non-zero betweenness centrality values, all of which have only 2 neighboring nodes. The rest of the nodes have a betweenness centrality of near zero, regardless of the number of neighbors.

Mobilizing ontology annotated data results

Assessing the practicality of retrieving ontology term annotated data from the example data store, we evaluated the relative amount of data which would be retrieved by users of various levels of querying expertise. Simulating the retrieval rates of data for users of the system with basic, intermediate and advanced querying expertise.

Two of these simulated querying proficiency experiment were conducted, the first queried for data which was about the exclusive AND intersection of two annotation terms. For example data about *snow and thickness*. The second simulation queried for data which is about a part of an input term, for example, *part of a glacier*. The results of the first simulation, displaying the percentages of annotation terms and data matrix columns retrieved

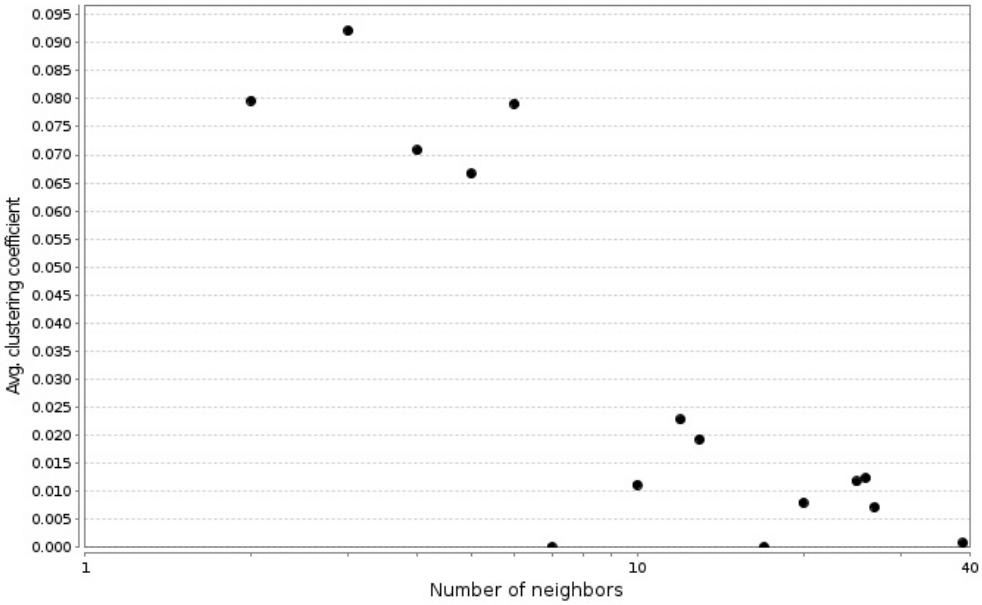


Figure 6: Average clustering coefficient as a function of number of neighboring nodes in the envoPolar subset analyzed as a network.

about the intersection of two terms, are presented in **Figure 7**.

Users with only basic querying expertise were only able to retrieve data and annotations from the *Bacteria and Archaea* case. Users with intermediate expertise were only able to retrieve data from 4 out of the 8 cases tested. Excluding the *Bacteria and Archaea* cases which was covered with 100% success by all three expertise classes, the percentage of annotations retrieved by intermediate users ranged from 40-66.7%. Whereas the percentage of data columns retrieved ranged from 50-80%. Advanced users were able to retrieve data columns and annotations from all 8 of the tested querying cases, with successes ranging from 57.1% to 100% of annotations and 66.7% to 100% of data matrix columns.

The results of the second simulation, displaying the percentages of data matrices and data points retrieved from *parts of* input terms are presented in **Figure 8**. Users with only basic querying expertise were only able to retrieve data matrices from the *parts of a marine biome* case, as well as data points from the *part of a centrally registered identifier symbol* case. In terms of the success rates of retrieving data matrices about *parts of a marine biome*, the basic user expertise case retrieved 25%, the intermediate expertise case retrieved 75% and advanced expertise case retrieved 100%. Intermediate expertise users were only able to retrieve data points from 4 out of the 8 *parts of cases*. Although advanced users were were able to retrieve data points from the majority of *parts of* query cases, they were unable to retrieve any from the *parts of a carbon atom* cases, and they were only able to retrieve 82.3% of data which is *part of a glacier*, and 94.9% of data annotated with an ontology term which is *part of an ocean*.

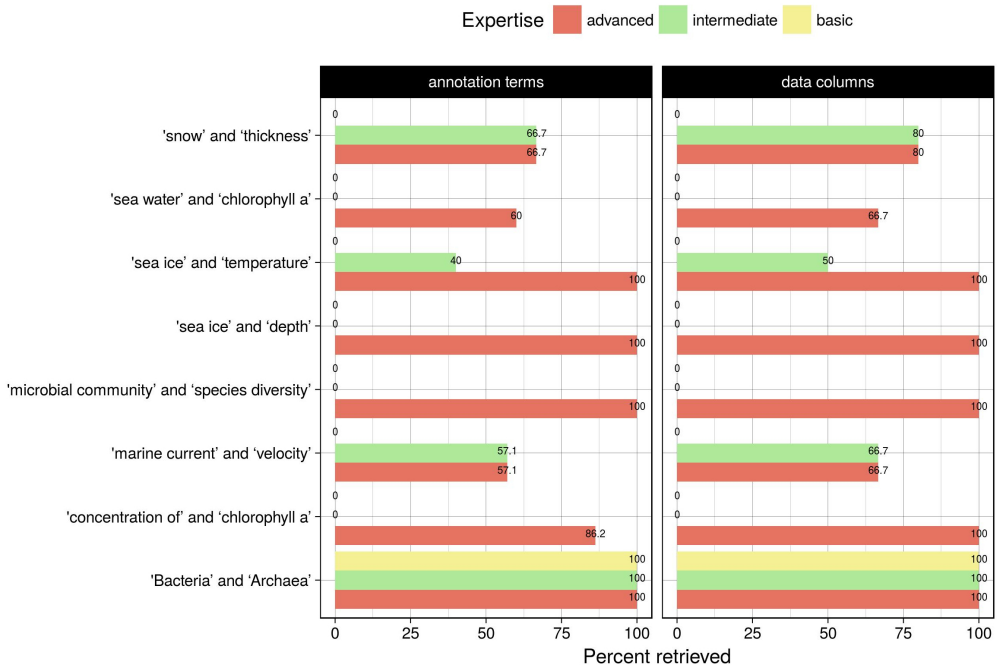


Figure 7: Analysis of querying expertise required to obtain data matrix columns and annotations when querying for data about subclasses of a term AND another term.

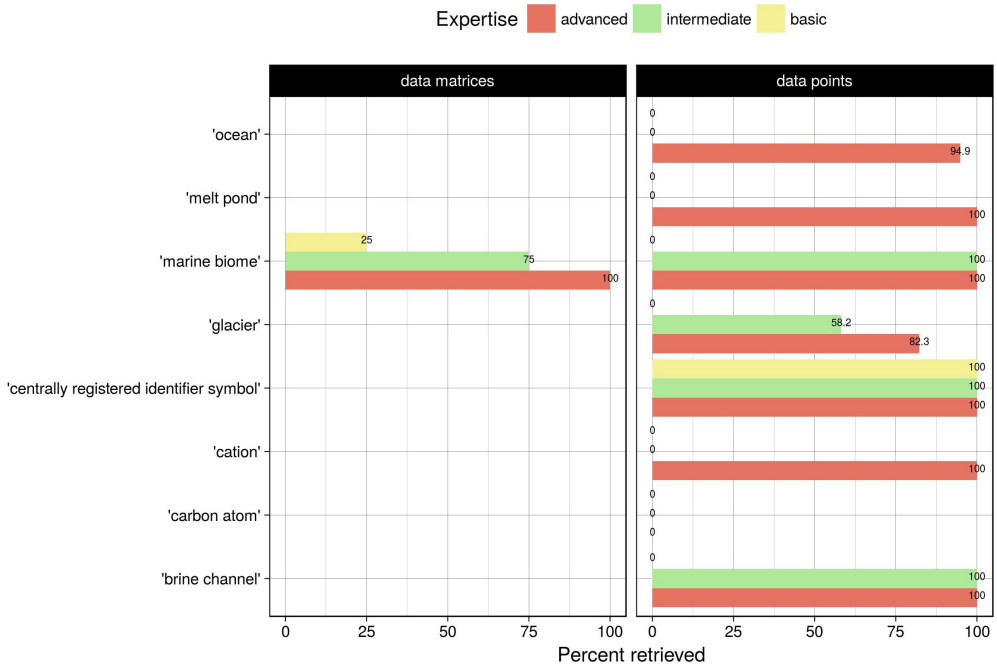


Figure 8: Analysis of querying expertise required to obtain data matrices and data points when querying for data about parts associated with an ontology term.

Ontological representation of plankton ecology related phenomena results

In the course of this research plankton ecology related ontology terms were created to enhance the Environment Ontology (ENVO) [14][15], the Phenotypic Quality Ontology (PATO) [52] and the Population and Community Ontology (PCO) [54]. For a complete description of these potential ontology see the link provided in the supplementary section.

An example of a term prepared for the Population and Community Ontology is:

phytoplankton bloom process

Defined as:

```
A plankton bloom process during which at least two of the populations in a
   community of phytoplankton, in a body of water, undergo rapid growth,
   resulting in high concentrations of phytoplankton that occur only
   periodically and briefly in that ecosystem, relative to their
   concentrations through the majority of the planetary orbital period.
```

Including subclass axioms:

`'plankton bloom process'`

`'has participant' some 'phytoplankton community'`

`'part of' some 'surface photoautotrophic biomass formation'`

`'has part' some 'population bloom'`

`'occurs in' some 'water body'`

database_cross_reference

Example terms prepared for the Phenotypic Quality Ontology are:

planktonic

Defined as:

```
An organismal quality inhering in a bearer by virtue of the bearer's inability
    to sustain directed movement to overcome displacement by physical forces
   such as currents.
```

Including the subclass axiom:

`'organismal quality'`

database_cross_reference

ice cover of a planetary surface

Defined as:

```
A physical quality which inheres in a land or water body by virtue of that
   land or water body having a two dimensional surface layer whose connection
    to some adjacent atmosphere or outer space is interrupted by ice.
```

Including subclass axioms:

'physical quality'

'inheres in' some 'planetary surface' or 'water body'

'has part' some 'surface layer' and ('adjacent to' some atmosphere or 'adjacent to' some 'outer space')

'adjacent to' some 'water ice'

'has exact synonym' 'ice cover'

'has exact synonym' 'ice coverage'

Example terms prepared for the Environment Ontology are:

marginal ice zone

Defined as:

```
An environmental zone in which is the site of the transition between the open
   ocean and sea ice.
```

Including subclass axioms:

'environmental zone'

'has quality' some 'partial ice cover'

overlaps some ('sea ice' and 'marine water body')

'causally upstream of, positive effect' some 'phytoplankton bloom process'

'causally upstream of, positive effect' some 'photoautotrophic biomass formation'

'has related synonym' 'sea ice edge'

database_cross_reference 1, database_cross_reference 2

marine environment determined by a phytoplankton community bloom

Defined as:

```
A marine environmental system which has a phytoplankton community bloom as
    part, such that the rapid growth of at least two of the populations in a
    blooming community of phytoplankton, exert a strong causal influence on
    the function of the marine environmental system, and the removal of the
    blooming phytoplankton community would cause the marine environmental
    system to collapse.
```

Including subclass axioms:

'marine environment determined by a community bloom'

'determined by' some 'phytoplankton bloom process'

'causally downstream of, negative effect' some 'marine water body stratification'

marine water body stratification

Defined as:

```
A water body stratification process during which water within a marine water
    body is separated by density into layers which sit atop one another.
```

Including subclass axioms:

'water body stratification'

'occurs in' some 'marine water body'

'results in formation of' some 'stratified marine water body'

'results in formation of' min 2 'marine layer'

sea ice melting

Defined as:

```
An icemelt process during which meltwater is produced by the melting of sea
    ice, increasing stratification in the underlying water column, and
    increasing the amount of electromagnetic radiation absorbed within the
    site previously occupied by sea ice, which acted as a medium by which to
    attenuate and reflect incoming electromagnetic radiation.
```

Including subclass axioms:

'ice melt'

'has input' some ('sea ice' and ('adjacent to' some troposphere)and ('adjacent to' some '
marine water body'))and 'decreased degree of illumination'

'has output' some (meltwater and ('adjacent to' some troposphere)and ('adjacent to' some 'marine water body'))and 'increased degree of illumination'

'causally upstream of' some 'marine water body stratification'

database_cross_reference 1, database_cross_reference 2, database_cross_reference 3

Discussion

Model polar datastore and scripts

OBO ontologies offer an interdisciplinary unified knowledge model, which provides the possibility of semantically linking data many type of data [27]. In this work we have created a model polar datastore in which data is stored in machine-readable formats and annotated with combinations of ontology terms. This allows for data to be mobilized through SPARQL queries, selectively retrieving data about terms of interest. Data annotations make use of the upper level semantic model Basic Formal Ontology (BFO) [10], and the Relations Ontology (RO) [13]. By using these upper level semantic models for data annotation, the contents of the datastore can be interconnected via the knowledge represented in the OBO ontology knowledge graph. This was realized by developing scripts which make use of the process causal algebra provided by the Relations Ontology and the well structured knowledge hierarchy provided by the Basic Formal Ontology. These scripts are able to retrieve classes pertaining to a phenomena of interest, as well as classes pertaining to connected phenomena. This allows for the interconnections between phenomena encoded into ontologies to be leveraged to interconnect interdisciplinary data. Enabling questions to be asked of interdisciplinary data. In the course of this work the semantically annotated datastore and the ontology querying scripts were employed to ask a variety of competency questions. Assessing the fitness for purpose of ontologies to interconnect interdisciplinary environmental and genomic datasets.

Interconnecting genomic and environmental data via ontology discussion

Make use of data annotated with interoperable Gene and Environent Ontology terms, data was mobilized to answer the question:

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

Table 2, showing the relative genomic and transcriptomic abundances of oxidation-reduction es in various types of marine benthic biomes, addresses the question of whether ontologies are fit for purpose to facilitate genomic comparisons based on environmental annotations. The results indicate as would be biologically expected that neritic samples are relatively enriched in photosynthesis, aerobic respiration and respiratory electron transport chain related genes. Relative to the deep abyssal and bathyal biome samples enriched in undifferentiated oxidation-reduction processes, and methanogenesis related genes. This preliminary comparison is a indicates the feasibility of using ontology term annotations to interlink and compare genomic data differentiated by source environment.

Further exploring the use of interoperable GO and ENVO semantics to compare genomic abundances of samples annotated with different ENVO terms, we asked another question:

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

Table 3, showing the relative abundance of vitamin biosynthetic process genes in various types of marine benthic biomes, further addresses the use of ontologies to interconnect genomic data. Analyzing the result of elevated vitamin biosynthetic genomic capacity in deep samples relative to the lower transcriptomic capacity in neritic samples. We made use of the ontology knowledge graph to search for the relative abundances of other biological processes and molecular functions which may help to explain the differences in vitamin biosynthesis. 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 [55][56], we investigated transition metal ion binding and transport subclasses, with the aid of the Gene Ontology knowledge hierarchy. The results of elevated ferrous ion binding and transport gene abundance in deep relative to neritic samples, combined with the elevated riboflavin biosynthetic process suggest a potential ecophysiological connection. Allowing us to hypothesize that riboflavin mediated iron reduction differentiates deep bathyal and abyssal sediments from neritic sediments. This example illustrates how using the interoperable Environment and Gene Ontologies, can be used to facilitate genomic comparisons, enabling more specific ecological questions to be asked of omic data.

As OBO ontologies adopt a realist philosophy, representing what exists in reality as opposed to conceptualizations of reality which are shared by knowledgeable agents [10]. Multiple competing hypothesis can be encoded into the ontology knowledge graph without the presumption of any being the absolute truth.

A hypothesis such as the interconnection between riboflavin production, iron binding and transport genes in deep marine sediments, could be semantically expressed and added to the ontology knowledge graph. This along with other hypotheses about covariation of gene abundances could subsequently be tested over larger collections of genomic data sets. Leveraging the ontology semantics to retrieve data to analyze gene covariation to support or reject batches of genomic hypotheses.

To further investigate potential knowledge which can be derived from the interconnection of data annotated with Gene Ontology and Environent Ontology terms, we asked the following question of the example datastore:

> "What organic acid biosynthetic process differentiate various types of marine benthic biomes?"

By drilling down into finer levels of granularity within the organic acid biosynthetic process hierarchy, see supplement figure x, we were able to pinpoint processes differentiating deep abyssal and bathyal samples from neritic. Figure 2 an examination of cellular amino acid biosynthetic processes, a subclass of organic acid biosynthetic process, shows a much clearer differentiation of samples, than does the higher level class in Figure 1. From this we were able to pinpoint even more specific subclasses to investigate differences between types of marine benthic biomes.

Figure 3 showing serine family amino acid biosynthetic processes illustrates a very clear and potentially biologically interesting difference. Glycine biosynthetic process, which has as subclass glycine biosynthetic process from serine, is more abundant in the deep samples, whereas cysteine biosynthetic process from serine is more abundant in the neretic samples.

The amino acid serine is precursor in the production of both glycine and cysteine. Therefore, from these finding we can hypothesize that organisms from neritic biomes tend to produce cysteine from serine, whereas organisms

from deep bathyal and abyssal biomes tend to produce glycine from serine. A possible explanation is that glycine is an important component in glycine betaine used by microbes as an osmoprotectant, helping to withstand osmotic stress [65]. Which may help cells to cope with high pressure deep environments.

What is most notable from this finding is that we were able to discover this potential difference based solely on information contained within the gene ontology. Having no preliminary ideas about what amino acid biosynthetic process could differentiate deep and neritic samples, nor knowledge of amino acids serine is a precursor to. This constituents an example of how ontologies are fit for purpose to interconnect disparate omic datasets and generate working hypotheses therefrom.

Ontology guided data assembly for ecological analysis discussion

To analyze how ontology terms can be used to facilitate the assembly of relevant knowledge and data about a phenomena of interest, we ask the following question:

> "What environmental factors have the greatest influence on the dynamics of a sea-ice associated phytoplankton community?"

Making use of the axioms contained in the hypothetical term **environment determined by a phytoplankton community associated with sea-ice** we retrieved an assembly of data about its constituents. The retrieved datasets contained some expected variables such as sea ice thickness, multi-year sea ice temperature, degree of illumination of sea ice. As well as some unexpected but potentially useful data such as sea water salinity, as well as seawater oxygen, phosphate and nitrate concentrations.

We conducted a PCA analysis on the retrieved data to simulate a real ecological analysis which could be performed on data sourced from the same spatiotemporal location. Giving incite into what environmental variable have an effect on the dynamics of an environment determined by a phytoplankton community associated with sea ice.

These various data may exert unexpected influence on the dynamics of the system in question. In an ideal case the ontology annotations would help to assemble data which researchers would not have thought to include in the analysis. For example the sea water salinity data which may give an indication of the existence of meltwater released from sea ice melting, which coinciding the available nutrient concentrations could indicate the beginning of a phytoplankton bloom.

Treating chlorophyll a as an indicators for primary production, additional analysis of such data could be performed using a redundancy analysis (RDA) in which the chlorophyll a concentration is treated as a response variable explained by the other variables. This workflow would serve to interconnect disparate datasets to answer ecological questions such as "What environment factors have the greatest influence on chlorophyll a concentration, (a proxy for phytoplankton growth) in an environment determined by a sea ice associated phytoplankton community?".

Connecting information contained within ontology terms to the term authors discussion

Internet technologies are changing the way in which scientific discoveries are made, enabling collaborative dissemination of knowledge and data [66]. Ontologies being semantic representations of expert knowledge have the potential to facilitate networking among scientists, allowing users to connect to the authors who have contributed their knowledge.

In order for ontologies to facilitate future scientific networking and discoveries, ontologies would benefit from more domain experts recording their knowledge into ontologies. To incentivize such actions, ontologies would benefit from micro-crediting knowledge contributions at the term level. To facilitate scientific networking, authors who contribute knowledge to ontologies should be micro-credited with an unambiguous personal identifiers. These identifiers would need to be connected to a living system which is queryable. Allowing for users to query the ontology knowledge for any authors who contributed knowledge related to specific input terminology of interest. Enable a query such as

> "Find the contract information for all authors who contributed knowledge to the sea ice terminology hierarchy."

A standard method by which to micro-credit authors within ontologies is to annotate term with a link to an Open Researcher and Contributor ID (ORCID) [57]. ORCIDs are persistent digital identifier serving as primary keys to distinguish researchers. ORCID satisfies the requirement of being a permanently maintained persistent living system. ORCID additionally provides an application programming interface by which user contact information can be queried. Authors may change affiliations or contact information multiple times throughout their career, however they would only ever use a single ORCID account. Hence ORCIDs provide a persistent unique identifier fit for the purpose of interconnecting authors of ontology terms to the knowledge they have helped to encode.

In order to evaluate the extent to which ontologies contributed to over the course of this work serve to interconnect people who contribute knowledge to the knowledge they have contributed, we asked the following question:

> "How well do the Environment Ontology and the Environment Ontology Polar subset connect authors of terms to the information they helped to encode?"

The results intricate that only 20 and 50% of terms from the Environment Ontology and its polar subset respectably contain a directly querable author annotation. Making it difficult to directly search for the author of an given term. Although ontologies such as the Environment make use of term ranges to identify authors, this information is stored in a separate meta-data owl file, which would be difficult to query for with no a priori knowledge of. The practice of using author ID ranges works for ontologies with smaller number of contributing authors, but constituents a cumbersome solution for the micro-crediting of many authors who may have only contributed to a minimum of one term. Directly annotating terms with links to contributing author ORCIDs provides a more easily scalable solution for future influxes of contributing authors. ORCIDs additionally provid direct querability for the author of a given term of interest. The direct annotation of ontology terms with links to a contributing authors ORCID, would facilitate future networking by connecting ontology terms to the contact information of

their authors through the ORCID API. From which ontology term author contact information can be pulled.

Connecting datasets and publications about an ontology term discussion

Jackson's (2012) bestiary of ignorance proposes four categories in an overview of knowledge or lack of knowledge about a subject [67]. The most subtle yet possibly most important of these categories describes unknown known knowledge. Referring to scientific knowledge which has been generated or recorded, but to which easy access is lacking [1]. Hortal et al. (2015) proposed bioinformatic methods could be employed to facilitate the access to stored but non easily search-able knowledge contained in specimen collections or scientific literature [1]. Semantic annotation of specimen collections published datasets and scientific publications, is one possible approach for such bioinformatic tools to facilitating access to unknown known information. By mobilizing known unknown information data and knowledge into a greater ontology knowledge graph, scientists may be able to overcome the limitation of known unknown knowledge.

Evaluating the fitness for purpose of ontologies to connect users to publications about datasets annotated with ontology terms, we pose the following question to our example datastore:

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

The results of which demonstrate the ontology knowledge graph can be used to direct users toward publications associated with datasets annotated with terms discovered from the ontology knowledge graph. Searching for publications about parts of a marine biome the ontology knowledge graph lead us to papers about a marine water body. In some cases this process lead us to publications written about the data set of interest. For example the publication *Influence of snow depth and surface flooding on light transmission through Antarctic pack ice* [61], about the *Influence of snow depth and surface flooding on light transmission through Antarctic pack ice, supplementary data.* dataset. This publication was retrieved due to the dataset being annotated as being a data matrix about some:

'physical quality' and ('inheres in' some ('marine water body' and ('adjacent to' some 'sea ice')))

In other cases publication less directly related to a dataset about a part of a marine biome, were retrieved. Such as for example the publication *An evaluation of the application of CHEMTAX to Antarctic coastal pigment data* [62], which made use of a subset of the data from the *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.* dataset. This publication was retrieved as it is referenced in a dataset annotated as being a data matrix about:

'chlorophyll a' and ('part of' some 'marine water body')

Although These examples constitute relatively uninteresting examples of retrieving publication referenced in a dataset about a term of interest. They demonstrate proof of concept for the interconnection of datasets and publications annotated using ontology terms. This process, could be applied to search for data annotated at a higher

level of granularity. For example to search for publications which use datasets about a prokaryotic phytoplankton community bloom occurring in icebergs calved from Antarctic glaciers in the Weddell Sea. Semantically expressed as:

phytoplankton community bloom and (composed primarily of some **prokaryotic organisms** and (overlap some output of some iceberg calving process and (located in some Weddell Sea))).

A semantic annotation such as this could be realized by using the open source gazetteer GAZ [68], an ontologically-oriented listing of place names. Which could be used to provide the semantic annotation for a specific geographic feature of interest such as the Weddell Sea. Terms such as **phytoplankton community bloom** or **prokaryotic organisms** could be provided by the Population and Community Ontology. The objective being to interconnect data sets and publications annotated at a very specific level of granularity. Allowing users to ask such as:

> "What publications reference datasets about prokaryotic phytoplankton community bloom occurring in icebergs calved from Antarctic glaciers in the Weddell Sea?".

These same semantic data annotation, querying and retrieval principals could also be used to facilitate the search for information about specimens. For example if natural history collections containing preserved Alveolata (dinoflagellates), were to encode information about the morphologies of such specimens in queryable formats with ontology term annotations. Providing an specimen annotation such as:

> NamedIndividual and (subClassOf some 'Alveolata' and ('has role' some 'specimen role' and ('has quality' some 'morphology')))

Users would be enabled to ask a question of the now queryable natural history specimen collections such as:

> "What are all possible morphologies of Alveolata species for which there are collected specimens?"

This would go along way toward facilitating the ease of access to knowledge about unconnected parts of the collective scientific knowledge base. Helping the scientific community to overcome the challenge of coping with unknown known knowledge.

Reflections upon the VoCamp Glacier Ontology Hackathon discussion

Participation in the Feb 2, 2018 VoCamp Glacier Ontology Hackathon [69], an event to create lightweight vocabularies and ontologies for the semantic web, was an opportunity to experiment with semantic models for the annotation of polar data. As well as to aid in the clarification of glacier related semantics.

An example of the clarification of glacier related semantics, which took place during the "hackathon" was coming to a consensus definition of **ablation**. In the **ablation** example the NOAA National Weather Service Glossary 2009 [70] stipulates the restriction that only melting and evaporation processes contribute to ablation. The Cogley et al. IACS-UNESCO Glacier Mass Balance 2011 [71] 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.

As ontologies take an agnostic stance when representing knowledge which has multiple definitions or which pertains to competing hypotheses [10]. A variety of approaches can be taken in parallel to incorporate such definitional discrepancies into the ontology knowledge graph. 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 both the pre and post composed versions of the term would be equivalent in terms of machine-searchability. 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.

Further knowledge gained form the VoCamp Glacier Ontology Hackathon was an analysis of the variety of semantic models for expressing glacial related terminology, proposed during the event. These proposed glacial related semantic models varied in their usage of upper level semantic models. Some proposed models involved the creation of new upper level semantic models, while others made use of existing upper level semantic models such as the Basic Formal Ontology and Relations Ontology.

We examined the feasibility of using these various proposed semantic models for the mobilization of polar data. Evaluating the extent to which the Basic Formal Ontology and Relations Ontology upper level semantic models are necessary to be able to retrieve data from our example datastore. To such effect, we asked the following questions:

> "What percentages of these data items about participants in sea ice formation processes would we be able to retrieve if we didn't use the Basic Formal Ontology or Relations Ontology upper level semantic models?"

To answer such questions we simulated the effects of not utilizing the Basic Formal Ontology or Relations Ontology upper level semantic models to retrieve data from the example polar datastore. The results of these simulations indicate that the upper level semantic models are crucially important in order to retrieve data about classes participating in a phenomena of interest. In the simulation, if fundamental changes were made to the Basic Formal

Ontology, only 50% of the data would be retrievable. Additionally, If a non standardized set of relations were to be used in place of the structured Relations Ontology relations, it would not be be possible to retrieve data about classes participating in a phenomena of interest from the model datastore.

These results imply that upper level semantic model are crucially important when using using ontology terms to mobilize linked data. We are not necessarily advocating for the use of specific upper level models such as those employed in OBO Foundry and Library ontologies, namely the Basic Formal Ontology and the Relations Ontology. What is essential, however, is upper level semantic models be standardized and consistently used. The OBO Foundry and Library ontologies would be a reasonable choice for such standardization, as they provide an already existing, interoperable semantic infrastructure. Including the arguably most successful and widely used biomedical ontology, the Gene Ontology [16]. Additionally, efforts are underway to align the OBO ontologies, primarily the Environment Ontology with the NASA Semantic Web for Earth and Environmental Technology (SWEET) ontologies [72]. The SWEET ontologies being the *de facto* standards for the formal representation of earth and environmental science domain knowledge [73]. Aligned OBO and SWEET semantic hold great potential to aid in the future interconnection of environmental and genomic data.

Interconnecting stated and unstated knowledge via an ontology knowledge graph discussion

We assessed if ontological knowledge graphs are fit for the purpose of connecting information or data explicitly stated by a researcher to new, unstated but related knowledge or data. Assuming connectivity within the envoPolar knowledge graph is analogous to the facility of researchers searching the network to discover new data and knowledge associated with their stated input knowledge. We created a network out of the Polar subset of the Environment Ontology and assessed its network parameters calculated as a directed graph to attempt to answer the following question:

> "Is the ontology knowledge graph of the envoPolar subset sufficiently well connected to be able to lead researchers to new knowledge via unstated linkages to identified knowledge?"

Our analysis of the properties of the network envoPolar created from the envoPolar subset is as follows. The network has a low number of components, with the vast majority of nodes and edges belonging to the largest components. Therefore the network should be analyzed as a relational based regime as opposed to a competent based regime. This is logical as the network makes use of structured upper level semantic models of the Basic Formal Ontology and the Relations Ontology as has been previous discussed.

We additionally examined the diameter of the network, the longest possible path between connected nodes in a network to gain incite into how well integrated the network is. Longer maximum path lengths equate to less well integrated networks [74]. In the envoPolar network, the maximum path length is only 7. Examining the distribution of path lengths, see **Figure 5** we observe that the majority of nodes have a path length of 2. Meaning that the average node in the network is only 2 steps away from most other nodes. Hence the overall network is very well connected.

Examining the in-degree distributions of nodes in the network, see supplemental **Figure 10**, we remark that some nodes have very many connections, while the majority of nodes have only a few (one or two) connections. A network containing very few highly connected nodes and very many poorly connected nodes, implies it bears a centralized network structure. This can also be seen in **Figure 6** the graph of average clustering coefficient as a function of number of neighboring nodes. Where we observe two distinct clusters of nodes. Nodes higher up in the hierarchy are well connected to a small number of neighbors. While nodes lower down in the hierarchy are poorly connected to a larger number of neighbors. A third network parameters additionally indicating a very centralized network structure is the distribution of betweenness centrality values. Supplemental **Figure 12** shows a distribution where only 3 nodes, each of which only have two neighbors, have elevated betweenness centrality values, reflecting the amount of control these nodes exerts over the the interactions of other nodes. Demonstrating that the network is highly centralized, with three very important central nodes which exert control over all other nodes in the network. Geographic feature is an example of one of these central nodes, having the largest degree of in-connectivity in the envoPolar graph. This is logical as this node is relatively high in the material entity hierarchy. The nodes in-degree value of 44 means there are 44 classes in the envoPolar network which fall underneath the geographic feature hierarchy.

Highly centralized networks are termed scale free or power law networks [75], which describe an exponential relationship between the degree of connectivity a node has and the frequency of it's occurrence. Examining the topology of the envoPolar network we observe a hierarchical and branched tree-like structure. Branching structures are typically much more efficient ways of connecting networks, as the branching structures provide an exponential growth in the number of nodes that can be reached relative to the path length traversed [76]. Allowing for a very short average path length within a very large network, which is what we observe in the envoPolar network.

In terms of robustness a scale free network won't be dramatically affected by removing or changing low degree nodes, however it would be very affected if the central nodes were removed or changed. If for example the Basic Formal Ontology hierarchy we no longer used and suddenly a very central node such as geographic feature were to be removed without replacement, the network would shatter into many unconnected components, rendering it unable to interconnect information. Structured as is, however, with the majority of nodes only being two steps away from highly centralized and well connected "hub" nodes. The network is very highly interconnected. This is due to the hierarchical organizational structure of the ontology. Resulting in a very well connected network which could serve to lead researchers to new knowledge via unstated linkages to identified knowledge.

Mobilizing ontology annotated data discussion

In order to asses the practicality of retrieving ontology annotated data from our example datastore, which contains data annotated with axiomatic structures of various levels of complexity. We asked the following question:

> What level of querying expertise is required to access the various types of data contained in the example polar datastore?

Lacking the scope to be able to conduct a proper experiment on a study group of scientists with various proficiencies for performing semantic queries to retrieve data from the example datastore. We evaluated this question by performing a series of data queries using three levels of omitted axiom property paths. As to simulate what percentage of total data would be retrieved by users with basic, intermediate and advanced querying expertise. The results of these simulations indicated that users of basic querying expertise were only able to retrieve a tiny fraction of annotated data, users of intermediate expertise less than half the data and even users advanced users could not fully retrieve all data. Possible reasons for such results may be due to non-uniformities in axiomatic structures used to annotate the example data.

Examining the semantic data annotation models employed to post-compositionally create owl axioms. Axioms which contain ontology terms which render the data search-able by querying for specific ontology terms or combinations of terms which make up the data annotation. We investigated the axiomatic patterns for their query-ability.

An example, of an axiomatic pattern which was successfully queryable by users of intermediate expertise, which annotates data about **snow thickness** and **mean snow thickness** is as follows:

> thickness and ('inheres in' some 'snow')

This example axiomatic annotation structure, about a thickness quality which is realized in the material entity snow is relatively straight forward. Employing the pattern of:

```
class A, relation 1, some class B
```

When creating an axiomatic data annotation there is trade-off between a complete and correct philosophical description of a subject v.s. a more pragmatic linked data approach intended to make data easily mobilizable. The example of data about **mean snow thickness**, presents an axiomatic pattern which is a reasonable comprise between a correct description and an easily mobilizable data annotation. In order to semantically represent data about a more complicated phenomena, questions arise about how to address such a trade-off. For example data about **mean snow thickness**, which was not retrieve by the advanced querying proficiency simulation, is annotated with the axiom:

> 'expected value' and 'is about' min 2 ('data item' and 'is about' some (thickness and 'inheres in' some 'snow'))

In this example the data is an expected value (a mean) of other data which are about a thickness quality which is realized in some snow. In this example the axiomatic annotation is not as straight forward taking the form:

```
class A, relation 1, cardinality value, class B, relation 2, some class C,
   relation 3 some class D
```

The non-uniformity of this query, which makes use of a cardinality value as opposed to the general "some" article, prevents general querying patterns from accessing this data. This is due to the the different owl syntax for handing restrictions about "some" v.s. "min" or "max". Performing SPARQL queryies on axioms which are chained together, as is the case for owl annotated data, requires the use of extended property path within the SPARQL

query. An example property path required for a SPARQL query to be able to accesss owl annotated data is as follows:

```
owl:someValuesFrom/owl:intersectionOf/rdf:rest*/rdf:first
```

Where the `owl:someValuesFrom` is the relation to navigate through the "some" article. Due to the constraint of needing to use extended property paths to perform SPARQL queries on owl annotations, individual cases would need to be created to cover every possible case, if for example the article were variable in the property path. This would be a very cumbersome solution impeding the retrieval of annotated data. It is worth considering the utility of expressing the fact that we are describing an expected value about at least two data points. Especially if it impedes data mobilization. A compromise solution would be to make use of a more standard annotation pattern, for example of the form:

```
class A and (any relation some class B and (any relation some class C and
    (...)))
```

A highly uniform and easily expendable pattern such as this would facilitate the retrieval of data, as well as provide sufficient depth to satisfactorily describe a phenomena of interest. Satisfying both the ontology philosopher and the pragmatic mobilizer of data.

Ontological representation of plankton ecology related phenomena discussion

Attempting to build upon the semantic framework proposed as necessary by Stec et al. (2017), for future modeling of plankton ecosystems [24]. Proposed plankon ecology related ontology terms were created. Evaluating the proposed ontology terms for the extend to which they encode knowledge about phenomena relating to plankton ecology, we asked the question:

> "What specific examples of expert knowledge about phenomena relating to plankton ecology are represented in the proposed ontology term contributions?"

Discussed are some examples of expert knowledge of plankton related phenomena which has been captured and represented in the proposed ontology terms. Represented in the definition of the proposed quality of an organism term **planktonic**, is the classic oceanography definition of plankton, characterizing them as drifting organisms unable to swim against a current [77].

Examining the phenomena of physical processes which effect phytoplankton blooms, we encoded the following knowledge. The onset of phytoplankton blooms have been shown to be dependent on the timing of the retreat of melting sea ice [78]. We have encoded the knowledge of such a relationship into the potential term **sea ice melting**, by specifying that it is `'causally upstream of' some 'marine water body stratification'`.

We have encoded the concept of the **marginal ice zone**, which is described as the transition zone between the open ocean and sea ice [79]. In the marginal ice zone, melting sea-ice has been shown to promote phytoplankton

growth by stratifying the water column [80]. To encode the knowledge of the process of stratification we created the potential class **marine water body stratification**, which results in the formation of at least two layers of marine water. In the marginal ice zone, stratification of the water body resulting from melting sea ice has been shown to be the location of maximum chlorophyll [80], controlling the onset of the seasonal phytoplankton blooms [78]. To encode knowledge about phytoplankton blooms, we created a variety of terms related to population and community blooms. An example of which is the potential term **phytoplankton bloom process**. To represent the knowledge that phytoplankton blooms tend to occur as a result of sea ice retreat in the marginal ice zone. We have encoded into the **marginal ice zone** term an axiom stating that marginal ice zones are `'causally upstream of, positive effect' some 'phytoplankton bloom process'`. As phytoplankton bloom processes have a profound impact on their surrounding environment, we have additionally created the term **marine environment determined by a phytoplankton community bloom**, which we have specified to be related to phytoplankton bloom processes with the axiom `determined by' some 'phytoplankton bloom process'`.

In the marginal ice zone, water body stratification has been argued to be a major limiting parameters controlling seasonal phytoplankton blooms which are mediated by the sea-ice cover [78]. To make the concept of sea-ice cover machine-actionable, we have created a potential new quality term for **ice cover of a planetary surface** to described the area of ice coverage as a two dimensional surface layer on a planetary surface in which ice interrupting the connection to the atmosphere. Such a term is intended to be used to annotate ice cover data.

Conclusion

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

- Conclusion: Nodes are organized such that the most nodes are connected only to a local hub which is well connected to other hubs

Hierarchal graph structure serves to make for very well connected “hub” nodes to access other nodes from; facilitating knowledge interconnection. Hierarchal graph structure ensures graph is well connected via “hubs”, facilitating knowledge linking

Outlook

//outlook? 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.

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.

creation of tools to help semantically annotate data using ontology terms.

Polar knowledge application ontology

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

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).

For example semantic representation of the causal linkages between sea ice melting, increasing stratification and decreasing vertical mixing could be used to mobilize and integrate relevant data to better understand ice-melt associated phytoplankton blooms. The knowledge layer could be programmed to link AWI autonomous buoy sea ice thickness and physical oceanography temperature and salinity data, with external NASA earth observatory and National Snow and Ice Data Center satellite chlorophyll and sea ice cover data. Semantically linked into a machine readable knowledge graph, in-depth questions concerning the interplay between sea ice melting and phytoplankton bloom processes could be asked of these data. Additionally, this semantic layer could be used to prototype a system which dynamically merges and analyses these data, integrating new data as it is published.

Although this work would provide data engineering services, the principle objective would be to produce semantic research outputs, forming and testing hypotheses about knowledge representation.

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

//keep for sure original MIxS paper: [81]

//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.

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

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.

[22] (citation for dpbedia)

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

linking ontology mobilized data to United Nations' Sustainable Development Goals / 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.

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.

As a final component, I would research the semantics required to link the Biological Oceanography section's publications and data outputs to the United Nations' Sustainable Development Goals, Development Targets and indicators, as well as the Essential Ocean Variables. This work would be connected to Dr. Buttigieg's work on the recently funded INTERNAS project concerning knowledge transfer between science and policy domains.

//maybe come to this stuff ESPI wanting to get more socioeconomic value out of earth science data. Realizing the socioeconomic Value of Data https://2018esipwintermeeting.sched.com/event/D6oC

role of data in 2015 - 2020 ESIP Strategic Plan

link to my log

http://wiki.esipfed.org/index.php/2015-2020_Strategic_Plan

References

1. **Hortal J, Bello F de, Diniz-Filho JAF, Lewinsohn TM, Lobo JM *et al.*** Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 2015;46:523–549.

2. **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.

3. **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.

4. **Wang M, Overland JE**. A sea ice free summer arctic within 30 years? *Geophysical Research Letters*;36. Epub ahead of print 2009. DOI: 10.1029/2009GL037820.

5. **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.

6. **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.

7. **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.

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

9. **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.

10. **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.

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

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

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

14. **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.

15. **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.

16. **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.

17. **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.

18. **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.

19. **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).

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

21. **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).

22. **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.

23. **Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C *et al.*** The monarch initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research* 2016;45:D712–D722.

24. **Stec KF, Caputi L, Buttigieg PL, D'Alelio D, Ibarbalz FM *et al.*** Modelling plankton ecosystems in the meta-omics era. Are we ready? *Marine Genomics* 2017;32:1–17.

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

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

27. **Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R *et al.*** Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE* 2014;9:e89606.

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

29. **Torres-Martinez E, Paules G, Schoeberg M, Kalb MW**. A web of sensors: Enabling the earth science vision. *Acta Astronautica* 2003;53:423–428.

30. Welcome to the OGC OGC. http://www.opengeospatial.org/ (accessed 28 February 2018).

31. **Bröring A, Echterhoff J, Jirka S, Simonis I, Everding T *et al.*** New generation sensor web enablement. *Sensors* 2011;11:2652–2699.

32. Data Publisher for Earth & Environmental Science. https://www.pangaea.de/ (accessed 22 February 2018).

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

34. **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).

35. **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).

36. **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.

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

38. **Team R**. Rdflib: RDFLib is a Python library for working with RDF, a simple yet powerful language for representing information. https://github.com/RDFLib/rdflib.

39. **Bushnell B**. BBMap. *SourceForge*. https://sourceforge.net/projects/bbmap/ (accessed 22 February 2018).

40. **Kopylova E, Noé L, Touzet H**. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012;28:3211–3217.

41. **Pruesse E, Peplies J, Glöckner FO**. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 2012;28:1823–1829.

42. **Rho M, Tang H, Ye Y**. FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research* 2010;38:e191–e191.

43. **Finn RD, Mistry J, Tate J, Coggill P, Heger A *et al.*** The pfam protein families database. *Nucleic Acids Research* 2009;38:D211–D222.

44. **Meinicke P**. UProC: Tools for ultra-fast protein domain classification. *Bioinformatics* 2014;31:1382–1388.

45. **Mitchell JB**. Enzyme function and its evolution. *Current Opinion in Structural Biology* 2017;47:151–156.

46. **R Core Team**. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/ (2013).

47. **Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P *et al.*** *Vegan: Community ecology package*. https://CRAN.R-project.org/package=vegan (2017).

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

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

50. **Shannon P**. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* 2003;13:2498–2504.

51. 9.7. Statistics — Mathematical statistics functions — Python 3.6.4 documentation. https://docs.python.org/3/library/statistics.html (accessed 23 February 2018).

52. Phenotypic Quality Ontology - Summary NCBO BioPortal. http://bioportal.bioontology.org/ontologies/PATO (accessed 11 May 2017).

53. Ecocore: An ontology of core ecological entities. https://github.com/EcologicalSemantics/ecocore (2018).

54. Population and Community Ontology - Summary NCBO BioPortal. https://bioportal.bioontology.org/ontologies/PCO (accessed 11 May 2017).

55. **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.

56. **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.

57. Credit where credit is due. *Nature* 2009;462:825–825.

58. **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.

59. **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.

60. **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, supplementary data. Epub ahead of print 2017. DOI: 10.1594/PANGAEA.870706.

61. **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.

62. **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.

63. **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.

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. **Meury J**. Glycine betaine reverses the effects of osmotic stress on dna replication and cellular division in escherichia coli. *Archives of Microbiology* 1988;149:232–239.

66. **Nielsen M**. *Reinventing discovery : The new era of networked science*. Princeton, N.J: Princeton University Press; 2012.

67. **Jackson ST**. Representation of flora and vegetation in quaternary fossil assemblages: Known and unknown knowns and unknowns. *Quaternary Science Reviews* 2012;49:1–15.

68. An open source gazetteer constructed on ontological principles. https://github.com/EnvironmentOntology/gaz (2015).

69. Contribute to Virtual-Hackahon-on-Glacier-topic development by creating an account on GitHub. https://github.com/Vocamp/Virtual-Hackahon-on-Glacier-topic (2018).

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

71. **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).

72. SWEET Overview SWEET. https://sweet.jpl.nasa.gov/ (accessed 26 February 2018).

73. **DiGiuseppe N, Pouchard LC, Noy NF**. SWEET ontology coverage for earth system sciences. *Earth Science Informatics* 2014;7:249–264.

74. **Labs C**. Network Diameter. *Complexity Labs*. http://complexitylabs.io/network-diameter/ (2016, accessed 26 February 2018).

75. Scale-free graph with preferential attachment and evolving internal vertex structure. *Journal of Statistical Physics* 2013;151:1175–1183.

76. **Kou J, Chen Y, Zhou X, Lu H, Wu F *et al.*** Optimal structure of tree-like branching networks for fluid flow. *Physica A: Statistical Mechanics and its Applications* 2014;393:527–534.

77. **Lalli C**. *Biological oceanography : An introduction*. Oxford England: Butterworth Heinemann; 1997.

78. **Janout MA, Hölemann J, Waite AM, Krumpen T, Appen W-J von *et al.*** Sea-ice retreat controls timing of summer plankton blooms in the eastern arctic ocean. *Geophysical Research Letters* 2016;43:12, 493–12, 501.

79. The marginal ice zone. *Norwegian Polar Institute*. http://www.npolar.no/en/facts/the-marginal-ice-zone.html (accessed 27 February 2018).

80. **Cherkasheva A, Bracher A, Melsheimer C, Köberle C, Gerdes R *et al.*** Influence of the physical environment on polar phytoplankton blooms: A case study in the fram strait. *Journal of Marine Systems* 2014;132:196–207.

81. **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.

82. **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.

83. **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.

84. **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.

85. **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.

86. **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.

87. **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.

88. **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.

89. **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.

90. **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.

91. **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.

92. Python Data Analysis Library — pandas: Python Data Analysis Library. https://pandas.pydata.org/ (accessed 23 February 2018).

93. **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.

94. **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.

95. **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.

96. **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.

97. **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.

98. **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.

99. **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.

100. **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.

101. **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.

102. **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.

Appendices

Model polar datastore creation

Datasets employed

The following datasets were included in the example datastore:

1. Inorganic nutrients measured on water bottle samples at AWI HAUSGARTEN during POLARSTERN cruise MSM29. [82]
2. Physical oceanography and current meter data from mooring TD-2014-LT. [83]
3. Chlorophyll a measured on water bottle samples during POLARSTERN cruise ARK-XXIV/2. [84][63]
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. [58][59]
5. Biogenic particle flux at AWI HAUSGARTEN from mooring FEVI7. [85][86]
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. [87][88]
7. Sea ice thickness at Ice Camp 1 on 2013-09-01 (GEM2IceTh_DiveHole_IceStation1). [89][61]
8. Influence of snow depth and surface flooding on light transmission through Antarctic pack ice, supplementary data. [60][61]
9. Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-2_11. [90][91]
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.
11. Unpublished transcriptomic data from shallow Helgoland Marine Sediments during a spring phytoplankton bloom, courtesy of David Probandt, and Matthew Schechter.

Model polar datastore annotations

Detailed description of axioms making use of ontology terms used to post-compositionally annotate the example polar datastore is available from here

Python scripts

The follow python scripts were developed and used in the course of this work.

1. `create_rdf_triples_from_csv_files.py` to convert csv files into RDF triple files formated using the turtle specification. The script makes use of the any23 rover tool [33]. Available from here.

2. `merge_triples_to_datastore.py` to create a single datastore file in turtle specification format by merging together data files and data annotation files. Available from here.

3. `query_annotation_of_data_files_data_or_columns_about_input.py` to query the local datastore for data matricies or data matrix columns which area annotated with input ontology terms. Available from here.

4. `query_data_set_references.py` to query the local datastore for all data from data matrix columns which are annotated with a has database cross reference hasDbXref. Available from here.

5. `query_for_classes_linked_by_input_classes_and_input_properties.py` to query the ontobee programmatic SPARQL endpoint for ontology classes which contain axiomatic linkages between a set of input ontology class terms and a set input ontology property terms. Returning a three column csv file with columns for input terms, connective properties, and classes connected to the input term via the input property. Available from here.

6. `query_for_data_about_exclusive_and.py` to query the local datastore for data matrix columns which are annotated with an ontology terms about both of two input classes. Returning the data matrix columns and assocaited data. Available from here.

7. `query_for_parts_associated_with_input_class.py` to query the ontobee programmatic SPARQL endpoint for parts associated with an input class. Returning a list of ontology classes. Available from here.

8. `query_for_subclasses_of_input_purl.py` which queries for subclasses of an input ontology class term against the ontobee programmatic SPARQL endpoint. Returning a list of ontology classe terms including the input class, its subclasses and their subclasses recursively. Available from here.

9. `query_for_subproperties_of_input_purl` which queries for subproperties of an input ontology property term against the ontobee programmatic SPARQL endpoint. Returning a list of ontology property terms including the input class, its subproperties and their subproperties recursively. Available from here.

10. `query_GO_annotation_of_data_files_csv_annotations_columns.py` to query the local datastore for for data matrix about an input list of annotation terms, for columns which annotated with GO terms matching an input list of ontology classes. The script makes use of the python Python Data Analysis Library (pandas) Version: 0.22.0 [92]. The script returns a data table with rows corresponding to samples, and columns corresponding to relative abundances of GO terms. Available from here.

Interconnecting genomic and environmental data via ontology supplemental

Abyssal and Bathyal metagenomic data provided by Jose Rapp consisted of four samples collected from Polarstern cruise PS85, at stations 470, 460, 464, 465. Samples 1 and 2 were collected from depths of 1244m and 2403m which best correspond to 'marine bathyal zone biome' Samples 3 and 4 were collected from depths of 3531m and 5525m which best correspond to 'marine abyssal zone biome'

Neritic transcriptomic data provided by Dr. David Probandt, were collected from shallow ~8m depth Helgoland Marine Sediments during a spring phytoplankton bloom. Sediments were characterize as being 'sandy sediment' from a 'marine neritic benthic zone biome'. The first 4 samples: labeled X1-X4 were used.

Table 8: Full results of the relative genomic and transcriptomic abundances of oxidation-reduction processes 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

Table 9: Full results of the relative genomic and transcriptomic abundances of transition metal ion transport process in various types of marine biomes.

label	marine abyssal zone biome	marine bathyal zone biome	marine neritic benthic zone biome
ferrous iron transport	0.04	0.04	0.02
mercury ion transport	0.01	0.00	0.00
nickel cation transmembrane transport	0.00	0.00	0.00
transition metal ion transport	0.00	0.00	0.00
iron ion transmembrane transport	0.00	0.00	0.00
iron ion transport	0.00	0.00	0.00
copper ion transmembrane transport	0.00	0.00	0.00
cobalt ion transport	0.00	0.00	0.00
copper ion transport	0.00	0.00	0.00

Table 10: Full results of the relative genomic and transcriptomic abundances of transition metal ion binding molecular functions in various types of marine biomes.

label	marine abyssal zone biome	marine bathyal zone biome	marine neritic benthic zone biome
transition metal ion binding	0.70	0.67	0.52
cobalt ion binding	0.69	0.73	1.07
ferric iron binding	0.22	0.21	0.62
nickel cation binding	0.12	0.13	0.12
molybdenum ion binding	0.05	0.05	0.10
manganese ion binding	0.04	0.04	0.02
iron ion binding	0.04	0.04	0.03
zinc ion binding	0.03	0.03	0.11
ferrous iron binding	0.02	0.03	0.00
copper ion binding	0.01	0.01	0.02
copper chaperone activity	0.00	0.00	0.00

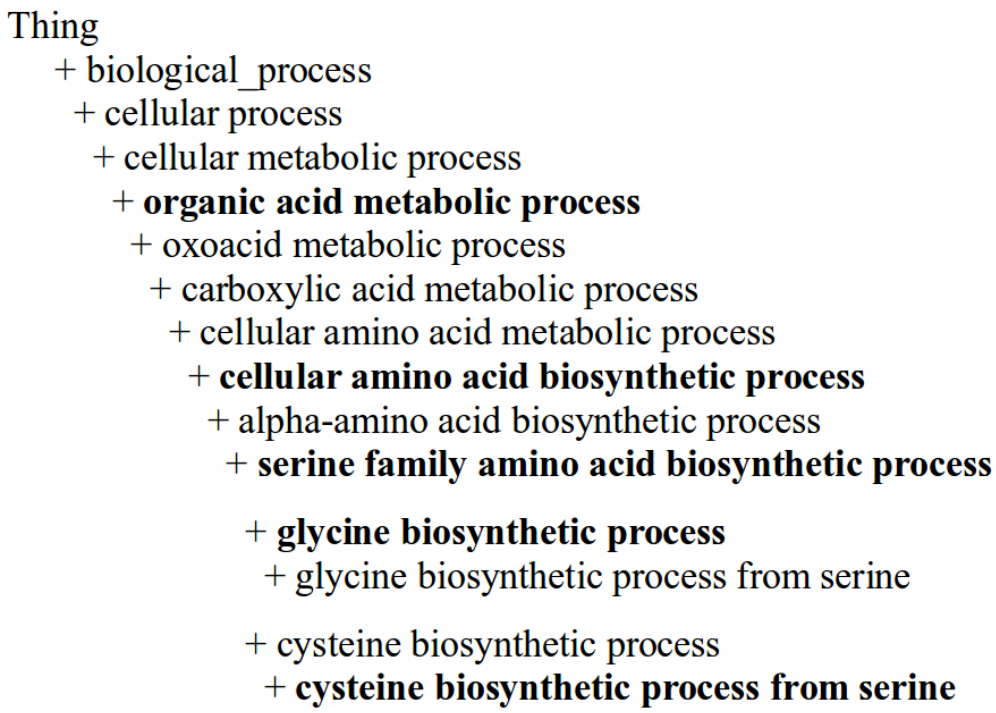


Figure 9: Gene Ontology biological process hierarchy differentiating neritic v.s. bathyal and abyssal benthic biome deep marine sediments.

Subclasses of boldfaced terms **organic acid biosynthetic process**, **cellular amino acid biosynthetic process**, and **serine family amino acid biosynthetic process** are the subjects of the Principal coordinate analyses plots in Figures 1, 2 and 3 respectably. Serine family amino acid biosynthetic process terms differentiating neritic v.s. bathyal and abyssal benthic biome deep marine sediments, shown in Figure 3, are **glycine biosynthetic process**, and **cysteine biosynthetic process from serine**

Code for solutions to competency questions available from here.

Ontology guided data assembly for ecological analysis supplemental

The hypothetical term **environment determined by a phytoplankton community associated with sea-ice** follows the design pattern of several preexisting terms in the Environment Ontology such as environment determined by a biofilm on a fungal surface within the environmental system hierarchy.

Data matrix columns utilized in the principal component analysis include:

1. influence_snow_depth.csvSignalStrength
2. inorganic_nutrients.csvNitrate
3. physical_oceanography.csvOxygen
4. inorganic_nutrients.csvPhosphate
5. physical_oceanography.csvSalinity

6. ice_algal_chlorophyll_myi.csvIceOrSnowTemperature
7. influence_snow_depth.csvSeaIceThickness

Code for solutions to competency questions available from here.

Connecting information contained within ontology terms to the term authors supplemental

Code for solutions to competency questions available from here.

Connecting datasets and publications about an ontology term supplemental

Table 11: Complete list of digital object identifiers of publications obtained querying for references of datasets which are about part of a marine biome.

data annotation	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 [62]
	10.1016/j.pocean.2009.07.011	Dimethylsulphide, DMSP-lyase activity and microplankton community structure inside and outside of the Mauritanian upwelling [93]
	10.5194/bg-9-1041-2012	Environmental control on the variability of DMS and DMSP in the Mauritanian upwelling region [94]
	10.3390/rs61010089	Global Retrieval of Diatom Abundance Based on Phytoplankton Pigments and Satellite Data [95]

data annotation	reference doi	reference title
	10.5194/bgd-10-12115-2013	Photophysiological state of natural phytoplankton communities in the South China Sea and Sulu Sea [96]
	10.1016/j.dsr.2014.12.010	Physiological characteristics of open ocean and coastal phytoplankton communities of Western Antarctic Peninsula and Drake Passage waters [97]
	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 [98]
	10.5194/bg-10-3297-2013	Sulphur compounds,methane, and phytoplankton: Interactions along a north-south transit in the western Pacific Ocean [99]
	10.3402/polar.v34.23349	Summertime plankton ecology in Fram Strait-a compilation of long- and short-term observations [63]
	10.5194/essd-5-109-2013	The MAREDAT global database of high performance liquid chromatography marine pigment measurements [100]

data annotation	reference doi	reference title
	10.1029/2003EO380001	Unique data repository facilitates ocean color satellite validation [101]
	10.5194/os-11-139-2015	Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations [102]
	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 [61]

Code for solutions to competency questions available from here.

Reflections upon the VoCamp Glacier Ontology Hackathon supplemental

Participants in the Feb 2, 2018 VoCamp Glacier Ontology Hackathon are listed in the following table x.

Table 12: Participants in the February VoCamp Glacier Ontology Hackathon

participant	affiliation
Pier Buttigieg	Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research
Brandon Whitehead	Centre for Agriculture and Biosciences International (London)
Siri Jodha Singh Khalsa	National Snow and Ice Data Center (Czech Republic)
Kai Blumberg	Max Plank Institute for Marine Microbiology
Gary Berg-Cross	Independent Consultant Potomac, MD
Ruth Duerr	Ronin Institute Boulder Colorado
Varanka Dalia	United States Geological Survey (Rolla Missouri)
Samantha Arundel	United States Geological Survey (Rolla Missouri)
Nancy Wiegand	University of Wisconsin-Madison
Torsten Hahmann	University of Maine
Brodaric Boyan	Natural Resources Canada
Gaurav Sinha	Ohio University
Charles F. Vardeman II	University of Notre Dame
Mark Schildhauer	University of California, Santa Barbara
Steven Chong	University of Arizona

Code for solutions to competency questions available from here.

Interconnecting stated and unstated knowledge via an ontology knowledge graph supplemental

The analysis is available for inspection from here

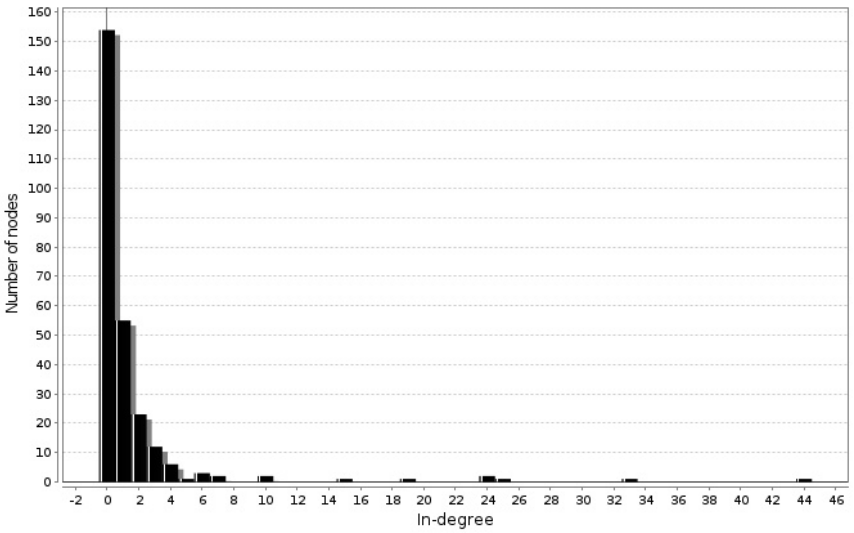


Figure 10: In degree distribution of the envoPolar subset analyzed as a network.

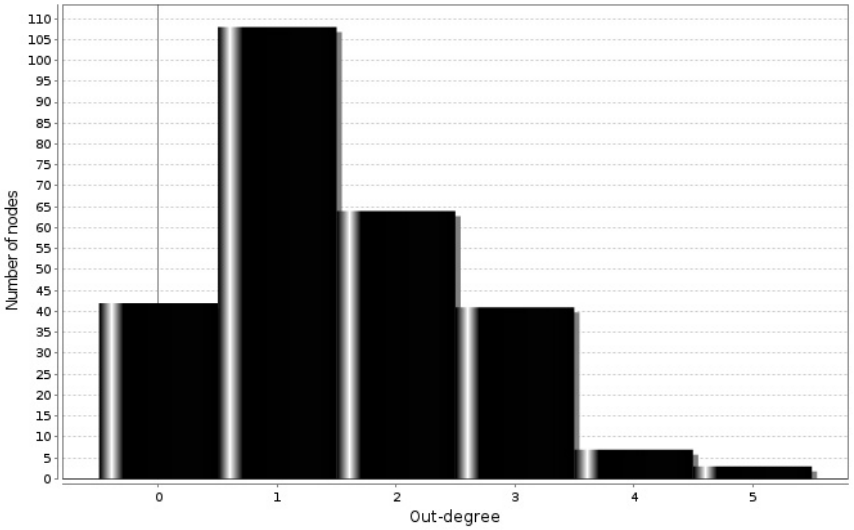


Figure 11: Out degree distribution of the envoPolar subset analyzed as a network.

Code for solutions to competency questions available from here.

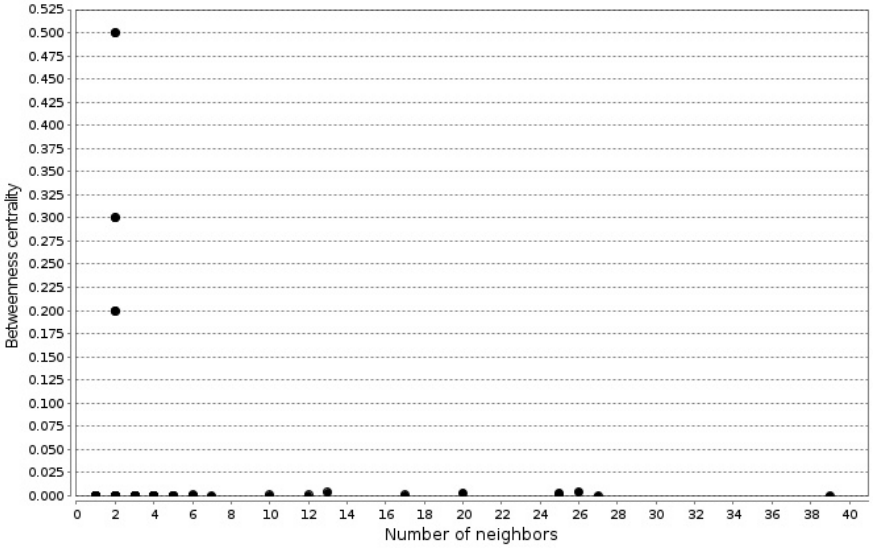


Figure 12: Plot of betweenness centrality as a function of number of neighbors of nodes from the envoPolar subset analyzed as a network.

Mobilizing ontology annotated data supplemental

The query expertise simulation for mobilizing data annotated with an exclusive AND intersection of two ontology terms made use of the following combinations of terms:

```
'Bacteria' and 'Archaea'
'concentration of' and 'chlorophyll a'
'marine current' and 'velocity'
'microbial community' and 'species diversity'
'sea ice' and 'depth'
'sea ice' and 'temperature'
'sea water' and 'chlorophyll a'
'snow' and 'thickness'
```

The querying expertise simulation for mobilizing data annotated as being about parts associated with an ontology term made use of the following input terms:

```
'brine channel'
'carbon atom'
'cation'
'centrally registered identifier symbol'
'glacier'
'marine biome'
'melt pond'
'ocean'
```

Table 13: Querying cases used to differentiate the expertise levels for the data mobilization simulations.

Expertise level	querying cases used
basic	1,2,3
intermediate	1,2,3,6
advanced	1,2,3,6,7,8,9
complete model	all

Querying cases included in each level of querying expertise:

Basic expertise:

```
X in (X)
X1-Xn in (X1 or X2 or ...)
X1-Xn in (X1 and X2 and ...)
```

Intermediate expertise:

```
X in (X)
X1-Xn in (X1 or X2 or ...)
X1-Xn in (X1 and X2 and ...)
Y in (X1 and X2 ... and ('any property' some Y))
```

Advanced expertise:

```
X in (X)
X1-Xn in (X1 or X2 or ...)
X1-Xn in (X1 and X2 and ...)
Y in (X1 and X2 ... and ('any property' some Y))
Y1-Yn in (X1 and X2 ... and ('any property' some (Y1 and Y2 and ... )))
Y1-Yn in (X1 and X2 ... and ('any property' some (Y1 or Y2 or ... )))
Z in (X1 and X2 ... and ('any property' some (Y1 and Y2 and ... and ('any
   property' some Z))))
```

complete model:

```
X in (X)
X1-Xn in (X1 or X2 or ...)
X1-Xn in (X1 and X2 and ...)
X1-Xn in ((X1 or X2 or ...) and ... )
Y and Z in (X1 and X2 ... and ('any property' some Y and 'any property' some Z
    ))
Y in (X1 and X2 ... and ('any property' some Y))
Y1-Yn in (X1 and X2 ... and ('any property' some (Y1 and Y2 and ... )))
Y1-Yn in (X1 and X2 ... and ('any property' some (Y1 or Y2 or ... )))
Z in (X1 and X2 ... and ('any property' some (Y1 and Y2 and ... and ('any
    property' some Z))))
Y in (W and ('any property' some ('any property' some X) and ('any property'
    some (Y and 'any property' some Z))))
Z in (W and ('any property' some ('any property' some X) and ('any property'
    some (Y and 'any property' some Z))))
X1-n in (W1 and W2 ... and 'any property' min N (X1 and X2 ... and  'any
    property' some (Y1 and Y2 and ... 'any property' some Z)))
Y1-n in (W1 and W2 ... and 'any property' min N (X1 and X2 ... and  'any
    property' some (Y1 and Y2 and ... 'any property' some Z)))
Z in (W1 and W2 ... and 'any property' min N (X1 and X2 ... and  'any property
    ' some (Y1 and Y2 and ... 'any property' some Z)))
```

Code for solutions to competency questions available from here, and here.

Ontological representation of plankton ecology related phenomena supplemental

Link to page containing additional plankton ecology related potential ontology terms, available from here